ИПМ им.М.В.Келдыша РАН

Абрау-2019 • Труды конференции

А.А. Печников

Свойства коммуникационного графа научно-образовательного Веба

Рекомендуемая форма библиографической ссылки

Печников А.А. Свойства коммуникационного графа научно-образовательного Веба // Научный сервис в сети Интернет: труды XXI Всероссийской научной конференции (23-28 сентября 2019 г., г. Новороссийск). — М.: ИПМ им. М.В.Келдыша, 2019. — С. 561-571. — URL: http://keldysh.ru/abrau/2019/theses/10.pdf doi:10.20948/abrau-2019-10

Размещена также презентация к докладу

Свойства коммуникационного графа научно-образовательного Веба

А.А. Печников

Институт прикладных математических исследований — обособленное подразделение ФИЦ "Карельский научный центр Российской академии наук"

Аннотация. Веб-граф является наиболее популярной моделью фрагментов реального Веба, применяемой в науке о Вебе. Исследование сообществ в веб-графе способствует лучшему пониманию организации фрагмента Веба и процессов, происходящих в нём. Сообщества, получаемые для веб-графов реальных фрагментов Веба, часто слабо различимы и с большим трудом поддаются содержательной интерпретации. Предлагается выделить в веб-графе коммуникационный граф, содержащий только те вершины (и дуги между ними), которые имеют встречные дуги, и уже в нём исследовать задачу разбиения на сообщества. По аналогии с социальными исследованиями связи, реализуемые через ребра в коммуникационном графе, предлагается называть «сильными», а все остальные «слабыми». На сильных связях строятся тематические сообщества, имеющие содержательную интерпретацию. В то же время слабые связи способствуют коммуникациям между сайтами, не имеющими общих признаков по сфере деятельности, географии, подчиненности и т.д., и в основном сохраняют связность фрагментов Веба даже при отсутствии сильных связей. Эксперименты, проведенные для фрагмента научно-образовательного Веба России, показывают возможность содержательной интерпретации полученных результатов и перспективность такого подхода.

Ключевые слова: веб-граф, коммуникационный граф, сообщество в графе, сила связей

Properties of communication graph of academic Web

A.A. Pechnikov

Institute of Applied Mathematical Research of the Karelian Research Centre of the Russian Academy of Sciences

Abstract. Web graph is the most popular model of real web fragments used in Web science. The study of communities in the web graph contributes to a better understanding of the organization of the web fragment and the processes taking place in it. Communities of the web-graph fragments of the real Web are often poorly differentiated and more difficult meaningful interpretation. It is proposed to select in the web graph a communication graph containing only those vertices (and arcs

between them) that have counter arcs, and already in it to investigate the problem of splitting into communities. By analogy with social studies, the ties implemented through the edges in the communication graph are proposed to be called “strong” and all others “weak”. Thematic communities with meaningful interpretation are built on strong ties. At the same time, weak ties facilitate communication between sites that do not have common features in the sphere of activity, geography, subordination, etc., and basically keep the web fragments connected even in the absence of strong ties. Experiments carried out for a fragment of the academic Web of Russia show the possibility of meaningful interpretation of the results and the prospects of this approach.

Keywords: web graph, communication graph, community in graph, strength of ties

1. Введение

Исследование графов реального (и виртуального) мира является важной задачей во многих областях, таких как биология, социология, социальные сети, вебометрика, и многие другие, поскольку позволяют понять структуру объектов и проанализировать их свойства. Достаточно давно известно, что вузовские и академические фрагменты Веба как России, так и других стран [1-3] обладают достаточно специфическими свойствами, характеризующими их структуру. В частности, в соответствующем веб-графе имеется большая компонента сильной связности и значительное количество «висячих» сайтов (имеющих либо только ссылки, сделанные с них, либо, что реже, ссылки сделанные только на них).

Компонента сильной связности сама по себе является интересным объектом исследований, позволяющим установить, например, наличие или отсутствие свойства «малого мира» [4], которая, однако, мало что дает в объяснении процессов возникновения гиперссылок между сайтами фрагментов Веба. В этом смысле гораздо больше пищи для размышлений дают такие структурные элементы графа, как сообщества сайтов, когда «внутри» сообществ сделано больше ссылок, чем между сообществами. Но опять-таки, попытки разбиения вершин, составляющих максимальную компоненту связности веб-графа на сообщества, приводят к сложно интерпретируемым результатам.

Основная идея, рассматриваемая в этой работе, заключается в том, чтобы построить некий «жесткий каркас» для фрагмента Веба, оставив только те сайты, которые имеют встречные гиперссылки, и уже на этом «каркасе» проверить свойства разбиения на сообщества (такой каркас далее будем называть коммуникационным графом). И уже далее с помощью известных алгоритмов исследуется вопрос о структуре сообществ вершин, коммуникационного графа и «слабого» веб-графа, из которого удален коммуникационный граф. В качестве реального объекта исследований взят научно-образовательный фрагмента Веба, для которого дается содержательная интерпретация полученных результатов.

2. Базовые понятия, методы и инструменты

Целевое множество сайтов задается прямым перечислением их доменных имен, и в результате сканирования находятся все связывающие их гиперссылки. В случае, когда сайты относятся к одному виду деятельности, такое множество называется тематическим. Соответственно, (тематический) фрагмент Веба – это целевое множество сайтов и множество связывающих их гиперссылок [3].

Веб-граф фрагмента Веба – это ориентированный граф без петель и кратных дуг, множество вершин которого представлено целевым множеством сайтов, а множество дуг строится следующим образом: две вершины связаны дугой, если есть хотя бы одна гиперссылка, связывающая соответствующие сайты.

Сообщество (кластер, модуль, группа) графа неформально можно определить как множество вершин с бóльшим количеством дуг между собой, чем с остальными вершинами графа [5].

Более строго разбиение графа на сообщества можно определить через понятие модулярности. Модулярность – это свойство графа и некоторого разбиения его на подграфы (модули-сообщества). Мера модулярности показывает, насколько данное разбиение качественно в том смысле, что существует много дуг, лежащих внутри подграфов-сообществ, и мало дуг, лежащих вне подграфов (соединяющих сообщества между собой).

Меру модулярности Q можно задать следующей формулой [5]:

$$Q = \frac{1}{2m}\sum_{ij}\left(A_{ij} - \frac{k_i k_j}{2m}\right) * \delta(c_i, c_j).$$

Здесь m – количество дуг в графе,

A_{ij} – элемент матрицы смежности графа,

k_i, k_j – степени соответствующих вершин,

$\delta(c_i, c_j)$ – символ Кронекера, вычисляемый по формуле: $\delta(c_i, c_j) = \begin{cases} 1 : if\ c_i = c_j \\ 0 : if\ c_i \neq c_j \end{cases}$, где c_i, c_j - метки класса соответствующих вершин.

Соответственно, наилучшим разбиением на сообщества считается разбиение, на котором достигается максимум Q.

Для сбора и анализа данных использовались программа для поиска и сбора внешних гиперссылок *BeeCrawler* [6] и базы данных внешних гиперссылок [7]. Для исследования веб-графов применялась открытая платформа для визуализации графов *Gephi* [8], в которой, среди многих, реализованы программы вычисления коэффициента модулярности и поиска сообществ в графе с использованием алгоритмов, предложенных в [9].

3. Фрагмент научно-образовательного Веба России

Исходное целевое множество сайтов фрагмента научно-образовательного Веба России сформировано на конец 2017 года, то есть после присоединения

Российской академии медицинских наук и Российской академии сельскохозяйственных наук (РАСХН) к РАН, но в процессе создания укрупненных структур типа федеральных исследовательских центров, а также укрупнения вузов. Такой «footprint» позволяет рассчитывать на то, что фрагмент Веба содержит достаточно устойчивые ссылки между сайтами, которые сформировались в течение нескольких предыдущих лет. Исходное целевое множество содержит 867 сайтов (279 университетов и 588 научных учреждений). Под научными учреждениями понимаются организации РАН от уровня институтов до региональных научных центров и отделений РАН (сайт собственно РАН *www.ras.ru* в целевое множество не включён, поскольку является мощным коммуникатором, существенно искажающим картину в свою пользу). Далее для краткости научные учреждения будем называть «институтами», понимая несколько расширительное толкование этого термина.

Сканирование всех 867 сайтов позволяет получить около 20000 гиперссылок, сделанных между этими сайтами, считая кратные ссылки. Заменим кратные ссылки на одинарные дуги и получим веб-граф, содержащий 867 вершин и 5030 дуг. Проверка на связность и сильную связность показывает наличие 73 изолированных вершин и 236 висячих (36 вершин не имеют входящих дуг и 200 (!) – исходящих).

В содержательном плане заметим, что 73 изолированные вершины в подавляющем большинстве относятся к сайтам институтов, ранее входившим в состав РАСХН.

Единственная компонента сильной связности (КСС) содержит 534 вершины и 4026 дуг. Веб-граф, равный максимальной КСС, имеет диаметр, равный 10 и коэффициент модулярности 0.398 [10], и при этом разбивается на 7 сообществ, содержащих от 40 до 139 вершин. Коэффициент модулярности говорит о невысоком стремлении сайтов организовываться в сообщества: значения Q принадлежат отрезку [-1, 1], а «хорошей» считается кластеризация где-то при Q больше 0.6.

Тем не менее, одно из построенных сообществ имеет хорошее содержательное объяснение. В него входят 40 вершин, соответствующих 4 сайтам университетов и 36 сайтам институтов, и всего 97 дуг (что в среднем существенно меньше, чем в веб-графе КСС). Из 40 сайтов 35 принадлежат институтам нынешнего Отделения сельскохозяйственных наук РАН, а также Красноярскому научному центру СО РАН, Кемеровскому технологическому институту пищевой промышленности, Кубанскому государственному технологическому университету, Воронежскому государственному лесотехническому университету и Санкт-Петербургскому горному университету.

Объяснения:

- Красноярский научный центр СО РАН на период исследования успел стать федеральным исследовательским центром, в состав которого входят 2 сельскохозяйственных института;

- Кемеровский технологический институт пищевой промышленности близок по роду деятельности к сельскому хозяйству;
- Кубанский государственный технологический университет имеет тесные связи с сельскохозяйственными институтами Кубани;
- Воронежский государственный лесотехнический университет на своем сайте имеет ссылку на сайт Центральной научной сельскохозяйственной библиотеки (которая попадает под термин «институты»);
- Санкт-Петербургский горный университет имеет на своем сайте ссылку на сайт Всероссийского института генетических ресурсов растений имени Н.И. Вавилова, где расположен раздел журнала «АПК», выписываемый библиотекой университета, и ещё одну ссылку на сайт Центральной научной сельскохозяйственной библиотеки.

Не считая последних двух случаев можно сказать, что данное сообщество имеет ярко выраженную сельскохозяйственно-агропромышленную тематику.

4. Коммуникационный граф веб-графа

Ньюман и соавторы при решении задачи о разбиении графа на сообщества избегают ориентированности графа, когда речь идет о веб-графах. Цитируя [10, стр. 026113-5]: «… Некоторые сети являются ориентированными, т. е. их ребра работают только одном только направлении. Веб является таким примером; ссылки в Вебе указывают в одном направлении, только от одной веб-страницы до другой. … Однако мы обнаружили, что во многих случаях лучше игнорировать направленный характер сети при нахождении структура сообществ. Часто ребро действует просто как указание на связь между двумя узлами, и направление не имеет значения».

А если направление (ориентация) имеет значение? Посмотрим рисунок 1, где изображено сельскохозяйственно-агропромышленное сообщество.

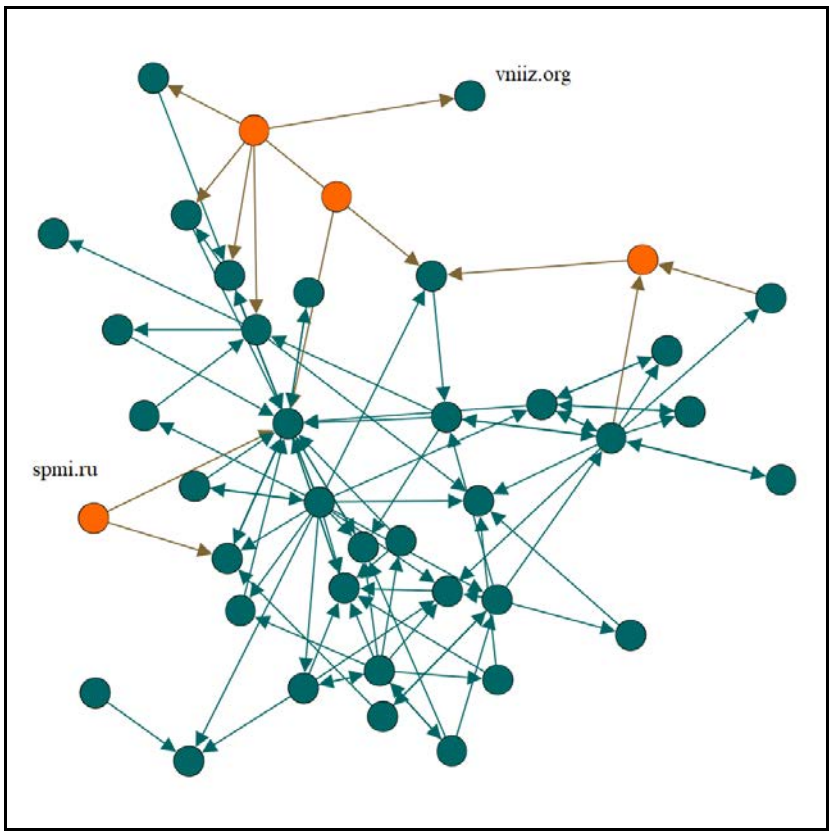

Рис. 1. Граф «сельскохозяйственно-агропромышленного сообщества»

Возникают вопросы: можно ли сайт Всероссийского НИИ зерна и продуктов его переработки (вершина *vniiz.org*) считать полноправным участником сообщества, хотя он не имеет ни одной ссылки на другие сайты? И можно ли считать участником сообщества сайт Санкт-Петербургского горного университета (*spmi.ru*), когда на него нет ни одной ссылки с других участников?

Термин «сообщество» (*Gemeinschaft*) возник в германской социологии в конце XIX века, и подразумевают совместную деятельность его участников, имеющих общие цели [11], и в этом контексте ответ на сформулированные вопросы должен быть отрицательный.

Поэтому далее рассмотрим неориентированный граф, который по построению подразумевает «совместную деятельность», а именно, встречные гиперссылки. Коммуникационный граф веб-графа – это неориентированный граф, имеющий то же самое множество вершин, что и веб-граф, а множество рёбер строится по следующему правилу: ребро (i,j) принадлежит множеству ребер коммуникационного графа тогда и только тогда, когда в веб-графе существуют дуги (i,j) и (j, i).

Коммуникационный граф, построенный таким образом, может иметь несколько компонент связности и/или изолированные вершины. В этом случае мы исключаем изолированные вершины (поскольку они не влияют на связность), и изучаем компоненты связности каждую по отдельности, начиная с максимальной.

Коммуникационный граф веб-графа научно-образовательного Веба содержит 313 вершин и 468 ребер (рис. 2).

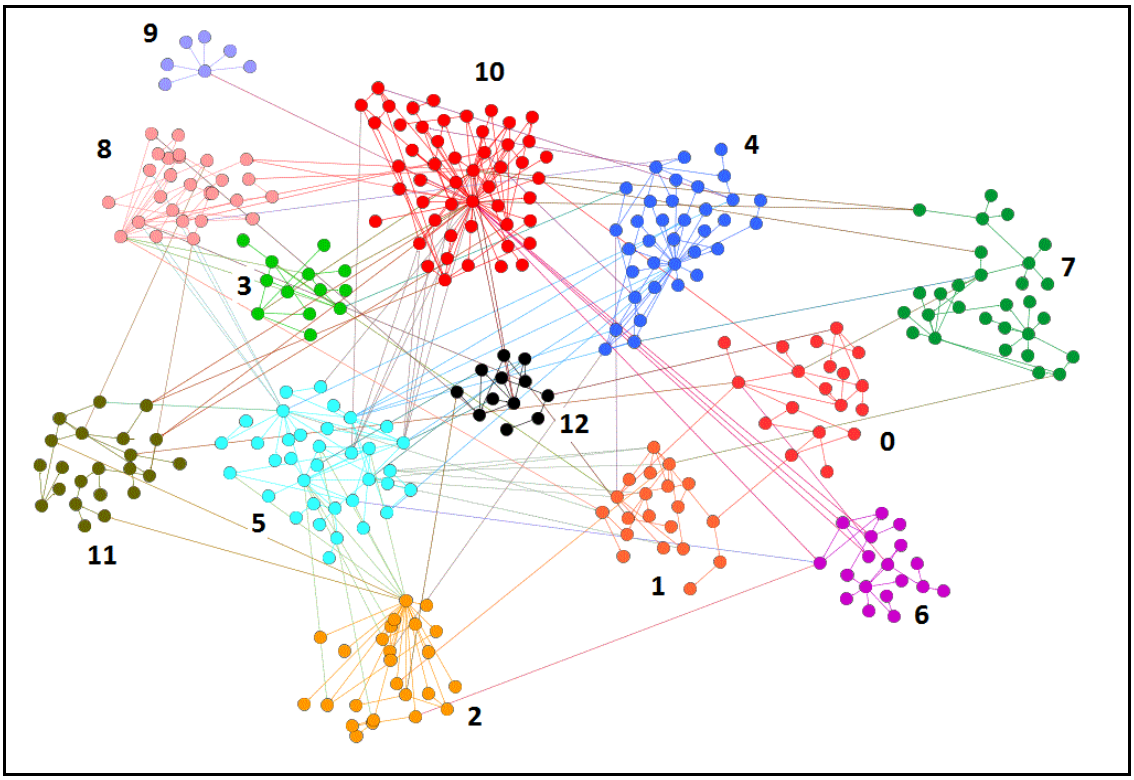

Рис. 2. Коммуникационный граф научно-образовательного Веба

Из 313 вершин 67 относятся к университетам, а 246 к институтам, то есть доля университетов в коммуникационном графе сократилась на одну десятую. Раскраска вершин дана в соответствии с полученным разбиением на сообщества.

Коэффициент модулярности данного разбиения равен 0.695, то есть достаточно высок. Построенные 13 сообществ содержат от 7 до 51 вершины, из них 4 можно достаточно точно идентифицировать по одному из двух признаков: география или научное направление.

Некоторым из построенных сообществ можно дать содержательную (тематическую) интерпретацию.

Четко выделяются 2 «географических» сообщества:

№2 – «Урал»: 1 университет (Пермский политехнический), 23 института УрО РАН и примкнувший к ним Научный центр волоконной оптики РАН из Москвы;

№10 – «Сибирь»: 4 сибирских университета, 47 сибирских институтов и 1 институт из Севастополя.

«Гуманитарное» сообщество №12 содержит 10 институтов (археология, этнография, лингвистические исследования, языкознание, история, мировая литература, славяноведение, философия) и два университета (Алтайский и Горно-Алтайский).

«Медицинское» сообщество №9 представляет собой «пучок» сайтов медицинских институтов Сибири и Дальнего востока, связанных с сайтом Сибирского отделения медицинских наук, но не связанных между собой.

Участники любого сообщества коммуникационного графа присутствуют в веб-графе в силу построения, хотя не обязательно, чтобы все участники из одного сообщества коммуникационного графа входили в одно и то же сообщество веб-графа (в нашем случае это верно для значительной части сообществ коммуникационного графа). Обратное неверно: сайты-участники агропромышленного сообщества веб-графа полностью отсутствуют в коммуникационном графе.

Теперь удалим из веб-графа КСС все пары встречных дуг, соответствующие ребрам коммуникационного графа. Полученный «слабый» веб-граф достаточно велик – 528 вершин и 3090 дуг, но имеет низкий коэффициент модулярности 0,356 и разбивается 8 сообществ, не имеющих убедительной содержательной интерпретации.

Найдем максимальную КСС «слабого» веб-графа, удалим все вершины, не вошедшие в КСС и получим КСС «слабый» веб-граф с 461 вершиной и 2744 дугами. Опять-таки, сохраняется низкий коэффициент модулярности.

5. Обсуждение результатов. Сильные и слабые связи.

Посмотрим динамику изменений доли институтов и университетов в рассмотренных графах, для чего сведем данные в таблицу 1.

Табл. 1. Доля сайтов институтов и университетов в графах

Граф	Кол-во вершин	Кол-во дуг (ребер)	Доля институтов	Доля университетов	Диаметр	Средняя длина пути
Начальный веб-граф	867	5030	0,68	0,32	-	-
КСС веб-граф	534	4026	0,67	0,33	10	3,6
Коммуникационный граф	313	936 (468)	0,78	0,22	9	3,45
«Слабый» веб-граф	528	3090	0,66	0,34	-	-
КСС «слабый» веб-граф	461	2744	0,63	0,37	11	4,2

Более 45 лет назад Грановеттером [12] была предложена концепция силы социальных связей. Все связи разделяются на две категории, – сильные и слабые, – с целью формализации межличностных отношений на основе длительности и частоты контактов. Например, сильная связь присуща друзьям, а слабая – соседям. Попробуем использовать понятие силы связи как аналогию для сайтов, хотя изначально понятно, что аналогия далеко не полная, поскольку Грановеттер считает, что связи симметричны.

Возьмем пару вершин i, j ориентированного графа. Если в графе существуют дуги (i,j) и (j,i), то будем говорить о сильной связи вершин i, j.

Если же для вершин *i, j* существует только одна из дуг (*i*,*j*) или (*j*,*i*), то будем говорить о слабой связи вершин *i* и *j*.

Начальный научно-образовательный веб-граф как и КСС веб-граф содержат сильные и слабые дуги, в коммуникационном графе отсутствуют слабые дуги, а в «слабом» веб-графе и КСС «слабом» веб-графе отсутствуют сильные дуги. Из табл. 1 видно, что соотношение институты/университеты практически одно и то же для всех графов, имеющих слабые или сильные и слабые дуги, но резко меняется для графа, содержащего только сильные дуги. Можно сделать вывод о том, что сильные связи более присущи институтам, чем университетам.

В [12, стр. 1362] отмечается, что «… чем сильнее связи, тем больше похожи индивиды друг на друга в разных аспектах». Это подтверждается в случае коммуникационного графа научно-образовательного Веба. Можно сказать, что сильные связи способствуют появлению устойчивых тематически «похожих» сообществ. И хотя таких сообществ немного, всего 4 из 13, они имеют четкую содержательную интерпретацию.

Однако если коммуникационный граф «нарастить» до КСС веб-графа (а тем более, до начального веб-графа), то оказывается, что ни одно из четырех сообществ коммуникационного графа не только не явилось основой для сообществ в веб-графах, но даже их участники попали в разные сообщества. Похоже, что слабые ссылки ведут к «размыванию» сообществ.

Правда, в КСС веб-графе мы обнаружили «сельскохозяйственно-агропромышленное» сообщество. Но вспомним, что в начальном веб-графе были найдены 73 изолированные вершины, также относящиеся к сельскохозяйственным институтам. Связывая эти два результата, возможно следующее объяснение такого исключения: сайты бывшей РАСХН просто не успели установить ссылки с другими сайтами целевого множества. Поэтому те сайты РАСХН, которые были как-то связаны между собой, имели мало ссылок «вовне» и организовали сообщество, а остальные остались изолированными.

Какова же роль слабых ссылок? По Грановеттеру в социологии слабые связи можно охарактеризовать как систему нерегулярных контактов, не охватывающих друзей индивида, а выходящих на представителей других тесно связанных групп, в которых он не состоит. Из этого следует, что индивид может устанавливать связи с людьми из значительного числа взаимно непересекающихся групп, при этом, не являясь членом каждой из них. В этом смысле слабые связи играют роль «мостов» в графе. Напомним, что мост – это ребро (неориентированного) графа удаление которого увеличивает число компонент связности [13]. Грановеттер в [12] говорит, что сильные связи почти никогда не являются мостами, как правило, мостами являются именно слабые связи.

В нашем случае косвенным подтверждением этого факта являются диаметр и средняя длина пути, которые очень близки в КСС «слабом» веб-графе, КСС веб-графе и коммуникационном графе (табл. 1). Получается, что

удаление сильных связей почти не сказывается на диаметре графов, а значит дуги, соответствующие им, чаще всего не являются мостами.

Отсюда следует вывод о том, что слабые связи фрагмента Веба служат для установления контактов сайтов из непересекающихся групп, что очень важно в смысле получения новой информации, не лежащей в круге интересов данной группы.

Работа выполнена при поддержке Российского фонда фундаментальных исследований, проект 18-07-00628-а.

Литература

1. Thelwall M., Wilkinson D. Graph structure in three national academic Webs: Power laws with anomalies // Journal of the American Society for Information Science and Technology. 2003. №54(8). P. 706-712.
2. Ortega J.L., Aguillo I.F. Visualization of the Nordic academic web: Link analysis using social network tools // Information Processing and Management. 2008. Vol. 44, Iss. 4. P. 1624-1633.
3. Печников А.А. Методы исследования регламентируемых тематических фрагментов Web // Труды Института системного анализа Российской академии наук. Серия: Прикладные проблемы управления макросистемами. 2010. Том 59. С. 134-145.
4. Watts D.J.; Strogatz S.H. Collective dynamics of 'small-world' networks // Nature. № 393. P. 440–442.
5. Ермолин Н.А., Мазалов В.В., Печников А.А. Теоретико-игровые методы нахождения сообществ в академическом Вебе // Труды СПИИРАН. 2017. Вып. 55. С. 237-254.
6. Pechnikov A.A., Chernobrovkin D.I. Adaptive Crawler for External Hyperlinks Search and Acquisition // Automation and Remote Control. 2014, Vol. 75, No. 3. P. 587–593.
7. 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.
8. The Open Graph Viz Platform. – URL: https://gephi.org.
9. Blondel V.D., Guillaume J-L., Lambiotte R., Lefebvre E. Fast unfolding of communities in large networks // Journal of Statistical Mechanics: Theory and Experiment. 2008. P 10008.
10. Newman M.E., Girvan M. Finding and evaluating community structure in networks // Physical Review E. 2004. Vol. 69(2). P 026113.
11. Левкина Л. И. Социально-историческая роль сообществ / Л. И. Левкина. – М.: Русайнс, 2016. 216 с.
12. Granovetter M.S. The Strength of Weak Ties // The American Journal of Sociology. 1973. №78 (6): 1360–1380.

13. Харари Ф. Теория графов. М.: Мир, 1973. 300 с.

References

1. Thelwall M., Wilkinson D. Graph structure in three national academic Webs: Power laws with anomalies // Journal of the American Society for Information Science and Technology. 2003. №54(8). P. 706-712.
2. Ortega J.L., Aguillo I.F. Visualization of the Nordic academic web: Link analysis using social network tools // Information Processing and Management. 2008. Vol. 44, Iss. 4. P. 1624-1633.
3. Pechnikov A.A. Metody issledovanija reglamentiruemyh tematicheskih fragmentov Web // Trudy Instituta sistemnogo analiza Rossiiskoi akademii nauk. Serija: Prikladnye problem upravlenija makrosistemami. 2010. Tom 59. S. 134-145.
4. Watts D.J.; Strogatz S.H. Collective dynamics of 'small-world' networks // Nature. № 393. P. 440–442.
5. Ermolin N.A., Mazalov V.V., Pechnikov A.A. Teoretiko-igrovye metody nahojdenija soobschestv v akademicheskom Webe // Trudy SPIIRAN. 2017. Vyp. 55. S. 237-254.
6. Pechnikov A.A., Chernobrovkin D.I. Adaptive Crawler for External Hyperlinks Search and Acquisition // Automation and Remote Control. 2014, Vol. 75, No. 3. P. 587–593.
7. Golovin A.S., Pechnikov A.A. Baza dannyh vneshnih giperssylok dlja issledovanija fragmentov Weba // Informacionnaja sreda vuza XXI veka: materialy VII Vserossiiskoi nauchno-prakticheskoi konferencii (23-27 sentjabrja 2013). Petrozavodsk, 2013. S. 55-57.
8. The Open Graph Viz Platform. – URL: https://gephi.org.
9. Blondel V.D., Guillaume J-L., Lambiotte R., Lefebvre E. Fast unfolding of communities in large networks // Journal of Statistical Mechanics: Theory and Experiment. 2008. P 10008.
10. Newman M.E., Girvan M. Finding and evaluating community structure in networks // Physical Review E. 2004. Vol. 69(2). P 026113.
11. Levkina L.I. Social’no-istoricheskaja rol’ soobschestv. M.: Rusains, 2016. 216 c.
12. Granovetter M.S. The Strength of Weak Ties // The American Journal of Sociology. 1973. №78 (6): 1360–1380.
13. Harary F. Toerija grafov. M.: Mir, 1973. 300 s.