УДК 519.178: 004.738.5

СООБЩЕСТВА В КОММУНИКАЦИОННОМ ГРАФЕ ТЕМАТИЧЕСКОГО ФРАГМЕНТА ВЕБА (НА ПРИМЕРЕ РАН)

В.В. Мазалов
Институт прикладных математических исследований,
ФИЦ «Карельский научный центр Российской академии наук»
Россия, 185910, Петрозаводск, Пушкинская ул., 11
E-mail: vmazalov@krc.karelia.ru

Н.Н. Никитина
Институт прикладных математических исследований,
ФИЦ «Карельский научный центр Российской академии наук»
Россия, 185910, Петрозаводск, Пушкинская ул., 11
E-mail: nikitina@krc.karelia.ru

А.А. Печников
Институт прикладных математических исследований,
ФИЦ "Карельский научный центр Российской академии наук"
Россия, 185910, Петрозаводск, Пушкинская ул., 11
E-mail: pechnikov@krc.karelia.ru

Ключевые слова: веб-граф, коммуникационный граф, сетевые сообщества, выделение сообществ в сети, метод максимального правдоподобия.

Аннотация: В работе исследуется вопрос о структуре сообществ коммуникационного графа веб-графа на примере сайтов институтов РАН. Для этого используется алгоритм кластеризации графов на основе метода максимального правдоподобия и алгоритм Гирвана-Ньюмана. Приводится сравнение работы данных алгоритмов и дается содержательная интерпретация полученных результатов.

1. Введение

Выделение сообществ в графах является важной задачей во многих прикладных областях: биологии, социологии, социальных сетях, вебометрике, и особенно актуальной для тех сетей, которые представлены графами большой размерности.

Исследования авторов показывают, что вузовские и академические фрагменты Веба России обладают достаточно специфическими свойствами, характеризующими их структуру. В частности, имеется достаточно большая компонента сильной связности и значительное количество «висячих» сайтов, имеющих либо только ссылки, сделанные с них, либо (что реже), ссылки сделанные только на них.

Наша идея заключается в том, чтобы построить некий «жесткий каркас» для такого фрагмента Веба, оставив только те сайты, которые имеют встречные гиперссылки и уже на этом «каркасе» проверить свойства разбиения на сообщества. С помощью из-

вестного алгоритма Гирвана-Ньюмана и авторского алгоритма выделения структуры сообществ на основе метода максимального правдоподобия исследуется вопрос о структуре сообществ коммуникационного графа веб-графа на примере РАН и дается содержательная интерпретация полученных результатов.

2. Основные используемые понятия, методы и алгоритмы

Пусть прямым перечислением задано некоторое целевое множество сайтов и в результате их сканирования найдены 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.

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

Коммуникационный граф веб-графа – это неориентированный граф, имеющий то же самое множество вершин, что и веб-граф, получаемый из веб-графа путём замены встречных дуг на ребра.

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

Для сбора и анализа данных использовались программа для поиска и сбора внешних гиперссылок BeeCrawler [1] и база данных внешних гиперссылок (БД ВГ) учреждений ФАНО на 2017 год (https://fano2017.boincfast.ru, гостевой вход guest/guest) [2].

Для исследования веб-графов применялась открытая платформа для визуализации графов Gephi [3], в которой, в том числе, реализована программа поиска сообществ в ориентированном графе с использованием алгоритма, предложенного в [4]. По сообщению авторов они успешно применяли его для анализа веб-графа с 118 миллионами вершин и более миллиарда дуг [4, стр.1].

Для поиска сообществ в неориентированном графе использовались программы, реализованные в системе Wolfram Mathematica на основе двух различных подходов. В первом случае это известный алгоритм иерархического разбиения Гирвана-Ньюмана [5], реализованный в пакете IGraph/M [6]. Второй подход, описанный в работе двух соавторов настоящей статьи [7], и основанный на использовании метода максимального правдоподобия, реализован авторами в Wolfram Mathematica.

3. Эксперименты с веб-графом РАН

3.1. Веб-граф и коммуникационный граф РАН

По данным БД ВГ был построен веб-граф научных учреждений РАН, вершинами которого являются сайты научных учреждений, связанные соответствующими дугами, содержащий 550 сайтов и 1590 дуг. Здесь нас интересуют связи между научными учреждениями, поэтому «административные» сайты ФАНО, РАН, его отделений и научных центров, а также сайты научных библиотек, и гиперссылки, инцидентные им, не рассматриваются, но для краткости этот веб-граф мы далее будем называть веб-граф РАН.

Каждой вершине был приписан признак научной деятельности в соответствии направлениями научных отделений, к которым относятся научные учреждения [8] (см. таблицу 1). Колонка «количество» будет объяснена позже.

Таблица 1. Классификатор видов научной деятельности.

№	Вид деятельности/научное отделение	признак	количество
1	Сельскохозяйственные науки	*agr*	0
2	Биологические науки	*bio*	15
3	Химия и науки о материалах	*chem*	11
4	Науки о Земле	*earth*	17
5	Энергетика, машиностроение, механика	*energ*	6
6	Глобальные проблемы и международ-	*intern*	4
7	Историко-филологические науки	*ist-fil*	16
8	Математические науки	*math*	3
9	Медицинские науки	*med*	0
10	Нанотехнологии и информационные технологии	*nano*	2
11	Физические науки	*phys*	8
12	Физиологические науки	*physiol*	
13	Общественные науки	*soc*	11

На рис. 1 приводится изображение построенного веб-графа РАН. Максимальная компонента связности, содержащая 434 вершины, выделена более темным цветом. Максимальная компонента сильной связности содержится «внутри» этой компоненты и содержит 203 вершины. По окружности рисунка более светлые вершины являются либо изолированными, либо попарно связными.

Рис. 1. Веб-граф научных учреждений РАН.

Попытки разбиения вершин веб-графа, составляющих максимальную компоненту связности на сообщества по методу [4], приводят к сложно интерпретируемым результатам (в смысле соответствия сообществ видам деятельности). В частности, 90% вершин *agr* попадают в одно сообщество, но также имеется крупное сообщество, содержащее вершины с признаками *agr, bio, chem, earth, energ, math, nano, phys* и *physiol*.

Коммуникационный граф этого веб-графа, построенный на максимальной компоненте сильной связности, изображен на рис. 2. На этом же рисунке приведено разбиение множества вершин на сообщества, которое понадобится далее.

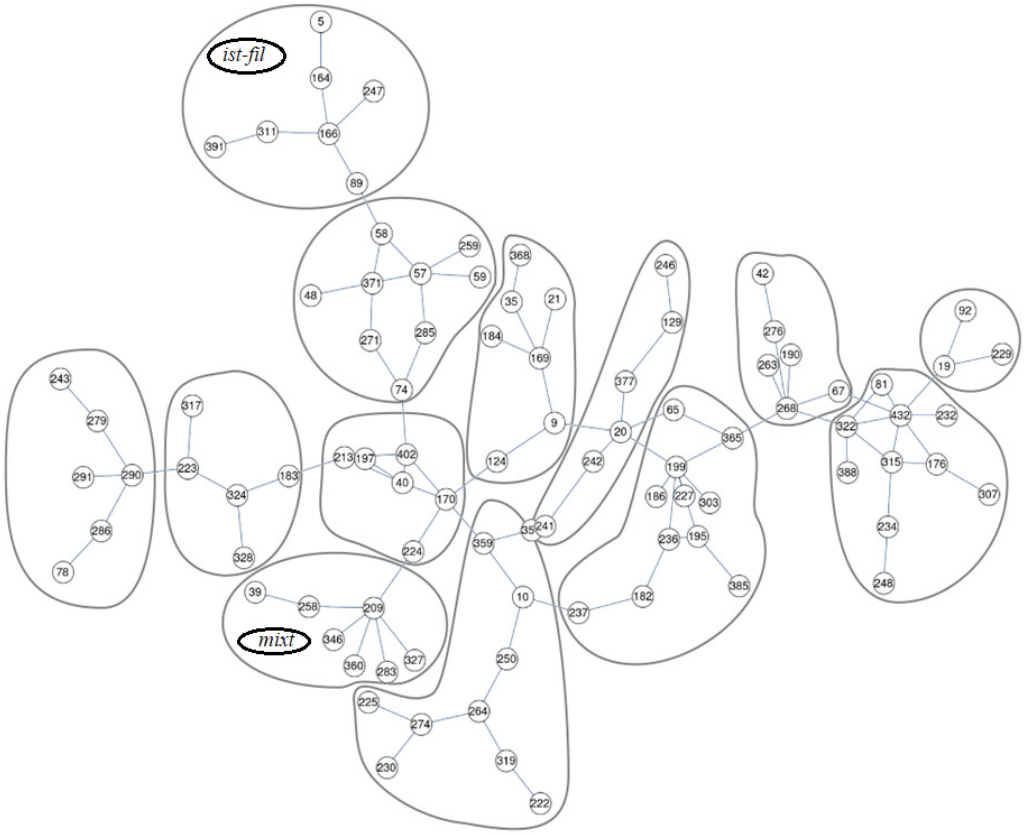

Рис. 2. Коммуникационный граф веб-графа научных учреждений РАН.

Коммуникационный граф веб-графа РАН содержит 93 вершины и 103 ребра, то есть он имеет «почти древовидную» структуру.

3.2. Разбиение коммуникационного графа на сообщества

Применение двух указанных выше программ разбиения множества вершин коммуникационного графа веб-графа РАН при заранее заданном количестве сообществ (по количеству видов научной деятельности) дало очень близкие результаты, причем сообщества во многом, но не в основном соответствуют видам деятельности. Количество вершин коммуникационного графа, имеющих одинаковый признак, дано в последней колонке табл. 1.

Например, алгоритм Гирвана-Ньюмана (разбиение на сообщества приведено на рис. 2) построил 5 «чистых» сообществ: *ist-fil* (7 вершин), *bio* (7), *chem* (6), *earth* (6), *earth* (11). Еще 2 сообщества имеют преобладающее количество вершин с одним и тем же признаком (*ist-fil* и *bio*). Но при этом 3 сообщества имеют существенно разнородную структуру, такую как, например *math* (2)+*phys* (3)+*chem* (1)+*energ* (1).

3.3. Содержательная интерпретация

Дадим ее на примере двух сообществ рис. 2.

Первое из них, помеченное *ist-fil*, содержит вершины, соответствующие официальным сайтам следующих институтов: Институт российской истории РАН, Архив РАН, Институт археологии РАН, Музей антропологии и этнографии им. Петра Великого (Кунсткамера) РАН, Институт истории материальной культуры РАН, Санкт-Петербургский институт истории РАН и Институт истории, археологии и этнографии ДагНЦ РАН. Совершенно очевидно, что сообщество сайтов исторических институтов.

Второе сообщество, помеченное *mixt*, содержит вершины, соответствующие официальным сайтам следующих институтов: Институт неорганической химии им. А.В. Николаева СО РАН, Институт ядерной физики им. Г.И. Будкера СО РАН, Институт лазерной физики СО РАН, Институт вычислительной математики и математической геофизики СО РАН, Институт систем информатики им. А.П. Ершова СО РАН, Институт автоматики и электрометрии СО РАН и Институт теоретической и прикладной механики им. С.А. Христиановича СО РАН. Конечно, можно говорить о принадлежности этих институтов Сибирскому отделению РАН, но хочется верить в то, что взаимодействие сайтов объясняется междисциплинарными исследованиями, проводимыми указанными институтами совместно.

4. Заключение

Разработан подход к нахождению сообществ академического фрагмента Веба, результаты которого имеют адекватную содержательную интерпретацию, что позволяет анализировать текущую структуру Веба, прогнозировать его развитие и делать рекомендации по его совершенствованию как по видам деятельности, так и по междисциплинарным исследованиям.

Работа выполнена при поддержке Российского научного фонда (17-11-01079).

Список литературы

1. 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.
2. 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.
3. Открытая платформа для визуализации графов Gephi. https://gephi.github.io.
4. 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. P10008.
5. Newman M. E., Girvan M. Finding and evaluating community structure in networks // Physical Review E. 2004. Vol. 69 (2). P026113.
6. IGraph/M | Zenodo. https://zenodo.org/record/2349281.
7. Мазалов В.В., Никитина Н.Н. Метод максимального правдоподобия для выделения сообществ в коммуникационных сетях // Вестник Санкт-Петербургского университета. Прикладная математика. Информатика. Процессы управления. 2018. Т. 14, Вып. 3. С. 200-214.
8. Список отделений РАН по направлениям наук. http://www.ras.ru/sciencestructure/departments.aspx.