УДК 519.7:004.738 **DOI** 10.15622/sp.55.10

Н.А. Ермолин, В.В. Мазалов, А.А. Печников

ТЕОРЕТИКО-ИГРОВЫЕ МЕТОДЫ НАХОЖДЕНИЯ СООБЩЕСТВ В АКАДЕМИЧЕСКОМ ВЕБЕ

Ермолин Н.А., Мазалов В.В., Печников А.А. **Теоретико-игровые методы нахождения сообществ в академическом Вебе.**

Аннотация. Исследуется задача нахождения сообществ в графе, представляющем собой фрагмент академического Веба, вершинами которого являются сайты научных организаций, а дугами — гиперссылки. Предлагается новый подход, основанный на методах коалиционной теории игр, применение которого приводит к устойчивому коалиционному разбиению. Для этого определяется функция предпочтения для любой пары вершин в графе, и тогда нахождение стабильного разбиения сводится к нахождению максимума потенциальной функции. Описан реализованный алгоритм поиска стабильного разбиения, даны оценки его сложности. Делается сравнение предлагаемого метода с двумя известными методами нахождения сообществ, в том числе эффективность нового метода показывается на разбиении на сообщества фрагмента Веба, состоящего из официальных сайтов Сибирского и Дальневосточного отделений РАН.

Ключевые слова: веб-пространство, граф, сообщество, модулярность, коалиционная теория игр.

1. Введение. Исследования Веба относятся к актуальным разделам такого направления компьютерных наук, как вебометрика [1], в последнее время все чаще называемого «наукой о Вебе» [2]. В России это направление активно развивается в Новосибирске и Санкт-Петербурге и представлено в работах [3-7]. Методы изучения принципов самоорганизации и связей веб-пространств, апробированные на примерах таких крупных организаций, как Сибирское отделение РАН и Санкт-Петербургский университет, могут быть использованы для анализа структурно-коммуникативной организации целого ряда профессиональных веб-сообществ.

Вследствие гигантской размерности Веба во многих случаях исследования проводятся на его достаточно узких фрагментах, таких как множества сайтов университетов Великобритании, научных учреждений России и так далее. Веб-пространство — это множество веб-сайтов, связанных посредством гиперссылок. В данной статье речь будет идти о веб-пространстве официальных сайтов научных организаций России, относящихся к Сибирскому и Дальневосточному отделениям РАН.

Математической моделью, успешно используемой для анализа веб-пространств, является веб-граф [8], в нашем случае построенный следующим образом: множество вершин соответствует сайтам организаций, а множество дуг — гиперссылкам, связывающим эти сайты. Такой граф является ориентированным графом без петель, имеющим кратные дуги.

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 [9]. Модулярность, в свою очередь, это свойство графа и некоторого разбиения его на подграфы. Для обозначения подграфов, на которые разбивается граф, используются различные термины, такие как кластер, модуль, сообщество [10, 11]. Применительно к Вебу более распространенным является использование термина «сообщество», и далее мы будем использовать его.

Мера модулярности показывает, насколько данное разбиение качественно в том смысле, что существует много дуг, лежащих внутри сообществ, и мало дуг, лежащих вне сообществ, но соединяющих их между собой.

Определение функции модулярности Q в терминологии случайных графов выглядит следующим образом [12]:

– пусть *G(V,E)* — граф с множеством вершин *V* и множеством дуг *E*;

– *A* — матрица инцидентности графа *G(V,E)*;

– A_{ij} — количество дуг из вершины *i* в вершину *j*;

– *m* — количество дуг в графе, $m = |E|$;

– *Pr*(.) — вероятность некоторого события;

– *S* — некоторое множество сообществ, на которые разбит граф *G(V,E)*;

– *s* — обозначение одного из модулей $s \in S$.

Тогда выражение для модулярности примет вид:

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

Модулярность часто используется для определения качества разбиения графа на сообщества, однако в данной работе такая задача не является основной.

Далее мы рассмотрим три метода нахождения сообществ, основанные на исследованиях, опубликованных в работах [13-15]; два первых — кратко с отсылкой к литературе, а третий более подробно. Будем называть эти методы по первым буквам фамилий авторов NG, BGLL и AKM соответственно.

Для методов NG и AKM в рамках работы были написаны реализующие их программы, а метод BGLL реализован в программном пакете Gephi [16], который и использовался в нашем случае.

Далее, мы сформировали фрагмент веб-пространства научных организаций России, содержащий организации, географически и админи-

стративно принадлежащие к двум разным отделениям РАН, построили для него веб-граф, нашли разбиения полученного веб-графа всеми тремя методами на заданное число сообществ, и сравнили их между собой. В первом случае мы проверили, насколько хорошо разбивается веб-граф на два сообщества в соответствии с известной нам принадлежностью к двум разным отделениям. Во втором случае сравниваются разбиения тремя методами на 5 сообществ, поскольку для метода BGLL именно разбиение на 5 сообществ дает максимальное значение функции модулярности (1).

Проведенные эксперименты позволяют сделать вывод о перспективности метода нахождения разбиений графа, основанного на принципиально новой идее коалиционных игр.

2. Методы NG и BGLL. Метод NG опишем, следуя [13]. Здесь рассматривается неориентированный граф, который получается из ориентированного графа следующим образом: сначала удаляются кратные дуги, а затем оставшиеся дуги заменяются на ребра.

Основная идея заключается в том, что наибольшей центральностью обычно обладают ребра, соединяющие разные сообщества, поэтому их надо удалять из графа в первую очередь. Для нахождения центральности ребер используется такой показатель центральности, как betweenness centrality (в русскоязычных публикациях используется также термин «центральность по посредничеству»). Центральность по посредничеству показывает, сколько кратчайших путей между всеми вершинами графа проходит через определенное ребро. После удаления нескольких ребер значения центральности ребер существенно меняются, поэтому их нужно пересчитывать.

Кратко опишем алгоритм:

1. Находим значения центральности всех ребер.
2. Удаляем ребро с максимальной центральностью.
3. Пересчитываем центральности ребер и переходим к шагу 2.

В какой-то момент прерывая этот процесс, получаем разбиение на сообщества, считая вершины каждой полученной компоненты связности отдельным сообществом.

Поскольку после каждого удаленного ребра проверка, не увеличилось ли число компонент связанности в графе, достаточно затратная, можно действовать по-другому, а именно следующим образом:

1. Сформируем список ребер в порядке удаления.
2. Поместим каждую вершину в отдельное сообщество.
3. Пройдем по списку удаленных ребер в обратном порядке, объединяя сообщества, которым принадлежат вершины очередного ребра.

Такая схема позволяет один раз сформировать список удаленных ребер (а это очень затратная процедура из-за необходимости

пересчета центральностей), а потом экспериментировать с количеством сообществ.

Метод BGLL опишем, следуя [14]. Идея алгоритма основана на свойстве самоподобия сложных сетей [17] и естественным образом включает понятие иерархии, так как в процессе его выполнения строятся сообщества сообществ.

Каждый проход алгоритма состоит из двух фаз, повторяющихся многократно.

Первый проход начинается с первой фазы, когда генерируется столько же сообществ, сколько вершин в графе, и каждое сообщество содержит по одной вершине.

Затем для каждой вершины i рассматриваются все вершины j, смежные с i, и оценивается прирост модулярности в том случае, если удалить i из своего сообщества и добавить его в сообщество j.

Вершина i добавляется в сообщество, для которого этот прирост является максимальным, но только если этот прирост будет положительным (в случае нескольких одинаковых приростов, например, берется последний из них). Если никаких положительных приростов нет, i остается в своем первоначальном сообществе.

Эта процедура применяется многократно и последовательно для всех вершин до тех пор, пока не будет достигнуто никакого дальнейшего улучшения приростов, и тогда первая фаза завершается.

Вторая фаза алгоритма заключается в создании нового графа, вершинами которого являются сообщества, найденные в результате первой фазы. Для этого ребрам между новыми вершинами присваиваются веса, которые вычисляются как сумма весов ребер между вершинами, принадлежащими двум разным сообществам (можно считать, что одиночное ребро изначально имеет вес 1). Ребра между вершинами одного и того же сообщества приводят к петлям для вершины, соответствующей этому сообществу в новом графе.

Как только вторая фаза будет завершена, можно будет выполнить следующий проход, начиная с применения первой фазы к полученной взвешенной сети.

Проходы повторяются итеративно до тех пор, пока значение модулярности растет, то есть до достижения максимума модулярности.

3. Нахождение сообществ как коалиционная игра. В работе [15] предложен теоретико-игровой подход для выделения сообществ в графе, основанный на методах коалиционных гедонических игр [18]. Вершины в графе интерпретируются как игроки, отношения между которыми определяются с помощью функции предпочтения. С их помощью можно построить устойчивые коалиционные разбиения.

Предположим, что множество игроков $N = \{1, 2, ..., n\}$ разбито на K коалиций $\Pi = \{S_1, ..., S_K\}$. Пусть $S_\Pi(i)$ обозначает коалицию $S_k \in \Pi$, где $i \in S_k$. Предпочтения игрока i представлены бинарным отношением $\succeq_i$ (рефлексивным и транзитивным) на множестве $\{S \subset N : i \in S\}$. Предпочтения являются аддитивно сепарабельными [18], если существует такая функция $v_i : N \to \mathbf{R}$, что $v_i = 0$ и:

$$S_1 \succeq_i S_2 \Leftrightarrow \sum_{j \in S_1} v_i(j) \geq \sum_{j \in S_2} v_i(j).$$

Предпочтения $\{v_i, i \in N\}$ являются симметричными, то есть $v_i(j) = v_j(i) = v_{ij} = v_{ji}$ для всех $i, j \in N$. Свойство симметрии является важным в теории гедонических игр.

Будем говорить, что коалиционное разбиение Π является Нэш-стабильным, если $S_\Pi(i) \succeq_i S_k \bigcup \{i\}$ для всех $i \in N, S_k \in \Pi \bigcup \{0\}$. В Нэш-стабильном разбиении никому из игроков невыгодно переходить в другую коалицию.

Потенциалом коалиционного разбиения $\Pi = \{S_1, ..., S_K\}$ называется функция:

$$P(\Pi) = \sum_{k=1}^{K} P(S_k) = \sum_{k=1}^{K} \sum_{i,j \in S_k} v_{i,j}.$$

Разбиение $\Pi = \{S_1, ..., S_K\}$, дающее максимум потенциала, является Нэш-стабильным. Для поиска стабильных разбиений можно воспользоваться следующей процедурой.

Начинаем с какого-то фиксированного разбиения $N = \{S_1, ..., S_K\}$. Выберем игрока i и произвольную коалицию S_k, отличную от $S_\Pi(i)$. Если $S_k \bigcup \{i\} \succeq_i S_\Pi(i)$, переместим вершину i в коалицию S_k; иначе оставляем разбиение неизменным и выбираем другую пару кандидатов, и так далее. Так как число вершин конечно, алгоритм закончится за конечное число шагов с каким-то локальным максимумом потенциала.

В работе [15] предложено использовать функцию предпочтения с параметром $\alpha \in [0,1]$ следующего вида:

$$v_{ij} = \begin{cases} 1 - \alpha, (i, j) \in E, \\ -\alpha, (i, j) \notin E, \\ 0, i = j. \end{cases}$$

Для любого подграфа $(S, E \mid S), S \subseteq N$, обозначим *n(S)* число вершин в *S*, и *m(S)* число ребер в *S*. Тогда потенциал можно представить как:

$$P(\Pi) = \sum_{k=1}^{K} \left(m(S_k) - \frac{n(S_k)(n(S_k)-1)\alpha}{2} \right). \quad (2)$$

Для потенциала такого вида справедливо следующее утверждение. *Теорема. Если* $\alpha = 0$, *то гранд-коалиция* $\Pi_N = \{N\}$ *дает максимум потенциала. При* $\alpha \to 1$ *максимум потенциала достигается на разбиении графа на максимальные клики.*

Также эффективным при разбиении на сообщества является использование при определении потенциала понятия модулярности, введенное в [13]. Для конфигурационной модели случайного графа формула модулярности (1) примет вид:

$$P(\Pi) = \sum_{k=1}^{K} \sum_{i,j \in S_k, i \neq j} \left(A_{ij} - \gamma \frac{d_i d_j}{2m} \right). \quad (3)$$

где A_{ij} — количество ребер между вершинами i и j, d_i — степень вершины i, m — общее число ребер в графе, γ — параметр, $0 < \gamma < 1$.

В качестве примера рассмотрим граф, изображенный на рисунке 1.

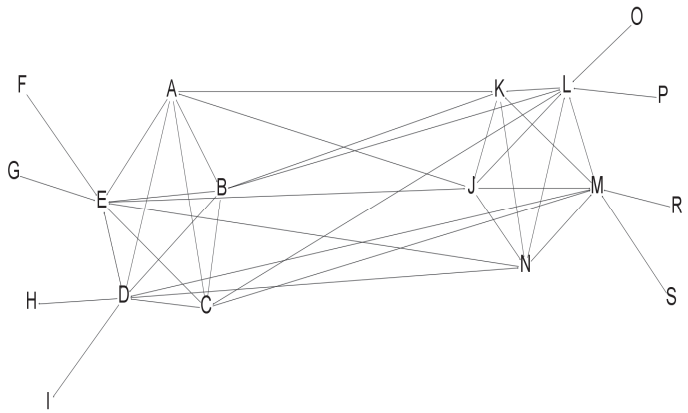

Рис. 1. Пример разбиения графа на сообщества

Граф состоит из двух полных подграфов: G_1={*A, B, C, D, E*} и G_2={*J, K, L, M, N*}. Помимо этого в графе есть 10 ребер между G_1 и G_2,

то есть таких (u,w), что $u \in G_1$ и $w \in G_2$. Наконец, есть еще 4 вершины, смежные с вершинами из G_1, и 4 вершины, смежные с G_2.

Наиболее естественным разбиением этого графа на два сообщества представляется $\Pi_N = S_L \bigcup S_R = \{A, B, ..., H, I\} \bigcup \{J, K, ..., R, S\}$, где первое сообщество — это вершины в левой части рисунка, а второе — вершины в правой части.

Если считать потенциал по формуле (3), то имеем: $P(S_L \bigcup S_R) = 28 - \gamma \cdot 31\frac{13}{19}$. Если перевести вершину J в другое сообщество, то получим $P(\{S_L \setminus J\} \bigcup \{S_R \bigcup J\}) = 26 - \gamma \cdot 32\frac{12}{19}$.

Важно, что неравенство $P(\{S_L \bigcup S_R) > P(\{S_L \setminus J\} \bigcup \{S_R \bigcup J\})$ верно для любых значений $0 < \gamma < 1$. Другие разбиения на сообщества, которые можно получить из $S_L \bigcup S_R$ путем перемещения только одной вершины, дают еще меньшие значения потенциала. То есть на $S_L \bigcup S_R$ достигается локальный максимум потенциала, а значит, мы имеем устойчивое по Нэшу разбиение.

Вычислим потенциал согласно (2): $P(S_L \bigcup S_R) = 28 - 72\alpha$. Если переместить вершину J в другое сообщество, то получим $P(\{S_L \setminus J\} \bigcup \{S_R \bigcup J\}) = 26 - 73\alpha$, то есть первое разбиение всегда предпочтительнее в смысле (2).

Для этого примера путем полного перебора вариантов несложно показать, что $P(S_L \bigcup S_R) > P(\Pi')$, где Π' — разбиение, получаемое из $S_L \cup S_R$ перемещением одной вершины, верно для всех Π' при любых значениях α. Значит, $S_L \bigcup S_R$ — стабильное по Нэшу разбиение в игре с потенциалом (2).

Но существуют и другие разбиения. Рассмотрим:

$$\Pi^* = \{A, B, C, D, E, J, K, L, M, N\} \bigcup$$
$$\bigcup \{F\} \bigcup \{G\} \bigcup \{H\} \bigcup \{I\} \bigcup \{O\} \bigcup \{P\} \bigcup \{R\} \bigcup \{S\}.$$

и

$$\Pi^{**} = \{A, B, C, D, E\} \bigcup \{J, K, L, M, N\} \bigcup$$
$$\bigcup \{F\} \bigcup \{G\} \bigcup \{H\} \bigcup \{I\} \bigcup \{O\} \bigcup \{P\} \bigcup \{R\} \bigcup \{S\}.$$

По формуле (2) имеем потенциал $P(\Pi^*) = 30 - 45\alpha > P(S_L \bigcup S_R)$ для любых α. С другой стороны, $P(\Pi^{**}) = 20 - 20\alpha > P(\Pi^*)$ при

$\alpha > \frac{2}{5}$. То есть при больших α разбиение на максимальные клики дает Нэш-стабильное разбиение. Это соответствует утверждению приведенной выше теоремы.

Если использовать (3), то $P(\Pi^*) = 30 - \gamma \cdot 42\frac{12}{19}$ и $P(\Pi^*) > P(S_L \bigcup S_R)$ выполняется при $\gamma < \frac{19}{108}$. Но при таких γ разбиение Π^* не является Нэш-стабильным, и алгоритм поиска устойчивого разбиения объединяет все вершины в одно сообщество $\Pi^* = \{A, B, ..., R, S\}$.

4. Алгоритм поиска Нэш-стабильного разбиения. Опишем алгоритм, который позволяет найти стабильное разбиение для функции потенциала (2) и фиксированного значения параметра α. На вход алгоритм получает произвольное стартовое разбиение Π_0.

На каждой итерации алгоритма мы будем рассматривать текущее разбиение Π и каждое разбиение Π_i' из тех, что получается переводом игрока $i \in S_{\Pi(i)}$ в коалицию $S_k \in \Pi$. Среди них будем выбирать такое разбиение Π', что $P(\Pi') - P(\Pi) \to \max$. Если $P(\Pi') - P(\Pi) > 0$, то текущим разбиением становится Π', иначе текущее разбиение Π и есть искомое стабильное разбиение.

Рассмотрим вначале потенциал в виде (2). Предположим, в текущем разбиении Π вершина *i* переводится из коалиции $S_{\Pi(i)}$ в какую-то коалицию S_k. В коалиции S_k игрок *i* приобретает $d_i(S_k \bigcup i)$ связей с весом $1-\alpha$ минус $m(S_k)+1-d_i(S_k \bigcup i)$ связей с весом α. Здесь $d_i(S)$ обозначает степень вершины *i* в графе на множестве вершин S.

При этом игрок *i* теряет в старой коалиции $d_i(S_{\Pi(i)})$ связей с весом $1-\alpha$ минус $m(S_{\Pi(i)}) - d_i(S_{\Pi(i)})$ связей с весом α. Таким образом, изменение потенциала равно:

$$P(\Pi') - P(\Pi) = d_i(S_k \bigcup i) - d_i(S_{\Pi(i)}) + \alpha(m(S_{\Pi(i)}) - m(S_k) - 1).$$

Если же игрок *i* станет индивидуальным, изменение потенциала станет равным:

$$P(\Pi') - P(\Pi) = -d_i(S_{\Pi(i)}) + \alpha(m(S_{\Pi(i)}) - 1).$$

Найдем изменение потенциала в виде (3).

Предположим, в текущем разбиении Π вершина i переходит из коалиции $S_{\Pi(i)}$ в какую-то коалицию S_k. В коалиции S_k потенциал (3) прирастает на $d_i(S_k \bigcup i)$ связей минус $\frac{\gamma \cdot d_i}{2m} \sum_{j \in S_k} d_j$. При этом игрок i теряет в старой коалиции величину $d_i(S_{\Pi(i)})$ связей минус $\frac{\gamma \cdot d_i}{2m} \sum_{j \in S_k \setminus i} d_j$.

$$P(\Pi') - P(\Pi) = -d_i(S_{\Pi(i)}) + d_i(S_k \bigcup i) + \frac{\gamma \cdot d_i}{2m} \left(\sum_{j \in S_{\Pi(i)}} d_j - \sum_{j \in S_k} d_j \right) - \frac{\gamma \cdot d_i^2}{2m}.$$

Теперь вернемся к тому, как реализовать структуру данных для хранения сообщества. Наша структура хранения сообщества содержит следующие поля:

1. Множество вершин. Множество целых чисел — номеров вершин в сообществе. Пусть n — количество вершин в сообществе. Тогда операции добавления, удаления или проверки наличия номера имеют временную сложность $O(log_2 n)$, а получить все номера в виде списка можно за $O(n)$. Множество с таким набором операций можно реализовать или использовать готовую реализацию на основе сбалансированного дерева поиска.

2. Список количеств общих ребер. Пусть N — количество вершин во всем графе. В этом списке длины N в i-ом элементе будем хранить кэшированное количество ребер, один конец которых — вершина i, а второй конец принадлежит сообществу.

3. Счетчик времени и массив временных отметок. После добавления или удаления вершины из сообщества, кэшированные значения количества общих ребер становятся неактуальными. Поскольку пересчитывать их все сразу очень долго, отмечать каким-то образом, что они не актуальны тоже долго ($O(N)$), то мы будем использовать счетчик времени и массив временных отметок (кэшированное значение актуально только тогда, когда его временная отметка совпадает со счетчиком времени). Отметим, что счетчик времени у каждого экземпляра структуры свой, не зависящий от других экземпляров.

4. Количество вершин и количество ребер в сообществе, сумма степеней вершин и квадратов степеней. Эти поля нам нужны для расчета потенциала сообщества.

Прокомментируем, как производить наименее тривиальные операции.

1. Создание пустого сообщества. Нам нужно выделить $O(N)$ памяти на хранение структуры (считаем, что это можно сделать за $O(1)$) и инициализировать $O(1)$ полей.

2. Удаление и добавление вершины в сообщество. При удалении или добавлении вершины нужно делать следующее:

(a) Изменить значения количества вершин в сообществе, суммы степеней и квадратов степеней вершин.

(b) Пересчитать число ребер в сообществе. Для этого нужно перебрать все ребра, выходящие из добавляемой или удаляемой вершины, и посчитать у скольких из них второй конец принадлежит множеству вершин сообщества. Именно эта операция имеет сложность $O(deg(v)log_2 n)$.

(c) Увеличить счетчик времени, чтобы кэшированное количество общих ребер для всех вершин стало неактуальным.

3. Получить количество ребер, инцидентных данной вершине, которые ведут в сообщество. Если временная отметка для данного номера вершины совпадает со счетчиком времени, то вернуть кэшированное значение. Иначе пересчитать и запомнить его за $O(deg(v)log_2 n)$.

5. Численные эксперименты. В качестве примера было рассмотрено множество сайтов научных учреждений, входящих в состав Сибирского и Дальневосточного отделений Российской академии наук (далее — СО РАН и ДВО РАН) в их дореформенной версии. Общее количество сайтов научных учреждений (научных отделений, центров, институтов, библиотек, проектов) равно 140, из них 102 относятся к учреждениям СО РАН и 38 — ДВО РАН. Сайты связаны между собой 2315 гиперссылками. Данные о гиперссылках были взяты из базы данных внешних гиперссылок [19]. Сканирование веб-сайтов выполнялось с мая 2013 по март 2014 года с использованием краулера для сбора внешних гиперссылок BeeBot [20].

Часть сайтов научных учреждений СО РАН и ДВО РАН приведена в таблице 1. Сайты упорядочены по их доменным именам, в колонке *branch* стоит принадлежность соответствующему отделению.

Таблица 1. Множество исследуемых сайтов научных учреждений

№	URL	branch	Название учреждения
1	*509.tig.dvo.ru*	dvo	Технический сайт ТИГ ДВО
2	*tig.dvo.ru*	dvo	Тихоокеанский институт географии ДВО РАН
3	*alley.iis.nsk.su*	so	Аллея памяти ИСИ СО РАН
4	*www.iis.nsk.su*	so	Институт систем информатики имени А.П. Ершова СО РАН
5	*bionano.niboch.nsc.ru*	so	Научная конференция посвященная 25-летнему юбилею ИХБиФМ СО РАН
6	*www.niboch.nsc.ru*	so	Институт химической биологии и фундаментальной медицины СО РАН
7	*ccu.kirensky.ru*	so	ЦКП КНЦ СО РАН
8	*kirensky.ru*	so	Институт физики им. Л.В. Киренского СО РАН
9	*www-sbras.nsc.ru*	so	Сибирское отделение РАН
..	....	…	…..
12	***crust.irk.ru***	**so**	**Институт земной коры СО РАН**
13	***sifibr.irk.ru***	**so**	**Сибирский институт физиологии и биохимии растений СО РАН**
14	*www.gs.nsc.ru*	so	Геофизическая служба СО РАН
15	*www.igc.irk.ru*	so	Институт геохимии им. А.П. Виноградова СО РАН
16	***www.irigs.irk.ru***	**so**	**Институт географии им. В.Б.Сочавы СО РАН**
17	***www.isc.irk.ru***	**so**	**Иркутский научный центр СО РАН**
18	*www.lin.irk.ru*	so	Лимнологический институт СО РАН
19	***www.sei.irk.ru***	**so**	**Институт систем энергетики им. Л.А. Мелентьева СО РАН**
20	*ecrin.ru*	dvo	Институт экономических исследований ДВО РАН
21	*forest.akadem.ru*	so	Институт леса им. В.Н. Сукачева СО РАН
22	*icarp.ru*	dvo	Институт комплексного анализа региональных проблем ДВО РАН
23	*ivep.as.khb.ru*	dvo	Институт водных и экологических проблем ДВО РАН
24	*www.febras.ru*	dvo	Дальневосточное отделение РАН
..	...	…	…..
46	***imbt.ru***	**so**	**Институт монголоведения, буддологии и тибетологии СО РАН**
47	***old.imbt.ru***	**so**	**Институт монголоведения, буддологии и тибетологии СО РАН (старый сайт)**
..	...	…	…..
128	***bibl.history.nsc.ru***	**so**	**Библиотека Института истории СО РАН**
129	***www.icc.irk.ru***	**so**	**Институт динамики систем и теории управления СО РАН**
130	*www.icms.kemsc.ru*	so	Институт углехимии и химического материаловедения СО РАН
131	***baikalmuseum.irk.ru***	**so**	**Байкальский музей Иркутского научного центра СО РАН**
..	...	…	…..
136	***med.isc.irk.ru***	**so**	**Больница Иркутского НЦ СО РАН**
..	...	…	…..
140	*gallery.hcei.tsc.ru*	so	Gallery ИСЭ СО РАН

Веб-граф G_{so+dvo} представляет собой, в зависимости от используемого метода, ориентированный или неориентированный граф без кратных дуг и петель, дуга в графе появляется в том случае, когда есть хотя бы одна гиперссылка, связывающая соответствующие сайты-вершины.

Вследствие удаления кратных дуг количество дуг в исследуемом графе равно 633, а количество вершин — 140. Неориентированный граф получается из ориентированного графа заменой дуг на ребра.

Из большого количества проведенных экспериментов приведем только результаты наиболее характерных разбиений на 2 и 5 сообществ. Как уже упоминалось ранее, методы, реализация в виде алгоритмов и полученные результаты будем обозначать с использованием первых букв фамилий авторов NG, BGLL и AKM.

Для метода AKM были реализованы два алгоритма, использующих формулы (2) и (3) для вычисления потенциалов, однако в статье оставлены только результаты по формуле (3), поскольку оба метода дают практически одинаковые разбиения.

Разбиение G_{so+dvo} на два сообщества методами NG, BGLL и AKM дает одинаковый результат: в первое сообщество входят все вершины, соответствующие сайтам СО РАН, а во второе — ДВО РАН.

Результаты разбиений G_{so+dvo} на пять сообществ сведены в таблицу 2, где *C1*-*C5* используются для обозначения сообществ.

Таблица 2. Основные результаты экспериментов

Метод	*C1*	*C2*	*C3*	*C4*	*C5*
NG	98	38	2	1	1
AKM	91	38	7	2	2
BGLL	46	38	21	18	10

В позиции на пересечении названия метода и сообщества указывается мощность данного сообщества. Сообщества упорядочены по мощности слева направо.

Заметим, что для всех трех методов полученные сообщества *C2* содержат по 38 вершин, в точности соответствующих веб-сайтам ДВО РАН. Естественно, сообщества *C1*, *C3*, *C4* и *C5* для всех трех случаев содержат сайты СО РАН, имея разные мощности в зависимости от используемого метода.

Начнем с последнего. Метод BGLL дает разбиения на сообщества, достаточно большие по количеству участников. Так *C1* содержит 46 сайтов, включая официальный сайт Сибирского отделения РАН и сайты 7 научных центров (из 8), больше ориентированных на административные связи. Сообщество *C3* можно назвать «физико-химическим», а *C4* — «компьютерно-моделирующим» по входящим в

них сайтам научных учреждений. Сообщество *C5* содержит 10 (из 12) сайтов Иркутского научного центра.

На рисунке 2 приведен пример разбиения на сообщества по методу АКМ. Сообщество в *C2* правом эллипсе целиком состоит из сайтов ДВО РАН. Внутри эллипса, названного СО РАН, содержатся 4 сообщества: *C1* — это 91 сайт, включая официальный сайт Сибирского отделения РАН.

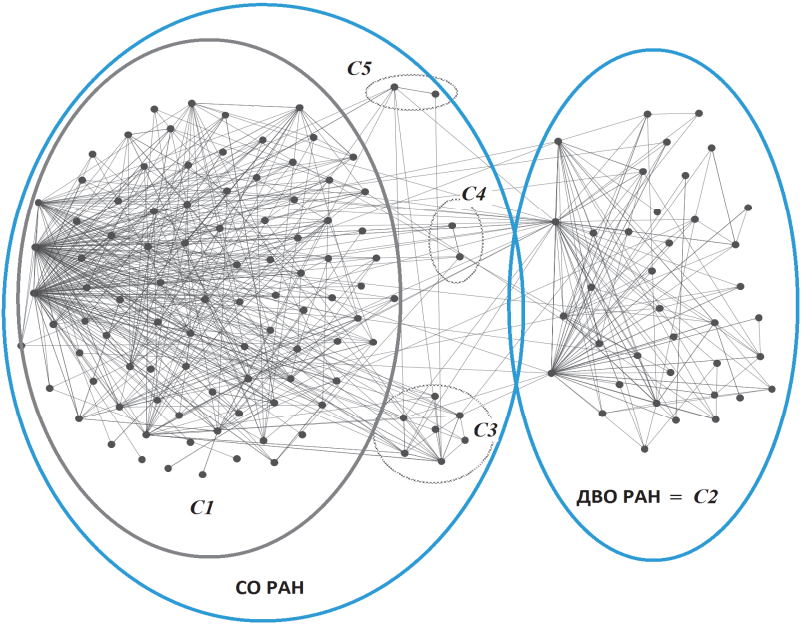

Рис. 2. Пример разбиения графа G_{so+dvo} на сообщества по методу АКМ

Сообщество *C3* представляет собой 7 сайтов организаций, входящих в состав Иркутского научного центра СО РАН:

– Института земной коры СО РАН;
– Сибирского института физиологии и биохимии растений СО РАН;
– Института географии им. В.Б. Сочавы СО РАН;
– Института систем энергетики им. Л.А. Мелентьева СО РАН;
– Института динамики систем и теории управления СО РАН;
– самого Иркутского научного центра СО РАН;
– и даже Больницы Иркутского научного центра СО РАН.

В таблице 1 данные об этих сайтах выделены полужирным шрифтом.

В состав *C4* входят два сайта: Лимнологического института СО РАН и «Информационные ресурсы Лимнологического института СО РАН».

В состав *C5* также входят два сайта: Омского научного центра СО РАН и Института проблем переработки углеводородов СО РАН, входящего в состав этого же центра.

На рисунке 3 приведен пример разбиения на сообщества по методу NG. Здесь так же, как и в случае АКМ, характерно наличие большого сообщества *C1*, содержащего 98 сайтов, включая официальный сайт Сибирского отделения РАН.

Как и в предыдущем примере сообщество *C2* целиком состоит из сайтов ДВО РАН. Кроме того, метод NG формирует еще три небольших сообщества, причем два из них вырожденные и содержащие по одному сайту.

Сообщество *C3* состоит из двух сайтов Института монголоведения, буддологии и тибетологии СО РАН (нового и старого). Сообщества *C4* и *C5* и вовсе состоят из одиночных сайтов Байкальского музея Иркутского научного центра СО РАН и Библиотеки Института истории СО РАН. В таблице 1 данные об этих сайтах также выделены полужирным шрифтом.

Заметим, однако, что данное разбиение не является Нэш-стабильным.

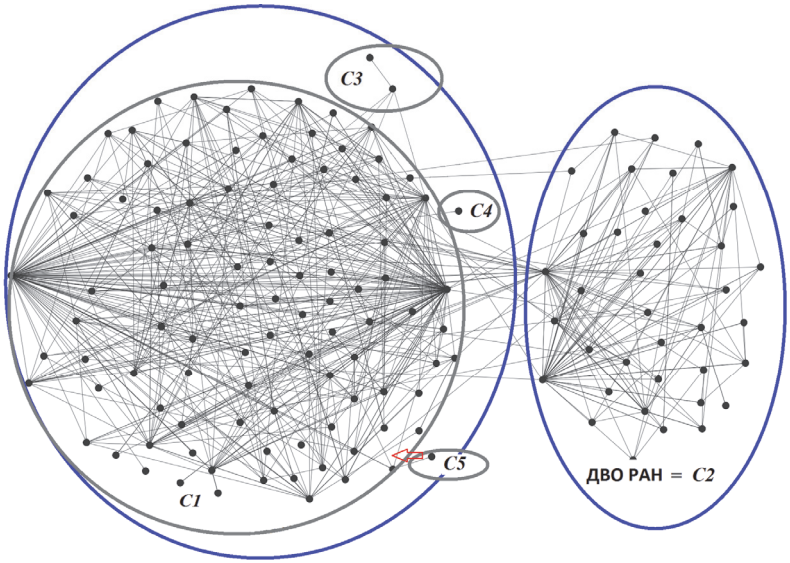

Рис. 3. Пример разбиения графа G_{so+dvo} на сообщества по методу NG

Действительно, потенциал такого разбиения равен $1190-\gamma\cdot\frac{472962}{633}$. Но если переместить, например, вершину, соответствующую сайту Библиотеки Института истории СО РАН (*C5*), в сообщество СО РАН (*C1*), — это продемонстрировано стрелкой на рисунке 3, — то значение потенциала увеличится и станет равным $1192-\gamma\cdot\frac{473903}{633}$, что превосходит предыдущее значение при $\gamma<\frac{633}{469}$, то есть при любых $0<\gamma<1$. То же самое относится и к разбиению на сообщества по методу BGLL.

6. Заключение. В работе предложен метод нахождения разбиений графа, основанный на принципиально новой идее коалиционных игр, который дает результаты, не уступающие по содержательной интерпретации двум известным и широко распространенным методам. Для этого определяется специальным образом коалиционная игра, в которой вершины графа являются игроками, и предпочтения игроков определяются бинарным отношением на множестве коалиций. После этого находится стабильное коалиционное разбиение, в котором никому из игроков не выгодно менять свою коалицию.

Этот метод реализован в виде алгоритма, который апробирован при кластеризации множества сайтов научных организаций Сибирского и Дальневосточного отделений РАН. Приведены результаты численных экспериментов на данном фрагменте Веба. При анализе веб-пространств организаций можно рекомендовать использование всех трех методов с возможностью получения окончательного результата как комбинации трех частных результатов на основе их содержательного анализа.

Литература

1. *Thelwall M.* Webometrics and Social Web Research Methods // University of Wolverhampton. 2013. 140 p.
2. *Hall W., Tiropanis T.* Web evolution and Web Science // Computer Networks. 2012. vol. 56, no. 18. pp. 3859–3865.
3. *Шокин Ю.И. и др.* Анализ веб-пространства академических сообществ методами вебометрики и теории графов // Информационные технологии. 2014. № 12. С. 31–40.
4. *Шокин Ю.И. и др.* Исследование научного веб-пространства Сибирского отделения Российской академии наук // Вычислительные технологии. 2012. Т. 17. № 6. С. 86–98.
5. *Веснин А.Ю., Константинова Е.В., Савин М.Ю.* О сценариях присоединения новых сайтов к веб-пространству СО РАН // Вестник Новосибирского государственного университета. Серия: Информационные технологии. 2013. Т. 11. № 4. С. 28–37.
6. *Клименко О.А.* Модели представления академического веб-пространства // Информационные и математические технологии в науке и управлении. 2016. № 2. С. 103–110.
7. *Корелин В.Н., Блеканов И.С., Сергеев С.Л.* Применение модифицированного алгоритма LSH для кластеризации внешнего окружения веб-пространства университетов // Научно-технические ведомости Санкт-Петербургского государственного политехнического университета. Информатика. Телекоммуникации. Управление. 2015. № 5 (229). С. 79–87.
8. *Bonato A., Graham F. C., Pralat P.* Algorithms and Models for the Web Graph // Proceedings of the 13th International Workshop (WAW 2016). 2016. vol. 10088. 165 p.

9. *Flake G.W., Lawrence S.R., Giles C.L., Coetzee F.M.* Self-Organization and Identification of Web Communities // IEEE Computer. 2002. vol. 35. no. 3. pp. 66–71.
10. *Labatut V., Balasque J.-M.* Detection and Interpretation of Communities in Complex Networks: Practical Methods and Application // Computational Social Networks. 2012. pp. 81–113.
11. *Avrachenkov K., El Chamie M., Neglia G.* Graph clustering based on mixing time of random walks // Proceedings of IEEE ICC 2014. 2014. pp. 4089–4094.
12. *Newman M.E.J.* Finding community structure in networks using the eigenvectors of matrices // Phys. Rev. 2006. vol. 74. no. 3. pp. 036104.
13. *Girvan M., Newman M.E.J.* Community structure in social and biological networks // Proc. of National Academy of Science. 2002. vol. 99(12). pp. 7821–7826.
14. *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. pp. P10008.
15. *Avrachenkov K.E., Kondratev A.Yu., Mazalov V.V.* Cooperative Game Theory Approaches for Network Partitioning // International Computing and Combinatorics Conference. 2017. LNCS 10392. pp. 591–602.
16. Gephi – The Open GraphViz Platform. URL: www.gephi.org (дата обращения: 09.06.2017).
17. *DeDeo S., Krakauer D.* Dynamics and Processing in Finite Self-Similar Networks // Journal of the Royal Society Interface. 2012. vol. 9. no. 74. pp. 2131–2144.
18. *Bogomolnaia A., Jackson M.O.* The stability of hedonic coalition structures // Games and Economic Behavior. 2002. vol. 38. no. 2. pp. 201–230.
19. *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. Петрозаводск. 2013. С. 55–57.
20. *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.

Ермолин Николай Александрович — студент, Петрозаводский государственный университет (ПетрГУ). Область научных интересов: алгоритмы и структуры данных. Число научных публикаций — 1. nikolayermolin@yahoo.com; пр. Ленина, 33, Петрозаводск, 185910, Республика Карелия, РФ; р.т.: +7(814-2)71-10-01.

Мазалов Владимир Викторович — д-р физ.-мат. наук, профессор, временно исполняющий обязанности директора, Институт прикладных математических исследований Карельского научного центра Российской академии наук (ИПМИ КарНЦ РАН), руководитель лаборатории математической кибернетики, Институт прикладных математических исследований Карельского научного центра Российской академии наук (ИПМИ КарНЦ РАН). Область научных интересов: теория игр, стохастическое динамическое программирование, математическая биология. Число научных публикаций — 156. vlmazalov@yandex.ru, http://www.krc.karelia.ru/HP/mazalov; ул. Пушкинская, 11, Петрозаводск, 185910, Республика Карелия, РФ; р.т.: +7(8142)78-11-08, Факс: +7(8142)76-63-13.

Печников Андрей Анатольевич — д-р техн. наук, доцент, руководитель лаборатории телекоммуникационных систем, Институт прикладных математических исследований Карельского научного центра Российской академии наук (ИПМИ КарНЦ РАН), главный научный сотрудник лаборатории телекоммуникационных систем, Институт прикладных математических исследований Карельского научного центра Российской академии наук (ИПМИ КарНЦ РАН). Область научных интересов: вебометрика, дискретная математика и математическая кибернетика, программные системы и модели. Число научных публикаций — 150. pechnikov@krc.karelia.ru, http://www.krc.karelia.ru/HP/pechnikov; ул. Пушкинская, 11, Петрозаводск, 185910, Республика Карелия, РФ; р.т.: +7(8142)78-11-08, Факс: +7(8142)76-63-13.

Поддержка исследований. Работа выполнена при финансовой поддержке РФФИ (проекты №№ 16-51-5506, 15-02-00352 и 15-01-06105).

DOI 10.15622/sp.55.10

N.A.ERMOLIN, V.V. MAZALOV, A.A. PECHNIKOV

GAME-THEORETIC METHODS FOR FINDING COMMUNITIES IN ACADEMIC WEB

Ermolin N.A., Mazalov V.V., Pechnikov A.A. **Game-Theoretic Methods for Finding Communities in Academic Web.**

Abstract. We consider the problem of community detection for the graph which is a fragment of the academic Web. The nodes of the graph are the sites of the scientific organizations, and its arcs are hyperlinks. We propose a new approach based on the methods of coalition game theory to derive the Nash-stable coalition partition. This is determined by a function of preferences for any pair of vertices in the graph. The problem of finding a stable partition is connected with finding a maximum of potential function. The algorithm for searching stable partitioning and evaluating its complexity is presented. The proposed method was compared with two well-known methods of finding communities. The efficiency of the new method is demonstrated on the fragment of the Web which consists of the official sites of the Siberian and Far East branches of RAS.

Keywords: web space, graph, community, modularity, coalition game theory.

Ermolin Nikolay Aleksandrovich — student, Petrozavodsk State University (PetrSU). Research interests: algorithms and data structures. The number of publications — 1. nikolayermolin@yahoo.com; 33, Lenin Str., 185910, Petrozavodsk, Republic of Karelia, Russia; office phone: +7(814-2)71-10-01.

Mazalov Vladimir Victorovich — Ph.D., Dr. Sci., professor, interim director, Institute of Applied Mathematical Research Karelian Research Centre of Russian Academy of Science, head of mathematical cybernetics laboratory, Institute of Applied Mathematical Research Karelian Research Centre of Russian Academy of Science. Research interests: webometrics, discrete mathematics, mathematical cybernetics, software systems and models. The number of publications — 156. vlmazalov@yandex.ru, http://www.krc.karelia.ru/HP/mazalov; 11, Pushkinskaya str., Petrozavodsk, 185910, Republic of Karelia, Russia, Russia; office phone: +7(8142)78-11-08, Fax: +7(8142)76-63-13.

Pechnikov Andrey Anatolievich — Ph.D., Dr. Sci., associate professor, head of telecommunications systems laboratory, Institute of Applied Mathematical Research Karelian Research Centre of Russian Academy of Science, chief researcher of telecommunications systems laboratory, Institute of Applied Mathematical Research Karelian Research Centre of Russian Academy of Science. Research interests: webometrics, discrete mathematics, mathematical cybernetics, software systems and models. The number of publications — 150. pechnikov@krc.karelia.ru, http://www.krc.karelia.ru/HP/pechnikov; 11, Pushkinskaya str., Petrozavodsk, 185910, Republic of Karelia, Russia, Russia; office phone: +7(8142)78-11-08, Fax: +7(8142)76-63-13.

Acknowledgements. This research is supported by RFBR (grants №№ 16-51-5506, 15-02-00352 and 15-01-06105).

References

1. Thelwall M. Webometrics and Social Web Research Methods. University of Wolverhampton. 2013. 140 p.

2. Hall W., Tiropanis T. Web evolution and Web Science.*Computer Networks*. 2012. vol. 56. no. 18. pp. 3859–3865.
3. Shokin Yu.I. et al. [Analysis of web-space of academic communities by methods of Webometrics and graph theory]. *Informacionnye tehnologii – Information technology.* 2014. vol. 12. pp. 31–40. (In Russ.).
4. Shokin Yu.I. et al. [A study of the academic web space of the Siberian branch of the Russian Academy of Sciences]. *Informacionnye tehnologii – Information technology.* 2012. vol. 17. pp. 86–98. (In Russ.).
5. Vesnin A.Yu., Konstantinova E.V., Savin M.Yu. [On scenarios of joining new sites to the webspace of the SB RAS]. *Vestnik Novosibirskogo gosudarstvennogo universiteta. Serija: Informacionnye tehnologii – Novosibirsk State University Journal of Information Technologies.* 2013. Issue 11. vol. 4. pp. 28–37. (In Russ.).
6. Klimenko O.A. [Models of the representation of academic web space]. *Informacionnye i matematicheskie tehnologii v nauke I upravlenii – Information and mathematical technologies in science and management.* 2016. № 2. pp. 103–110. (In Russ.).
7. Korelin V.N., Blekanov I.S., Sergeev S.L. [The modified algorithm LSH for clustering external web space universities]. *Nauchno-tehnicheskie vedomosti Sankt-Peterburgskogo gosudarstvennogo politehnicheskogo universiteta. Informatika. Telekommunikacii. Upravlenie – Scientific-technical Bulletin of Saint-Petersburg state Polytechnic University. Informatics. Telecommunications. Management.* 2015. vol. 5 (229). pp. 79–87. (In Russ.).
8. Bonato A., Graham F. C., Pralat P. Algorithms and Models for the Web Graph. Proceedings of the 13th International Workshop (WAW 2016). 2016. vol. 10088. 165 p.
9. Flake G.W., Lawrence S.R., Giles C.L., Coetzee F.M. Self-Organization and Identification of Web Communities. IEEE Computer. 2002. vol. 35. no. 3. pp. 66–71.
10. Labatut V., Balasque J.-M. Detection and Interpretation of Communities in Complex Networks: Practical Methodsand Application. *Computational Social Networks*. 2012. pp. 81–113.
11. Avrachenkov K., El Chamie M., Neglia G. Graph clustering based on mixing time of random walks. Proceedings of IEEE ICC 2014. 2014. pp. 4089–4094.
12. Newman M.E.J. Finding community structure in networks using the eigenvectors of matrices. *Phys. Rev*. 2006. vol. 74. no. 3. pp. 036104.
13. Girvan M., Newman M.E.J. Community structure in social and biological networks. Proc. of National Academy of Science. 2002. vol. 99(12). pp. 7821–7826.
14. 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. pp. P10008.
15. Avrachenkov K.E., Kondratev A.Yu., Mazalov V.V. Cooperative Game Theory Approaches for Network Partitioning. International Computing and Combinatorics Conference. 2017. LNCS 10392. pp. 591–602.
16. Gephi – The Open Graph Viz Platform. Available at: www.gephi.org (accessed: 09.06.2017).
17. DeDeo S., Krakauer D. Dynamics and Processing in Finite Self-Similar Networks. *Journal of the Royal Society Interface.* 2012. vol. 9. no. 74. pp. 2131–2144.
18. Bogomolnaia A., Jackson, M.O. The stability of hedonic coalition structures. *Games and Economic Behavior.* 2002. vol. 38. no. 2. pp. 201–230.
19. Golovin A.S., Pechnikov A.A. [Database external hyperlinks for the study of fragments of the Web]. *Informacionnaja sreda vuza XXI veka: materialy VII Vserossiiskoi nauchno-prakticheskoi konferencii* [The information environment of the University of the XXI century: materials of VII all-Russian scientific-practical conference]. Petrozavodsk. 2013. pp. 55–57. (In Russ.).
20. 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.