УДК 519.178 Вестник СПбГУ. Прикладная математика. Информатика... 2022. Т. 18. Вып. 1
MSC 05C70

Методы теории кооперативных игр в задаче ранжирования текстов

В. В. Мазалов[1,2], *В. А. Хитрая*[1,2], *А. В. Хитрый*[1]

[1] Федеральный исследовательский центр «Карельский научный центр Российской академии наук», Российская Федерация, 185910, Петрозаводск, ул. Пушкинская, 11
[2] Петрозаводский государственный университет, Российская Федерация, 185910, Петрозаводск, ул. Ленина, 33

Для цитирования: *Мазалов В. В., Хитрая В. А., Хитрый А. В.* Методы теории кооперативных игр в задаче ранжирования текстов // Вестник Санкт-Петербургского университета. Прикладная математика. Информатика. Процессы управления. 2022. Т. 18. Вып. 1. С. 63–78. https://doi.org/10.21638/11701/spbu10.2022.105

Предлагается метод ранжирования корпуса текстов новостного портала на основе мер центральности графа. Каждому тексту ставится в соответствие вершина некоторого графа, а его структура определяется на основе семантической связности текстов. В качестве меры центральности используется значение Майерсона в кооперативной игре на графе, где как характеристическая функция выбрано число простых путей в подграфе определенной длины m. Для значений m ранжирование на базе значений Майерсона будет различным. Для окончательного ранжирования предлагается применить процедуру ранжирования с помощью турнирной матрицы. Работа алгоритма ранжирования проиллюстрирована на численных примерах, связанных с конкретным порталом новостей.

Ключевые слова: корпус текстов новостей, граф, мера центральности, значение Майерсона, турнирная матрица, процедура ранжирования.

1. Введение. В связи с развитием Интернета значительно возросло количество текстовой информации. Агрегаторы, поисковики, а также различные интеллектуальные системы структурирования информации применяют множество методов для связи информационных потоков. Одними из наиболее часто используемых подходов являются построение графовых моделей и их дальнейший анализ. Такие модели могут быть построены на основе выделения семантических связей в корпусе текстов [1].

Существует устоявшийся процесс обработки текстов, который позволяет выделять семантическую связанность на основе содержания. Он состоит из очистки текста, приведения слов к минимальной форме и построения векторной модели корпуса [2], которая дает возможность изучения текстов уже безотносительно их конкретного содержания.

Классической задачей анализа корпуса текстов или набора документов (веб-страниц) является ранжирование. Данная процедура позволяет определить, какой из документов является ключевым в наборе, и выдать релевантную информацию в виде сортированного по рангу списка документов. К таким алгоритмам относится PageRank, разработанный Сергеем Брином и Ларри Пейджем в рамках исследовательского проекта BackRub для анализа связей между веб-страницами [3].

В графовых моделях для ранжирования можно использовать меры центральности вершин и ребер графа [4–10]. Центральность вершины отражает, насколько хорошо рассматриваемая вершина расположена на путях, которые соединяют остальные

вершины, и соответственно, насколько активно вершина участвует в процессе распространения информации между остальными вершинами в графе. Для определения меры центральности вершин предлагается применять методы теории кооперативных игр, в частности значение Майерсона [10–14]. В работе [15] для ранжирования рекомендован метод максимального правдоподобия.

В настоящей работе продемонстрировано, как методы теории кооперативных игр могут быть использованы для решения задачи определения центральности вершин графа и последующего их ранжирования. Была построена графовая модель корпуса текстов новостного портала, для которой проведена процедура ранжирования тремя методами, и сравнены полученные результаты.

Статья построена следующим образом. В п. 2 рассмотрен алгоритм построения графовой модели для анализа корпуса текстов, в п. 3 описаны методы теории кооперативных игр, которые могут быть применены для ранжирования текстов при использовании графовой модели. В п. 4 описан процесс построения графовой модели для эксперимента на основе коллекции новостных текстов, в п. 5 приведены результаты ранжирования текстов новостного портала с помощью модифицированного метода Майерсона, алгоритма PageRank, а также по правилу Борда.

2. Построение графовой модели связей множества текстов. Изучим коллекцию документов D. Заранее неизвестно, объединены ли эти документы конкретной тематикой и схожи ли они по смыслу. Определим степень сходства документов из коллекции D.

В первую очередь необходимо очистить документы от слов и символов, которые не представляют реального интереса при построении модели. Удаляются пунктуационные символы и так называемые «стоп-слова» (местоимения, предлоги, междометия и т. п.). Оставшиеся слова формируют множество слов, содержащихся в коллекции D. Обозначим его W.

Пример 1.1. Продемонстрируем этапы обработки текста на примере отрывка новости. Исходный текст:

Эскиз балета «Анна Каренина». Музыка Р. Щедрина, хореограф-постановщик Кирилл Симонов. Главы из романа Л. Толстого «Анна Каренина» читает Марина Перелешина (Москва) (16+). Музыкальный театр. По входным билетам.

На первом этапе произведена техническая очистка текста:

эскиз балета анна каренина музыка р щедрина хореограф постановщик кирилл симонов главы из романа л толстого анна каренина читает марина перелешина москва музыкальный театр по входным билетам

Для каждого слова $w_i \in W$ выделим лексему l_j [16], которая представляет собой совокупность всех значений и грамматических форм слова в языке. Общее число лексем n не превосходит общее число слов коллекции. Выделенные лексемы формируют множество лексем L. Одним из методов выделения лексем является стемминг [17]: слово, если это возможно, приводится к минимальной форме, для чего удаляются суффиксы, приставки, окончания. При необходимости процесс может быть повторен несколько раз. Наиболее распространен метод стемминга Портера, подробное описание которого можно найти в [18].

Пример 1.2. Для текста, приведенного выше, запишем набор слов, из которых формируются лексемы после удаления «стоп-слов»:

эскиз балет ан каренин музык р щедрин хореограф постановщик кирилл симон глав рома л толст ан каренин чита марин перелешин москв музыкальн театр входн билет.

Одним из методов анализа документа без учета синтаксической структуры или порядка слов в тексте является подсчет количества появлений слова в тексте. Для этого с помощью полученных лексем может быть сформирован так называемый мешок слов [19] документа $d \in D$:

$$b(d) = (b_{l_1}, b_{l_2}, ..., b_{l_n}),$$

где b_{l_j} — число слов в документе d, порождающих лексему l_j.

Построенный таким образом мешок слов имеет явно выраженную особенность. Существуют лексемы, которые будут часто встречаться в любых случайно выбранных текстах, причем в зависимости от размера документа такие лексемы могут иметь бóльший вес, чем те, которые определяют его семантику. Для того чтобы понизить их значимость в мешке слов, были созданы специальные методы, основанные на рассмотрении корпуса целиком, а не отдельного документа. Один из таких методов — применение статистической меры TF-IDF для оценки важности лексемы при анализе коллекции документов. В простейшем случае важность лексемы (или ее вес) прямо пропорциональна частоте использования лексемы в конкретном документе $tf_{l,d}$ и обратно пропорциональна частоте, с которой она встречается в коллекции документов, df_l [20]. Обычно в формуле для вычисления меры используется обратное значение $idf_l = \frac{1}{df_l}$. Тогда значение меры для лексемы l в документе d определяется следующим образом:

$$f_{l,d} = tf_{l,d} \cdot idf_l.$$

Введем вектор F_d — вектор мер лексем, содержащихся в документе d:

$$F_d = (f_{1,d}, ..., f_{n,d}).$$

На основе набора векторов мер лексем для M документов из коллекции D можно рассчитать симметричную матрицу сходства векторов C, состоящую из элементов

$$c_{i,j} = \cos\left(F_{d_i}, F_{d_j}\right) = \frac{F_{d_i} \cdot F_{d_j}}{\|F_{d_i}\| \|F_{d_j}\|},$$

которую можно интерпретировать как матрицу сходства документов из коллекции D:

$$C = \begin{pmatrix} c_{1,1} & ... & c_{M,1} \\ ... & & ... \\ c_{1,M} & ... & c_{M,M} \end{pmatrix}.$$

В свою очередь, по матрице C может быть построена матрица I. Для этого зададим пороговое значение c_{th} такое, что

$$I_{i,j} = \begin{cases} 1, & \text{если } c_{i,j} > c_{th},\ j \neq i, \\ 0, & \text{иначе,} \end{cases} \qquad i, j = 1, ..., M.$$

Пороговая величина c_{th} выбирается эмпирически в зависимости от определения достаточной семантической связи между документами. Чем она больше, тем более семантически связанными оказываются пары документов, для которых в матрице получены значения, равные 1. Построенная таким образом матрица I может рассматриваться как матрица смежности для некоторого неориентированного графа G, который может применяться в качестве графовой модели корпуса текстов.

3. Ранжирование текстов на основе теории кооперативных игр.

3.1. Значение Майерсона. При использовании графовой модели ранжирование текстов представляет собой поиск значений центральности вершин графа G, описываемого с помощью матрицы смежности I. Для определения центральности вершин в графе может быть применен теоретико-игровой подход. Определим на графе $G = (V, E)$, где V — множество вершин и E — множество ребер, а также кооперативную игру $\Gamma =< V, v >$, $|V| = n$. В этой игре вершины служат игроками, а характеристическая функция $v(K), K \subset V$, определяется как количество простых путей длины m в подграфе, порожденном множеством вершин K. Число $m = 1, 2, \ldots$ фиксировано. Очевидно, что функция $v(K)$ является монотонной, т. е. $v(K_1) \leqslant v(K_2)$, если $K_1 \subset K_2$. Тогда для ранжирования вершин в графе можно использовать решение кооперативной игры в виде значения Шепли — Майерсона. В работах [12, 13] было показано, что в случае, когда характеристическая функция $v(K), K \subset V_\eta$, определена как количество всех простых путей длины m в подграфе K, то значение Майерсона для игрока i равно

$$\varphi_i^m = \frac{a_m(i)}{m+1}, \tag{1}$$

где $a_m(i)$ есть число простых путей длины m, проходящих через вершину i. Пути $i_1, i_2, ..., i_k$ и $i_k, ..., i_2, i_1$ считаются одинаковыми.

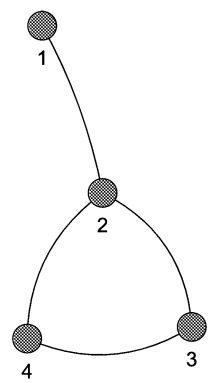

Рис. 1. Граф из четырех вершин

Например, для графа из четырех вершин, представленного на рис. 1, для $m = 2$ значение Майерсона для игрока 2 равно $\varphi_2^2 = 4/3$, поскольку через вершину 2 проходят 4 простых пути: $(1, 2, 3)$, $(1, 2, 4)$, $(2, 3, 4)$ и $(2, 4, 3)$. Следовательно, $\varphi_2^2 = 4/3 = 1.333$. А для игрока 1 величина Майерсона $\varphi_1^2 = 2/3 = 0.666$, поскольку через вершину 1 проходят лишь два пути длины 2: $(1, 2, 3)$ и $(1, 2, 4)$.

Вершины с высоким значением Майерсона играют большую роль в структуре графа, вершины с небольшим — не так важны. Покажем это на примере графа, изображенного на рис. 2. В табл. 1 приведены значения Майерсона, вычисленные для различных значений m. Видно, что, например, для вершины 5 центральность сильно варьируется. Если для $m = 1$ эта вершина имеет минимальный ранг, то при $m = 4$ он максимальный.

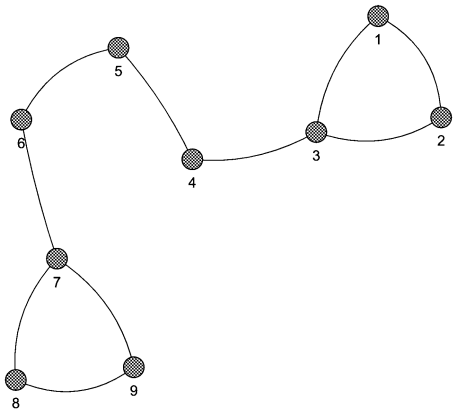

Рис. 2. Изменение величин m в графе из девяти вершин

Таблица 1. **Значения центральности вершин графа для различных m**

m	Вершина								
	1	2	3	4	5	6	7	8	9
1	1	1	1.5	1	1	1	1.5	1	1
2	1.33	1.33	2	1	1	1	2	1.3	1.3
3	0.75	0.75	1.25	1.5	1.5	1.5	1.25	0.75	0.75
4	0.6	0.6	1	1.4	1.8	1.4	1	0.6	0.6
5	0.5	0.5	1	1.33	1.33	1.33	1	0.5	0.5
6	0.57	0.57	0.86	0.86	0.86	0.86	0.86	0.57	0.57
7	0.75	0.75	1	1	1	1	1	0.75	0.75

3.2. Решение многокритериальной задачи. В работе [9] было предложено в качестве характеристической функции рассматривать многочлен вида

$$v(K,r) = \sum_{m=1}^{\infty} A_m(K) r^m, \quad K \subseteq N, \quad 0 \leqslant r \leqslant 1,$$

здесь $A_m(K)$ — количество простых путей длины m в подграфе, порожденном множеством вершин K. Тогда значение Майерсона для игрока i будет равно [12, 13]

$$\varphi_i(r) = \sum_{m=1}^{\infty} \frac{a_m(i)}{m+1} r^m, \quad i \in N,$$

где $a_m(i)$ — число простых путей длины m, проходящих через вершину i. Например, для графа, представленного на рис. 1, значение Майерсона для игрока 2 будет равно

$$\varphi_2(r) = \frac{3}{2}r + \frac{4}{3}r^2 + \frac{2}{4}r^3. \tag{2}$$

Заметим, что выражение (2) зависит от параметра r. Варьируя параметр от 0 до 1, можно получать различные порядки ранжирования вершин.

Можно предложить альтернативный подход, используя для ранжирования вершин альтернативные методы на основе турнирной матрицы, которая составляется следующим образом.

Рассмотрим двух игроков (две вершины) i и j. Сравнивая значения Майерсона для различных m, в матрице R_{ij} ставим 1 (выигрыш), если ранг i в большинстве

случаев выше, чем у j, ставим 0 (проигрыш), если ранг i в большинстве случаев ниже, чем у j, и 1/2 (ничья), иначе. Так, для графа, изображенного на рис. 2, по значениям, приведенным в табл. 1, турнирная матрица имеет вид

$$R = \begin{pmatrix} \cdot & \frac{1}{2} & 0 & 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \cdot & 0 & 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ 1 & 1 & \cdot & 0 & 0 & 0 & \frac{1}{2} & 1 & 1 \\ 1 & 1 & 1 & \cdot & 0 & \frac{1}{2} & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & \cdot & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & \frac{1}{2} & 0 & \cdot & 1 & 1 & 1 \\ 1 & 1 & \frac{1}{2} & 0 & 0 & 0 & \cdot & 1 & 1 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & 0 & 0 & \cdot & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & 0 & 0 & \frac{1}{2} & \cdot \end{pmatrix}. \tag{3}$$

Например, если сравнить игроков 5 и 3, то игрок 5 проигрывает игроку 3 для $m = 1, 2$, выигрывает у игрока 3 для $m = 3, 4, 5$ и у игроков ничья для $m = 6, 7$. Поэтому $R_{53} = 1$, соответственно $R_{35} = 0$.

В работе [21] был предложен теоретико-игровой подход ранжирования игроков на основе турнирной матрицы. При таком подходе характеристическая функция может быть определена следующим образом. Положим $v(N) = 1$ и для коалиции $K \subset N$ получим, что

$$v(K) = \max_{i \in K} \min_{j \in N \setminus K} R_{ij}.$$

Затем в качестве ранжирования можно использовать любое решение кооперативной игры: значение Шепли, N- или C-ядро.

Можно также воспользоваться правилом Борда. В соответствии с этим правилом нужно сложить в каждой строке турнирной матрицы все ранги и ранжировать игроков согласно этому суммарному рангу. В частности, для турнирной матрицы (3) эти ранги равны соответственно $1.5, 1.5, 4.5, 6.5, 8, 6.5, 4.5, 1.5, 1.5$. Таким образом, для графа, представленного на рис. 2, победителем является вершина 5, поскольку игрок 5 выигрывает у любого из игроков (так называемый победитель Кондорсе).

Заметим, что в качестве значений центральности было рассмотрено значение Майерсона. Для его вычисления по формуле (1) нужно найти число всех простых путей определенной длины, проходящих через данную вершину. Это не простая вычислительная задача. Была предложена модификация значения Майерсона [14], которая проще для вычислений. Идея этого представления заключается в том, что в каждом простом пути вершина i, для которой вычисляется вектор Майерсона, встречается один раз. Тогда $a_m(i)$ можно трактовать как число появлений вершины i во всех простых путях длины m. В [14] описана модификация центральности по Майерсону, где центральность k-го порядка вершины i — это число появлений вершины i в путях длины k, включая циклы. Вектор $\sigma(k)$ — вектор центральностей вершин графа G, i-я компонента которого равна

$$\sigma_i(k) = \frac{s_i(k)}{k+1},\ i = 1, ..., n,$$

где $s_i(k)$ — суммарное число появлений вершины i в путях длины k, вычисляемое по формуле

$$s_i(k) = \sum_{j=1}^{n} a_{ij}^{(k)} + \sum_{l=1}^{n} \left[a_{li} \sum_{j=1}^{n} a_{ij}^{(k-1)} + a_{li}^{(2)} \sum_{j=1}^{n} a_{ij}^{(k-2)} + \ldots + a_{li}^{(k)} \right].$$

В этом выражении $a_{ij}^{(k)}$ — элементы матрицы смежности, возведенной в соответствующую степень k. Именно этот метод и будет использоваться при ранжировании текстов в данной работе.

Существующие меры центральности имеют как положительные, так и отрицательные стороны. Например, мера betweenness centrality [4] учитывает только кратчайшие пути между вершинами, в то же время как информация может распространяться по произвольным путям в сети. Наиболее популярной является мера PageRank [3], основанная на предельном распределении случайного блуждания по вершинам графа. Ее вычисление требует нахождения обратной фундаментальной матрицы гигантского размера. Однако это один из самых быстрых методов. Но его оценки также вызывают критику. Так, центральности всех вершин графа, изображенного на рис. 3, по методу PageRank одинаковы и равны 1/6. В то же время значение Майерсона для вершин 1 и 3 различаются и равны, в частности, для $m = 2$ соответственно $\varphi_1^2 = 6/3 = 2$, $\varphi_3^2 = 9/3 = 3$. Это более отвечает нашему представлению о центральности вершин 1 и 3 в таком графе.

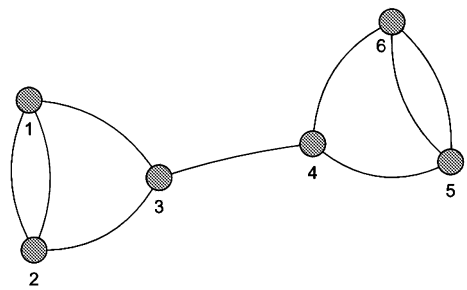

Рис. 3. Сравнение величин центральности вершин графа 1 и 3, полученных по методу Page Rank и значению Майерсона

4. Построение графа на основе коллекции новостных текстов. Программная часть реализована в среде Jupyter на языке Python 3 и использованием инструментов NumPy, Scikit-learn, NLTL, а также различных библиотек для обработки и представления данных. Для хранения данных применялись базы SQLite и PostgreSQL, а также формат GraphML для представления графовых моделей.

Для практического эксперимента была выбрана коллекция из 500 статей новостного портала «Петрозаводск Говорит». Для каждой новости были собраны параметры `node_id` — идентификатор новости на сайте, `title` — заголовок новости, `body` — текст новости.

Получение данных с сайта было реализовано при помощи кравлера, выполнявшего последовательный проход по страницам, базируясь на инкрементации `node_id`, которая в случае отсутствия статьи с таким идентификатором игнорировалась. Все статьи записывались в исходном виде в базу данных для последующей обработки.

Тексты новостей прошли предварительную обработку в несколько этапов:

1) удаление новостей, которые связывают по искусственным характеристикам. Это такие новости, как «Топ-5» и «Обзор недели», которые создавали нерелевантные связи, так как основывались на популярности или попадании во временной отрезок, а не на содержании;

2) техническая очистка текста, т. е. удаление символов, которые не относятся к русскому алфавиту, склеивание двойных пробелов в один, удаление табуляций и

переносов, приведение всех символов к строчным. Данный этап необходим для упрощения модели, так как удаляемые параметры не имеют влияния на семантические связи;

3) формирование множества слов коллекции документов было произведено при помощи пакета ruts с дополнительной опцией удаления слов, которые являются междометиями, союзами и прочими несодержательными для анализа конструкциями. Затем при помощи пакета NLTK были выделены лексемы.

Мешок слов построен с помощью TF-IDF токенайзера из пакета sklearn. В данном пакете мера IDF вычисляется по следующей формуле:

$$idf_w = \ln \frac{1 + M}{1 + df_w},$$

где M — число документов в обрабатываемом корпусе; df_w — количество документов, содержащих слово w. Полученный вектор — мешок слов — нормализуется при помощи евклидовой нормы.

Для сокращения размерности векторов $b(d_i)$ из полученного множества лексем L были удалены лексемы l_j, для которых $b_{l_j} < 0.09$. Это значение было выбрано эмпирически, на основе визуального анализа значимости удаляемых элементов.

Затем при помощи вычисления косинусной меры из библиотеки `scipy` была построена матрица сходства C, которая была преобразована в матрицу смежности I с пороговым значением $c_{th} = 0.72$, также определенным эмпирически с помощью анализа выделяемых групп. На рис. 4 представлен граф, построенный на основе полученной матрицы I.

В результате были выбраны девять подграфов, содержащих более пяти вершин. Дальнейший анализ выявил, что тексты, образовавшие такие подграфы, тесно связаны семантически: G_1 — граф, содержащий документы о прогнозе погоды, G_2 — новости о розыске МВД, G_3 — дорожно-транспортные происшествия, G_4 — граф, образованный документами, содержащими новости о бывшем вице-премьере правительства Республики Карелия, G_5 — театральные новости, G_6 — автомобильные пожары, G_7 — документы, описывающие новости мэрии Петрозаводска, G_8 — бюджет Республики Карелия, G_9 — новости о военнослужащих. Выбранные подграфы иллюстрирует рис. 5, *а–з*.

5. Результаты ранжирования текстов. В данном разделе приведены результаты ранжирования текстов новостного портала с помощью модифицированного метода Майерсона, алгоритма PageRank и по правилу Борда.

Рассмотрим граф G_1, представленный на рис. 6. Для его вершин вычислим величины центральностей порядка k, на их основе проведем ранжирование по правилу Борда, а также рассчитаем значения PageRank со стандартными параметрами (табл. 2).

Для сравнения результатов значения центральности вершин, полученные с помощью модифицированного метода Майерсона при $k = 5$, числовые величины, характеризующие важность вершин, определенные с помощью алгоритма PageRank, и ранги, построенные на основе турнирной матрицы по правилу Борда, нормируются. Кривые на рис. 7 позволяют визуально оценить результаты ранжирования.

Предложенный график дает возможность однозначно определить схожесть результатов, полученных тремя методами. Стоит отметить, что наибольшие значения получены для одной и той же вершины. Таким образом, ранжирование текстов с помощью модифицированного метода Майерсона или по правилу Борда может быть предложено в качестве альтернативы алгоритму PageRank.

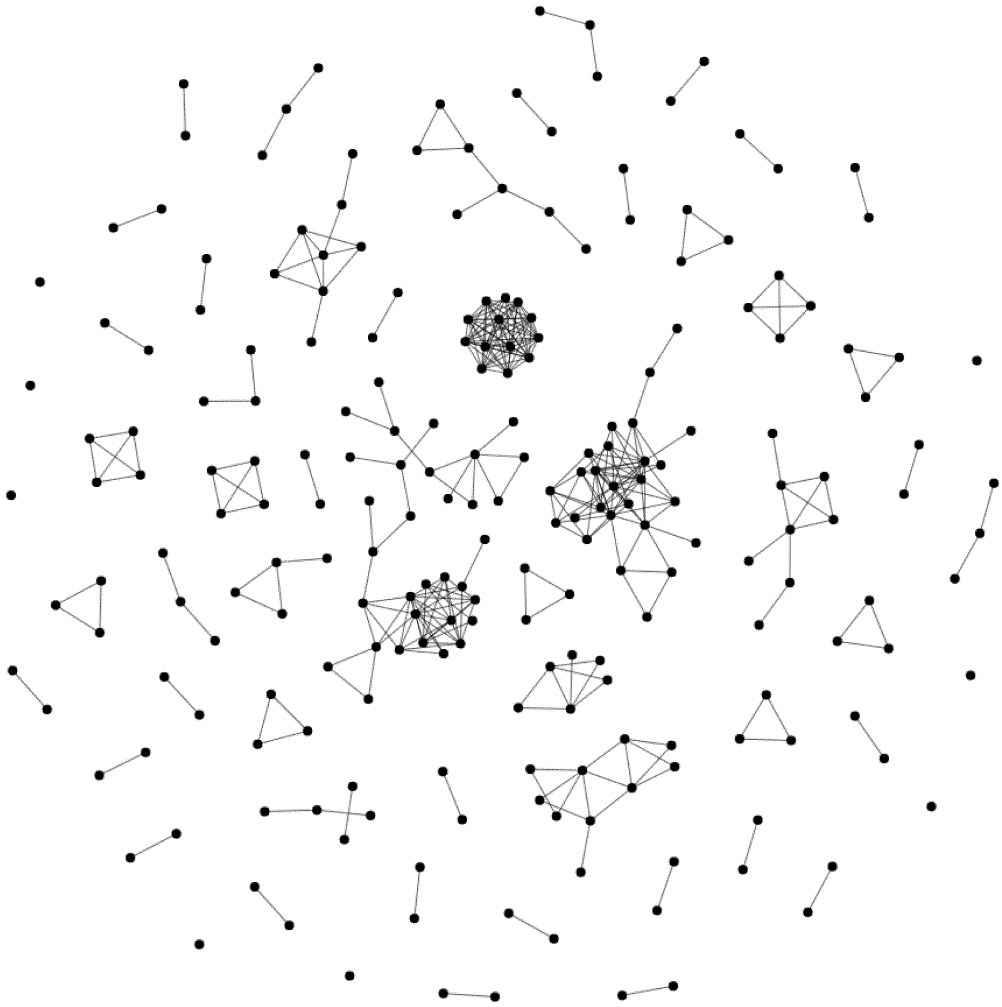

Рис. 4. Общий граф коллекции новостных текстов

Таблица 2. **Результаты ранжирования вершин графа** G_1

Вершина	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$	PageRank	Правило Борда
1	5	36	265	2036	15 698	0.035	13
2	12	110	928	7632	62 023	0.074	23
3	2	6	27	159	1029	0.013	1
4	13	118	1000	8321	67 952	0.08	25
5	4	20	131	907	6614	0.031	7
6	4	22	145	1009	7337	0.027	9
7	12	110	947	7909	64 931	0.075	24
8	3	13	75	483	3365	0.020	6
9	5	47	363	2977	23 804	0.035	18
10	10	86	711	5767	46 446	0.063	21
11	5	36	262	1992	15 267	0.032	12
12	2	12	69	494	3477	0.018	5
13	4	36	269	2167	17 010	0.030	14

Окончание таблицы 2

Вершина	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$	PageRank	Правило Борда
14	5	33	231	1703	12 842	0.034	10
15	6	47	362	2861	22 533	0.040	17
16	1	1	4	20	122	0.017	0
17	7	55	434	3449	27 315	0.044	19
18	1	9	53	387	2727	0.012	4
19	9	75	600	4804	38 306	0.069	20
20	2	5	28	166	1156	0.027	2
21	1	6	40	265	1883	0.013	3
22	12	104	869	7134	57 800	0.076	22
23	5	36	257	1976	15 104	0.040	11
24	6	38	289	2172	16 703	0.038	15
25	5	45	339	2760	21 808	0.034	16
26	3	21	144	1051	7889	0.024	8

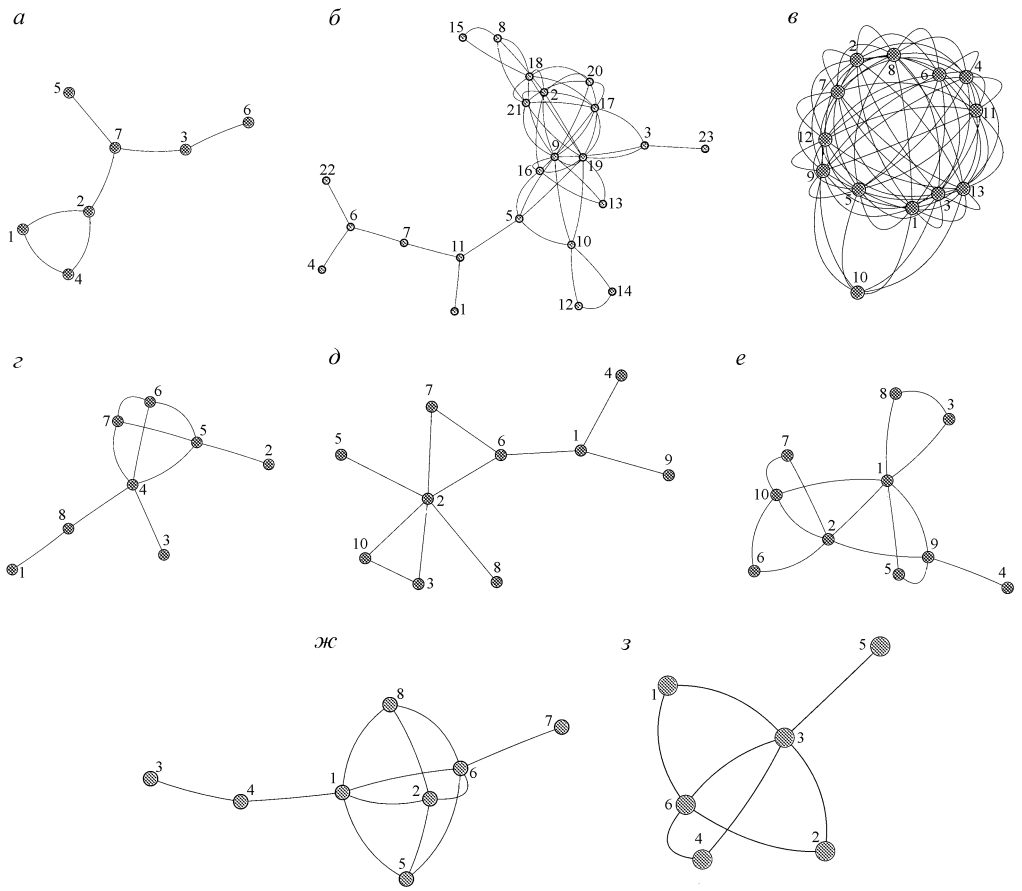

Рис. 5. Выбранные подграфы, содержащие более пяти вершин
а — G_2, *б* — G_3, *в* — G_4, *г* — G_5, *д* — G_6, *е* — G_7, *ж* — G_8, *з* — G_9

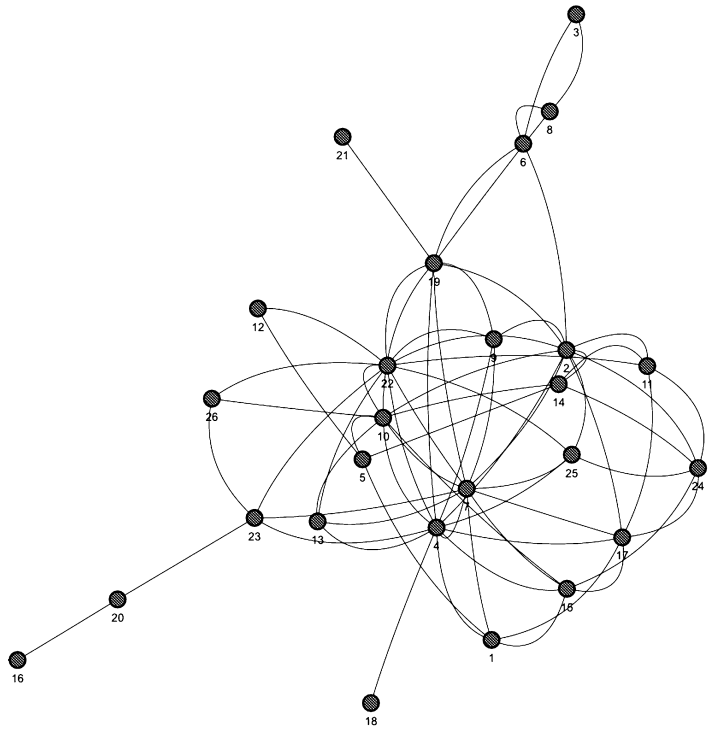

Рис. 6. Граф G_1

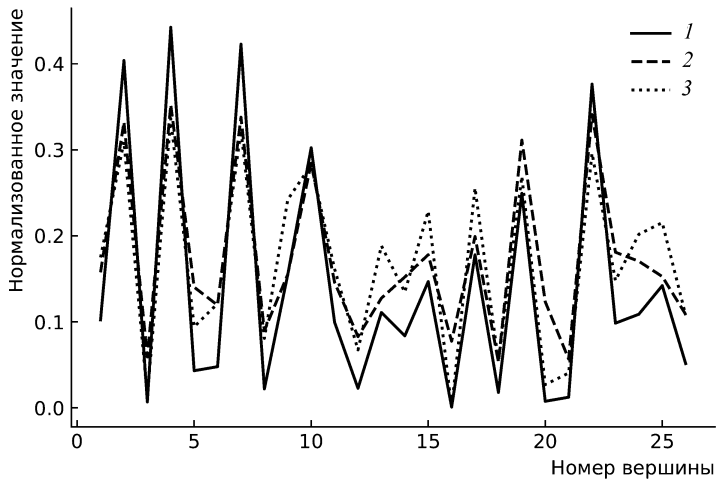

Рис. 7. Значения центральности вершин, полученные с помощью модифицированного метода Майерсона при $k = 5$ (*1*), алгоритма PageRank (*2*) и на основе турнирной матрицы по правилу Борда (*3*)

Сравнение значений на остальных графах по итогам работы с коллекцией новостных текстов дает аналогичную картину. Результаты представлены на рис. 8, *а–з*.

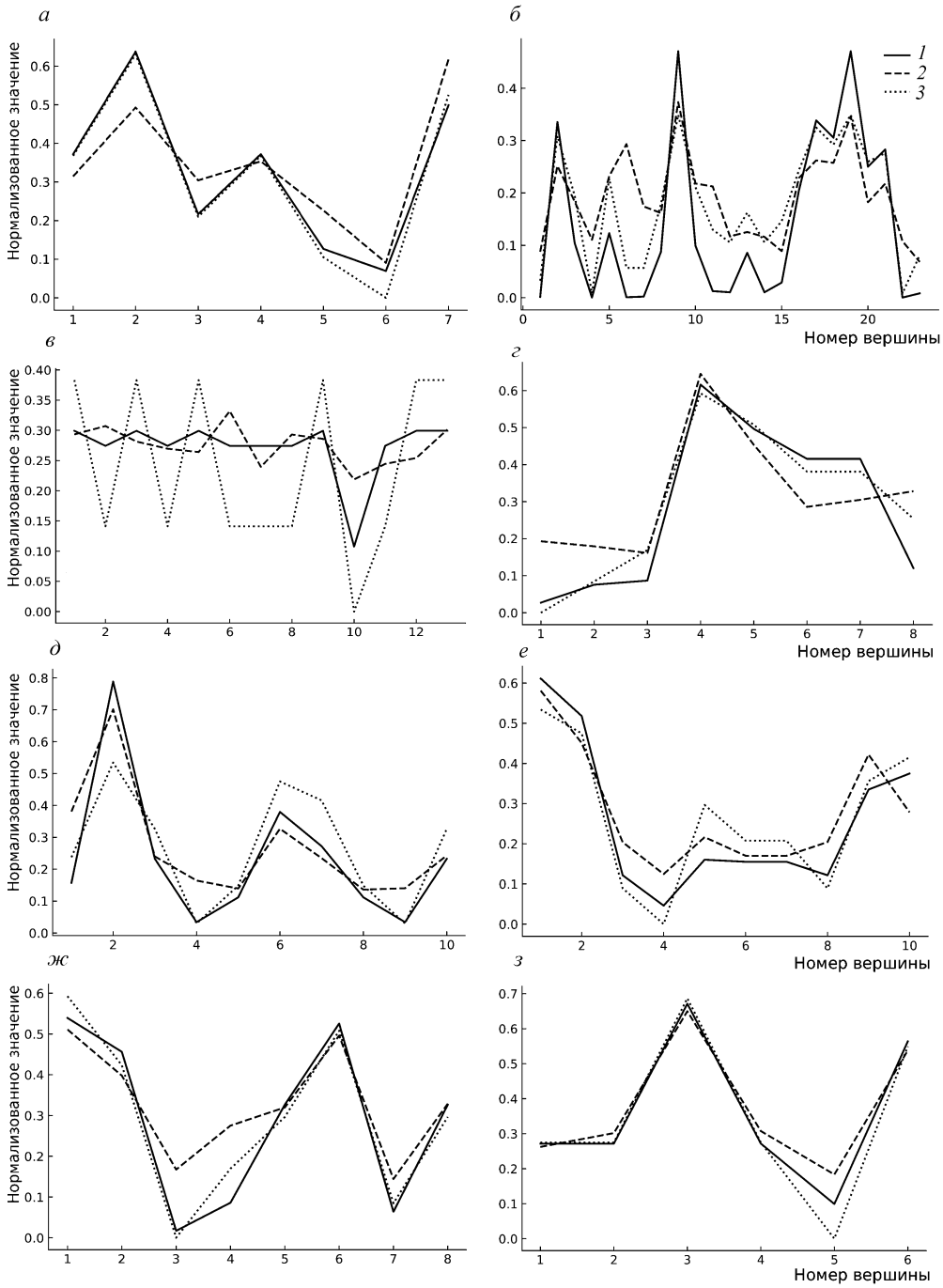

Рис. 8. Сравнение метода Майерсона (*1*), алгоритма PageRank (*2*) и правила Борда (*3*)

6. Заключение. В работе продемонстрировано, что в задачах ранжирования текстов применим теоретико-игровой подход. Описан процесс работы с корпусом текстов новостного портала, построена графовая модель данного корпуса. В качестве методов ранжирования предлагаются модифицированный метод Майерсона в кооперативной игре, опирающийся на число путей в графе фиксированной длины, включая циклы, а также правило Борда на основе турнирной матрицы. Схожесть результатов ранжирования позволяет прийти к выводу о возможности применения модифицированного метода Майерсона и правила Борда для анализа корпуса текстов наравне с повсеместно распространенным алгоритмом PageRank.

Литература

1. *Silva A., Lozkins A., Bertoldi L. R., Rigo S., Bure V. M.* Semantic textual similarity on Brazilian Portuguese: An approach based on language-mixture models // Вестник Санкт-Петербургского университета. Прикладная математика. Информатика. Процессы управления. 2019. Т. 15. Вып. 2. С. 235–244. https://doi.org/10.21638/11701/spbu10.2019.207

2. *Jones K. S.* A statistical interpretation of term specificity and its application in retrieval // J. Documentation. 2004. Vol. 60. N 5. P. 493–502. https://doi.org/10.1108/00220410410560573

3. *Page L., Brin S., Motwani R., Winograd T.* The pagerank citation ranking: Bringing order to the Web // Proceedings of the 7^{th} International World Wide Web Conference. Brisbane, Australia. 1998. P. 161–172. URL: http://citeseer.nj.nec.com/page98pagerank.html (дата обращения: 15 июля 2021 г.).

4. *Freeman L. C.* A set of measures of centrality based on betweenness // Sociometry. 1977. Vol. 40. N 1. P. 35–41. http://dx.doi.org/10.2307/3033543

5. *Brandes U.* Centrality measures based on current flow // STACS 2005. 22^{nd} Annual Symposium on Theoretical Aspects of Computer Science. Stuttgart, Germany. February 24–26, 2005. Proceedings. Eds. by V. Diekert, B. Durand. Vol. 3404 of Lecture Notes in Computer Science. Stuttgart: Springer, 2005. P. 533–544. https://doi.org/10.1007/978-3-540-31856-9_44

6. *Avrachenkov K., Litvak N., Medyanikov V., Sokol M.* Alpha current flow betweenness centrality // Algorithms and Models for the Web Graph. 10^{th} International Workshop (WAW 2013). Cambridge, MA, USA. December 14–15, 2013. Proceedings. Eds. by A. Bonato, M. Mitzenmacher, P. Pralat. Vol. 8305 of Lecture Notes in Computer Science. Cambridge: Springer, 2013. P. 106–117. https://doi.org/10.1007/978-3-319-03536-9_9

7. *Avrachenkov K. E., Mazalov V. V., Tsynguev B. T.* Beta current flow centrality for weighted networks // Computational Social Networks. 4^{th} International Conference (CSoNet 2015). Beijing, China. August 4–6, 2015. Proceedings. Lecture Notes in Computer Science. N 9197. 2015. P. 216–227. https://doi.org/10.1007/978-3-319-21786-4_19

8. *Newman M. E. J.* A measure of betweenness centrality based on random walks // Social Networks. 2005. Vol. 27. P. 39–54. http://dx.doi.org/10.1016/j.socnet.2004.11.009

9. *Jackson M. O.* Social and economic networks. Princeton, USA: Princeton University Press, 2008. 504 p. https://doi.org/10.1515/9781400833993

10. *Gomez D., Gonzalez-Aranguena E., Manuel C.* et al. Centrality and power in social networks: a game theoretic approach // Math. Soc. Sci. 2003. Vol. 46, N 1. P. 27–54. https://doi.org/10.1016/S0165-4896(03)00028-3

11. *Skibski O., Tomasz P., Talal R.* Axiomatic characterization of game theoretic centrality // J. Artif. Intell. Res. 2018. Vol. 62. P. 33–68. https://doi.org/10.1613/jair.1.11202

12. *Mazalov V. V., Trukhina L. I.* Generating functions and the Myerson vector in communication networks // Diskr. Mat. 2014. Vol. 26. N 3. P. 65–75. https://doi.org/10.1515/dma-2014-0026

13. *Avrachenkov K., Kondratev A. Yu., Mazalov V. V., Rubanov D. G.* Network partitioning as cooperative games // Computational social networks. 2018. Vol. 5. N 11. P. 1–28.

14. *Мазалов В. В., Хитрая В. А.* Модифицированное значение Майерсона для определения центральности вершин графа // Математическая теория игр и ее приложения. 2019. Vol. 11. № 2. С. 19–39.

15. *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. https://doi.org/10.21638/11701/spbu10.2018.302

16. *Korobov M.* Morphological analyzer and generator for russian and ukrainian languages // Analysis of Images, Social Networks and Texts / Eds. by M. Yu. Khachay, N. Konstantinova, A. Panchenko et al. Cham: Springer International Publ., 2015. Vol. 542 of Communications in Computer and Information Science. P. 320–332. http://dx.doi.org/10.1007/978-3-319-26123-231

17. *Lovins J. B.* Development of a stemming algorithm // Mech. Transl. Comput. Linguistics. 1968. Vol. 11. N 12. P. 22–31. URL: http://www.mtarchive.info/MT1968-Lovins.pdf (дата обращения: 15 июля 2021 г.).

18. *Van Rijsbergen C. J., Robertson S. E., Porter M. F.* New models in probabilistic information retrieval // Computer Laboratory. Cambridge, USA: Cambridge University Press, 1980. 613 p.

19. *Harris Z.* Distributional structure // Word. 1954. Vol. 10, N 2–3. P. 146–162. URL: https://link.springer.com/chapter/10.1007/978-94-009-8467-71 (дата обращения: 15 июля 2021 г.).

20. *Manning C. D., Raghavan P., Schutze H.* Introduction to information retrieval. Cambridge, USA: Cambridge University Press, 2008. 535 p.

21. *Kondratev A. A., Mazalov V. V.* Ranking procedure with the shapley value // Intelligent Information and Database Systems. 9th Asian Conference (ACIIDS 2017). Kanazawa, Japan. April 3–5, 2017. Proceedings. P. II / Eds. by N. T. Nguyen, S. Tojo, L. M. Nguyen, B. Trawinski. 2017. Vol. 10192 of Lecture Notes in Computer Science. P. 691–700. https://doi.org/10.1007/978-3-319-54430-4_66

Статья поступила в редакцию 21 августа 2021 г.
Статья принята к печати 1 февраля 2022 г.

Контактная информация:

Мазалов Владимир Викторович — д-р физ.-мат. наук, проф.; vmazalov@krc.karelia.ru

Хитрая Виталия Андреевна — аспирант, ст. преп.; nadezhda_ego@mail.ru

Хитрый Андрей Владимирович — аспирант; andrey.khitryy@gmail.com

Cooperative game theory methods for text ranking

V. V. Mazalov[1,2], *V. A. Khitraya*[1,2], *A. V. Khitryi*[1]

[1] Federal Research Center "Karelian Research Center of the Russian Academy of Sciences", 11, Pushkinskaya ul., Petrozavodsk, 185910, Russian Federation
[2] Petrozavodsk State University, 33, ul. Lenina, Petrozavodsk, 185910, Russian Federation

For citation: Mazalov V. V., Khitraya V. A., Khitryi A. V. Cooperative game theory methods for text ranking. *Vestnik of Saint Petersburg University. Applied Mathematics. Computer Science. Control Processes*, 2022, vol. 18, iss. 1, pp. 63–78.
https://doi.org/10.21638/11701/spbu10.2022.105 (In Russian)

A method of ranking the corpus of texts of a news portal, based on measures of graph centrality, is proposed. Each text is assigned a vertex of a certain graph, and its structure is determined based on the semantic connectivity of the texts. As a measure of centrality, the Myerson value is used in a cooperative game on a graph, where the number of simple paths in a subgraph of a certain length m is chosen as a characteristic function For different values of m, the ranking based on the Myerson value will be different. For the final ranking, it is proposed to use the ranking procedure based on the tournament matrix. The operation of the ranking algorithm is illustrated by numerical examples related to a specific news porta.

Keywords: text corpus of news, graph, centrality measure, Myerson value, tournament matrix, ranking procedure.

References

1. Silva A., Lozkins A., Bertoldi L. R., Rigo S., Bure V. M. Semantic textual similarity on Brazilian

Portuguese: An approach based on language-mixture models. *Vestnik of Saint Petersburg University. Applied Mathematics. Computer Science. Control Processes*, 2019, vol. 15, iss. 2, pp. 235–244. https://doi.org/10.21638/11701/spbu10.2019.207 (In Russian)

2. Jones K. S. A statistical interpretation of term specificity and its application in retrieval. *J. Documentation*, 2004, vol. 60, no. 5, pp. 493–502. https://doi.org/10.1108/00220410410560573

3. Page L., Brin S., Motwani R., Winograd T. The pagerank citation ranking: Bringing order to the Web. *Proceedings of the 7^{th} International World Wide Web Conference*. Brisbane, Australia, 1998, pp. 161–172. Available at: citeseer.nj.nec.com/page98pagerank.html (accessed: July 15, 2021).

4. Freeman L. C. A set of measures of centrality based on betweenness. *Sociometry*, 1977, vol. 40, no. 1, pp. 35–41. http://dx.doi.org/10.2307/3033543

5. Brandes U. Centrality measures based on current flow. *STACS 2005. 22^{nd} Annual Symposium on Theoretical Aspects of Computer Science*. Stuttgart, Germany, February 24–26, 2005. Proceedings, Eds. by V. Diekert, B. Durand, vol. 3404 of Lecture Notes in Computer Science. Stuttgart, Springer Publ., 2005, pp. 533–544. https://doi.org/10.1007/978-3-540-31856-9_44

6. Avrachenkov K., Litvak N., Medyanikov V., Sokol M. Alpha current flow betweenness centrality. *Algorithms and Models for the Web Graph. 10^{th} International Workshop (WAW 2013)*. Cambridge, MA, USA, December 14–15, 2013. Proceedings. Eds. by A. Bonato, M. Mitzenmacher, P. Pralat, vol. 8305 of Lecture Notes in Computer Science. Cambridge, Springer Publ., 2013, pp. 106–117. https://doi.org/10.1007/978-3-319-03536-9_9

7. Avrachenkov K. E., Mazalov V. V., Tsynguev B. T. Beta current flow centrality for weighted networks. *Computational Social Networks. 4^{th} International Conference (CSoNet 2015)*. Beijing, China, August 4–6, 2015. Proceedings, Lecture Notes in Computer Science, no. 9197, 2015, pp. 216–227. https://doi.org/10.1007/978-3-319-21786-4_19

8. Newman M. E. J. A measure of betweenness centrality based on random walks. *Social Networks*, 2005, vol. 27, pp. 39–54. http://dx.doi.org/10.1016/j.socnet.2004.11.009

9. Jackson M. O. *Social and economic networks*. Princeton, USA, Princeton University Press, 2008, 504 p. https://doi.org/10.1515/9781400833993

10. Gomez D., Gonzalez-Aranguena E., Manuel C. et al. Centrality and power in social networks: a game theoretic approach. *Math. Soc. Sci.*, 2003, vol. 46, no. 1, pp. 27–54. https://doi.org/10.1016/S0165-4896(03)00028-3

11. Skibski O., Tomasz P., Talal R. Axiomatic characterization of game theoretic centrality. *J. Artif. Intell. Res.*, 2018, vol. 62, p. 33–68. https://doi.org/10.1613/jair.1.11202

12. Mazalov V. V., Trukhina L. I. Generating functions and the Myerson vector in communication networks. *Diskr. Mat.*, 2014, vol. 26, no. 3, pp. 65–75. https://doi.org/10.1515/dma-2014-0026

13. Avrachenkov K., Kondratev A. Yu., Mazalov V. V., Rubanov D. G. Network partitioning as cooperative games. *Computational social networks*, 2018, vol. 5, no. 11, pp. 1–28.

14. Mazalov V. V., Khitraya V. A. Modificirovannoe znachenie Maiersona dlya opredeleniya centralnosti vershin grafa [Modified Mayerson value for determining the centrality of graph vertices]. *Matematicheskaia teoriia igr i ee prilozheniia* [*Mathematical theory players and its supplements*], 2019, vol. 11, no. 2, pp. 19–39. (In Russian)

15. Mazalov V. V., Nikitina N. N. Metod maksimalnogo pravdopodobia dlya vydelenya soobshestv v commukacionnih setyah [Maximum likelihood method for detecting communities in communication networks]. *Vestnik of Saint Petersburg University. Applied Mathematics. Computer Science. Control Processes*, 2018, vol. 14, no. 3, pp. 200–214. https://doi.org/10.21638/11701/spbu10.2018.302 (In Russian)

16. Korobov M. Morphological analyzer and generator for Russian and Ukrainian languages. *Analysis of Images, Social Networks and Texts*. Eds. by M. Yu. Khachay, N. Konstantinova, A. Panchenko et al. Cham, Springer International Publ., 2015, vol. 542 of Communications in Computer and Information Science, pp. 320–332. http://dx.doi.org/10.1007/978-3-319-26123-231

17. Lovins J. B. Development of a stemming algorithm. *Mech. Transl. Comput. Linguistics*, 1968, vol. 11, no. 12, pp. 22–31. Available at: http://www.mtarchive.info/MT1968-Lovins.pdf (accessed: July 15, 2021).

18. Van Rijsbergen C. J., Robertson S. E., Porter M. F. New models in probabilistic information retrieval. *Computer Laboratory*. Cambridge, USA, Cambridge University Press, 1980, 613 p.

19. Harris Z. Distributional structure. *Word*, 1954, vol. 10, no. 2-3, pp. 146–162. Available at: https://link.springer.com/chapter/10.1007/978-94-009-8467-71 (accessed: July 15, 2021).

20. Manning C. D., Raghavan P., Schutze H. *Introduction to information retrieval*. Cambridge, USA, Cambridge University Press, 2008, 535 p.

21. Kondratev A. A., Mazalov V. V. Ranking procedure with the shapley value. *Intelligent Information and Database Systems. 9^{th} Asian Conference (ACIIDS 2017)*. Kanazawa, Japan, April 3–5, 2017. Proceedings, P. II / Eds. by N. T. Nguyen, S. Tojo, L. M. Nguyen, B. Trawinski, 2017, vol. 10192 of Lecture Notes in Computer Science, pp. 691–700. https://doi.org/10.1007/978-3-319-54430-4_66

Received: August 21, 2021.
Accepted: February 01, 2022.

Authors' information:

Vladimir V. Mazalov — Dr. Sci. in Physics and Mathematics, Professor; vmazalov@krc.karelia.ru

Vitalia A. Khitraya — Postgraduate Student, Senior Lecturer; dobvitalia@yandex.ru

Andrei V. Khitryi — Postgraduate Student; andrey.khitryy@gmail.com