

SPECTRAL ANALYSIS OF THE JOURNAL CITATION NETWORK

S. V. Bredikhin, V. M. Lyapunov, N. G. Shcherbakova

Institute of Computational Mathematics and Mathematical Geophysics SB RAS,
630090, Novosibirsk, Russia

In this paper we investigate methods of spectral clustering for analysis of the journal citation networks. Clustering problem is reduced to min-cut graph partitioning: to find a partition of the graph such that the edges between different groups have very low weights and the edges within a group have high weights. That means that objects in different clusters are dissimilar from each other and objects within the same cluster are similar to each other, see C. J. Alpert, S.-Z. Yao (1995). Graph partitioning problems can be solved exactly in polynomial time, so for practical applications approximate solution methods have been developed. One of the widely used is the spectral partitioning method. The spectral methods usually involve taking the eigenvectors of some matrix based on relations between data elements. Most spectral clustering algorithms cluster the data with the help of eigenvectors of graph Laplacian matrices.

We study two major versions of spectral clustering, so called “unnormalized” and “normalized” spectral clustering that reveal the relationship of the object function formulation and the matrix used in the eigenvalue equation. Unnormalized spectral bi-clustering algorithms use the Laplacian matrix $L = D - A$ for solving the problem $L\mathbf{v} = \lambda\mathbf{v}$ and assigning vertices to clusters according to the signs of elements of the eigenvector \mathbf{v} corresponding to the second smallest eigenvalue. The simplified versions of the unnormalized spectral bi-clustering method is presented as the techniques of the consistency confirmation of the approach. As shown in M. E. J. Newman, M. Girvan (2004) this class of spectral clustering is only consistent under strong additional assumptions, which are not always satisfied in real data. Most of normalized spectral bi-clustering algorithms use the symmetric normalized Laplacian matrix $L^{sym} = D^{-1/2}LD^{-1/2}$ for these purposes, see J. Shi, J. Malik (2000). As shown in M. Meila, J. Shi (2001) the same results can be obtained by using the largest eigenvector of the matrix $P = D^{-1}A$. Spectral k -way clustering uses not only the second but also the next few eigenvectors to construct a partition.

The journal citation network on study is built on the basis of the bibliographic information extracted from the DB *RePEc*. The main component of the corresponding weighted digraph G has 1729 vertices (journals) and 135702 arcs (citations). We analyze the work of two spectral clustering algorithms in the context of three versions of transformation of digraph G to an undirected form. So, we examine the graphs represented by matrices $A + A^T$ (graph G^U), AA^T (graph G^{bib}) and $A^T A$ (graph G^{coc}), where A is the journal-journal citation matrix. Algorithm *WTR* P. Pons, M. Latapy (2005) is the agglomerative algorithm based on random walk matrix $P = D^{-1}A$. Algorithm *LEV* M. E. J. Newman (2006) is the bi-clustering algorithm based on the modularity matrix. The algorithms are implemented with use of the *igraph* packet (*C* library). We use *NMI*, *RAND*, *ADJUSTED_RAND* indexes as the measures of similarity of two data clusterings. For G^U clustering the similarity is low, as an example *ADJUSTED_RAND* = 0,07. The most similarity is reached for graph G^{bib} . *WTR* clusters of small size (less than 200) can be interpreted in terms of thematic fields. The results are presented in the tables (1–6). We can see that results strongly depend on the digraph transformation and the algorithm used.

Key words: journal citation network, co-citation network, bibliographic coupling network, weighted directed graph, graph partitioning, spectral clustering.

References

1. BRANDES U., GAERTLER M., WAGNER D. Experiments on graph clustering algorithms // Proc. of the 11th Annual European Symposium on Algorithms (ESA '03). 2003. P. 568–579.
2. ALPERT C. J., YAO S.-Z. Spectral partitioning: The more eigenvectors, the better // Proc. of the 32nd annual ACM/IEEE Design Automation Conference. 1995. P. 195–200.
3. GAREY R. R., JOHNSON D. S. Computers and intractability: A guide to the theory of NP-completeness. 1990. NY: W. H. Freeman & Co.
4. GOULD P. The Geographical Interpretation of Eigenvalues // Institute of British Geographers Transactions. 1967. V. 42. P. 53–85.
5. DONATH W. E., HOFFMAN A. Algorithms for partitioning of graphs and computer logic based on eigenvectors of connection matrices // IBM Technical Disclosure Bulletin. 1972. V. 15, iss. 3. P. 938–944.
6. BARNES E. An Algorithm for Partitioning the Nodes of a Graph // SIAM J. Alg. Disc. Math. 1982. V. 3, iss. 4. P. 541–550.
7. SARKAR S., BOYER K. L. Quantitative measures of change based on feature organization: Eigenvalues and eigenvectors // Computer Vision and Image Understanding. 1998. V. 71, N1. P. 110–136.
8. HALL K. M. An r -dimensional Quadratic Placement Algorithm // Management Science. 1970. V. 17. P. 219–229.
9. MOHAR B. The Laplacian Spectrum of Graphs. Graph Theory & Application. Wiley, 1991. P. 871–898.
10. NEWMAN M. E. J. Finding community structure using the eigenvectors of matrices // Phys. Rev. E 74, 036104 (2006).
11. WEST D. B. Introduction to Graph Theory. Prentice Hall, 1996.
12. DONETTI L., MUÑOZ A. Detecting network communities: a new systematic and efficient algorithm // J. of Statistical Mechanics. 2004. P. 10012.
13. BARNARD S., POTHEN A., SIMON H. A spectral algorithm for envelope reduction of sparse matrices // Numer. Linear Algebra Appl. 1995. V. 2. P. 317–334.
14. GUATTERY S., MILLER G. On the quality of spectral separators // SIAM J. Matrix Anal. Appl. 1998. V. 19. P. 701–719.
15. WEI Y.-C., CHENG C.-K. Toward efficient hierarchical designs by ratio cut partitioning // Proc. of the IEEE International Conference on Computer Aided Design. 1989. P. 298–301.
16. SHI J., MALIK J. Normalized cut and image segmentation // IEEE Transactions on Pattern Analysis and Machine Intelligence. 2000. V. 22, iss. 8. P. 888–905.
17. MEILA M., SHI J. A random walks view of spectral segmentation // Proc. International Workshop on AI and Statistics (AISTAT) 2001. [Electron. resource]. <https://dblp1.uni-trier.de/db/conf/aistats/aistats2001.html>.
18. ROSVALL M., BERGSTROM C. T. Maps of random walks on complex networks reveal community structure // Proc. Natl. Acad. Sci. USA. 2008. V. 105, N 4. P. 1118–1123.
19. PONS P., LATAPY M. Computing communities in large networks using random walks // J. of Graph Algorithms and Applications. 2006. V. 10, N 2. P. 191–218.
20. WARD J. H. Hierarchical grouping to optimize an objective function // J. of the American Statistical Association. 1963. V. 58, N 301. P. 236–244.
21. NEWMAN M. E. J., GIRVAN M. Finding and evaluating community structure in networks // Phys. Rev. 2004. E 69 (2) 026113.

22. RAGHAVAN U. N., ALBERT R., KUMARA S. Near linear time algorithm to detect community structures in large-scale networks // *Phys. Rev. E* 76, 036106.
23. BRANDES U., DELLING D., GAERTLER M., GORKE R., HOEFER M., NIKOLOSKI Z., WAGNER D. On Modularity Clustering // *IEEE Transactions on Knowledge and Data Engineering*. 2008. V. 20, iss. 2. P. 172–188.
24. BLONDEL V., GUILLAUME J., LAMBIOTTE J., LEFEBVRE E. Fast unfolding of communities in large networks // *J. Stat. Mech.* 2008, P10008.
25. CHUNG F., LU L. Connected components in random graphs with given degree sequences // *Annals of Combinatorics*. 2002. V. 6. P. 125–145.
26. LUCZAK T. Sparse random graphs with a given degree sequence // *Proc. of the Symposium on Random Graphs*. Poznac, 1989. NY: John Wiley, 1992. P. 165–182.
27. MOLLOY M., REED B. A critical point for random graphs with a given degree sequence // *Random Structures and Algorithms*. 1995. V. 6. P. 161–179.
28. REPEC. General principles. [Electron. resource]. <http://repec.org/>.
29. BREDIKHIN S. V., LYAPUNOV V. M., SCHERBAKOVA N. G. Cluster Analysis of the Citation Network of Scientific Journals // *Problemy informatiki*. 2017. N 3. P. 38–52.
30. IGRAPH – The network analysis package. [Electron. resource]. <http://igraph.org/c/doc/ix01.html>.
31. FRED A. L. N., JAIN A. K. Robust data clustering // *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR, Minneapolis (USA), June 16–22, 2003*. P. 128–136.
32. RAND W. M. Objective criteria for the evaluation of clustering methods // *J. Amer. Statistical Association*. 1971. V. 66, N 336. P. 846–850.
33. HUBERT L., ARABIE P. Comparing partitions // *J. Classification*. 1985. V. 2, iss. 1. P. 193–218.
34. JEL classification system / *EconLit Subject Descriptors*. 2016. [Electron. resource]. <https://www.aeaweb.org/econlit/jelCodes.php?view=jel>.

СПЕКТРАЛЬНЫЙ АНАЛИЗ СЕТИ ЦИТИРОВАНИЯ НАУЧНЫХ ЖУРНАЛОВ

С. В. Бредихин, В. М. Ляпунов, Н. Г. Щербакова

Институт вычислительной математики и математической геофизики СО РАН,
630090, Новосибирск, Россия

УДК 001.12+303.2

Исследуются спектральные методы анализа сети научных публикаций, организованной на отношении цитирования и представленной оргграфом $G^D = (V, E)$. Сравниваются результаты работы двух спектральных алгоритмов кластеризации. Оргграф G^D преобразуется в три неориентированных графа: $A + A^T$ (граф G^U), $A \times A^T$ (граф G^{bib}) и $A^T \times A$ (граф G^{coc}); здесь A — матрица смежности G^D . Кластеризации графов G^U , G^{bib} и G^{coc} выполнены с помощью алгоритмов *WTR* и *LEV*. Агломеративный алгоритм *WTR* основан на матрице случайного блуждания $P = D^{-1}A$, алгоритм бикластеризации *LEV* — на матрице модульности. Для сравнения результатов разбиения используются индексы *NMI*, *RAND*, *ADJUSTED_RAND*. В результате исследования выявлена зависимость результатов кластеризации от способа приведения G^D к неориентированному виду; кластеры журналов, построенные с помощью алгоритма *WTR*, могут быть проинтерпретированы в терминах принадлежности к тематическим областям. Результаты представлены в виде таблиц.

Ключевые слова: сеть цитирования журналов, сеть коцитирования, сеть библиографического сочетания, взвешенный ориентированный граф, разбиение графа, спектральная кластеризация.

Введение. Рассматривается задача разбиения множества вершин неориентированного связного графа $G = (V, E)$ на непересекающиеся подмножества (кластеры) $C^k = \{C_1, \dots, C_k\}$, такие что число ребер внутри кластеров велико по сравнению с числом ребер между кластерами. Кластер C_h можно отождествлять с индуцированным подграфом $G[C_h] = (C_h, E(C_h))$, где $E(C_h) := \{(u, v) \in E \mid u, v \in C_h\}$. Тогда $E(C) := \cup_h E(C_h)$ ($h = 1, \dots, k$) является множеством внутрикластерных ребер, а $\neg E(C) := E - E(C)$ — множеством межкластерных ребер. Разбиение на два кластера $C^2 = \{C_1, C_2\}$, $C_2 = V - C_1$ называется разрезом, а число межкластерных ребер $Cut(C_1, C_2)$ — размером разреза. Разрез с минимальным значением Cut называется минимальным разрезом [1]. Если граф взвешенный, то Cut определяется как сумма весов межкластерных ребер. Разбиение на два кластера будем называть бикластеризацией. По аналогии разбиение на k кластеров с минимальным значением Cut будем называть минимальным k -разрезом или k -кластеризацией.

Если задачу разбиения множества вершин G на кластеры представить как задачу определения минимального разреза, то решение будет тривиальным, поскольку Cut минимально, если поместить все вершины в один кластер. Отсюда вытекает требование к числу кластеров. Кроме того, нежелательно сокращать размер разреза за счет деления на небольшие кластеры. Поэтому предлагаемые алгоритмы разбиения должны балансировать между размером разреза и числом кластеров.

Решение подобных задач опирается на результаты анализа матрицы смежности G , для изучения структурных свойств которой, как правило, используются спектральные методы. Пусть $A = (A_{ij})$ — матрица смежности неориентированного графа $G = (V, E)$, $|V| = n$ (вершины занумерованы), $|E| = m$. Заданы желаемое число кластеров k и верхняя и нижняя границы размеров кластеров: $\forall (h, 1 \leq h \leq k) L_h \leq |C_h| \leq W_h$. Требуется найти сбалансированное разбиение C^k , при котором размеры кластеров находятся в нужных границах и которое минимизирует функцию

$$Cut = f(C^k) = 1/2 \sum_{h=1}^k E_h, \text{ где } E_h = \sum_{v_i \in C_h} \sum_{v_j \notin C_h} A_{ij}, \quad (1)$$

т. е. E_h — сумма весов ребер разреза, соответствующих кластеру C_h [2] (множитель $1/2$ возникает из-за того, что каждое ребро учитывается дважды). Задача нахождения сбалансированного разбиения является NP-полной [3]. Использование алгоритмов полиномиальной сложности неприемлемо для больших сетей, поэтому рассматриваются приближенные методы решения задачи.

К задаче (1) для неориентированного графа может быть сведена любая задача кластеризации множества объектов x_1, \dots, x_n , для которых установлена мера подобия $s_{ij} = s(x_i, x_j)$, согласно некоторой функции подобия, являющаяся симметричной и неотрицательной. Соответствующая матрица подобия $S = (s_{ij})$. Представление данных в форме “графа подобия”: объекту x_i соответствует вершина v_i (считаем, что вершины занумерованы и далее пользуемся номерами вершин), если $s_{ij} > 0$, то ребро (i, j) имеет вес s_{ij} . Взвешенную матрицу смежности соответствующего графа обозначаем A и далее рассматриваем задачу кластеризации как задачу разбиения графа. Аналогично, к задаче (1) для орграфа может быть сведена задача кластеризации объектов x_1, \dots, x_n , для которых установлено несимметричное отношение.

1. Кластеризация на основе матрицы Лапласа. Существуют различные варианты спектральной кластеризации, основанные на вычислении собственных векторов матрицы смежности [4–7]. Использование матрицы Лапласа обусловлено рядом ее важных свойств, отражающих структурные особенности графа [8, 9].

Рассмотрим задачу кластеризации $C^2 = \{C_1, C_2\}$ вершин неориентированного невзвешенного связного графа $G = (V, E)$ с $(0, 1)$ матрицей смежности A . Воспользуемся рассуждениями, приведенными в работе [10]. Здесь и далее предполагаем, что граф не имеет кратных ребер и петель. Разбиение можно представить с помощью индикаторного вектора (индекс-вектора) \mathbf{s} с элементами

$$s_i = \begin{cases} +1, & \text{если } i \in C_1 \\ -1, & \text{если } i \in C_2. \end{cases} \quad (2)$$

Заметим, что $\mathbf{s}^T \mathbf{s} = n$. Тогда

$$\frac{1}{2}(1 - s_i s_j) = \begin{cases} 1, & \text{если } (i \in C_1 \text{ and } j \in C_2) \text{ or } (i \in C_2 \text{ and } j \in C_1), \\ 0, & \text{если } (i, j \in C_1) \text{ or } (i, j \in C_2). \end{cases} \quad (3)$$

Таким образом, (1) можно представить в виде

$$Cut = \frac{1}{4} \sum_{i,j} (1 - s_i s_j) A_{ij}.$$

Определим степень вершины i :

$$deg_i = \sum_j A_{ij}.$$

Представим сумму элементов матрицы с учетом того, что $s_i^2 = 1$, как

$$\sum_{i,j} A_{ij} = \sum_i deg_i = \sum_i s_i^2 deg_i = \sum_{i,j} s_i s_j deg_i \delta_{ij},$$

где $\delta_{ij} = 1$, если $i = j$, и $\delta_{ij} = 0$ в противном случае. Тогда

$$Cut = \frac{1}{4} \sum_{i,j} s_i s_j (deg_i \delta_{ij} - A_{ij}).$$

Отсюда

$$Cut = \frac{1}{4} \mathbf{s}^\top L \mathbf{s}, \quad (4)$$

где L — симметричная матрица с элементами $L_{ij} = deg_i \delta_{ij} - A_{ij}$, т. е.

$$L_{ij} = \begin{cases} deg_i, & i = j, \\ -1 & i \neq j \text{ и } A_{ij} = 1, \\ 0 & \text{в противном случае.} \end{cases} \quad (5)$$

Таким образом, L — матрица Лапласа ($L = D - A$, где D — диагональная матрица, $D_{ii} = deg_i$).

Представим \mathbf{s} как сумму нормализованных собственных векторов \mathbf{v}_i матрицы L :

$$\mathbf{s} = \sum_{i=1}^n \alpha_i \mathbf{v}_i,$$

где $\alpha_i = \mathbf{v}_i^\top \mathbf{s}$, так как $\mathbf{s}^\top \mathbf{s} = n$, имеем:

$$\sum_{i=1}^n \alpha_i^2 = n. \quad (6)$$

Отсюда

$$Cut = \sum_i \alpha_i \mathbf{v}_i^\top L \sum_j \alpha_j \mathbf{v}_j = \sum_{i,j} \alpha_i \alpha_j \lambda_j \delta_{ij} = \sum_i \alpha_i^2 \lambda_i, \quad (7)$$

где λ_i — собственное значение матрицы L , соответствующее собственному вектору \mathbf{v}_i с учетом того, что $\mathbf{v}_i^\top \mathbf{v}_j = \delta_{ij}$ (если векторы нормализованы, то $\mathbf{v}_i^\top \mathbf{v}_i = 1$). Будем считать, что собственные значения упорядочены по убыванию:

$$\lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_n; \quad (8)$$

а $\mathbf{v}_1, \dots, \mathbf{v}_n$ — соответствующие собственные векторы.

Теперь задача минимизации Cut (1) может быть сформулирована как задача выбора неотрицательных величин α_i^2 (см. (7)) так, чтобы “большие” значения соответствовали “малым” собственным значениям, а “малые” — “большим”, и в то же время выполнялось (6).

Сумма каждой строки матрицы L равна нулю:

$$\sum_j L_{ij} = \sum_j (deg_i \delta_{ij} - A_{ij}) = deg_i - deg_i = 0.$$

Таким образом, вектор $\mathbf{v} = (1, 1, \dots, 1)$ является собственным вектором матрицы L , соответствующим собственному значению ноль. У матрицы L все собственные значения неотрицательны, поэтому значение ноль является минимальным, $\lambda_1 = 0$. Такой выбор эквивалентен помещению всех вершин в один кластер, но это решение не представляет интереса.

Рассмотрим вариант деления на два кластера размеров n_1 и n_2 , тогда:

$$\alpha_1^2 = (\mathbf{v}_1^\top \mathbf{s})^2 = \frac{(n_1 - n_2)^2}{n}.$$

Поскольку нет возможности варьировать этот коэффициент, то обращаем внимание на другие элементы суммы (7). Если нет другого ограничения на \mathbf{s} , кроме $\mathbf{s}^\top \mathbf{s} = n$, то Cut может быть минимизирован выбором \mathbf{s} , пропорциональным второму минимальному собственному вектору \mathbf{v}_2 матрицы L . Граф связный, поэтому λ_2 (8) отлично от нуля (так как кратность значения ноль соответствует числу связных компонент [11]). Тогда (7) зависит от λ_2 , остальные члены суммы будут равны нулю ввиду ортогональности собственных векторов L . Согласно (2) элементы \mathbf{s} равны либо $+1$, либо -1 , т. е. вектор \mathbf{s} может быть не параллелен \mathbf{v}_2 . Путем аппроксимации можно определить \mathbf{s} так, чтобы вектор был “почти” параллелен \mathbf{v}_2 . Это эквивалентно максимизации модуля:

$$|\mathbf{v}_2^\top \mathbf{s}| = \left| \sum_i v_2(i) s_i \right| \leq \sum_i |v_2(i)|, \quad (9)$$

где $v_2(i)$ — элемент i вектора \mathbf{v}_2 . Неравенство (9) следует из неравенства треугольника, равенство достигается, если все члены первой суммы имеют одинаковый знак. Другими словами, максимум $|\mathbf{v}_2^\top \mathbf{s}|$ достигается, когда $v_2(i) s_i \geq 0$ для всех i , что эквивалентно тому, что множители имеют один знак:

$$s_i = \begin{cases} +1, & \text{если } v_2(i) \geq 0, \\ -1, & \text{если } v_2(i) < 0. \end{cases}$$

В случае произвольного размера кластеров деление производится на основании знаков элементов вектора \mathbf{v}_2 . Для кластеров фиксированного размера n_1, n_2 следует упорядочить элементы \mathbf{v}_2 от наибольших положительных до наименьших отрицательных и распределить согласно размерам.

Базовую схему спектральной бикластеризации S_1 можно представить в виде:

Шаг 1 S_1 . Построить матрицу Лапласа L (см. (5)).

Шаг 2 S_1 . Найти второй минимальный собственный вектор, являющийся решением уравнения $L\mathbf{v} = \lambda\mathbf{v}$.

Шаг 3 S_1 . $C_1 = \{i; v_2(i) \geq 0\}; C_2 = \{i; v_2(i) < 0\}$.

Увеличения числа кластеров можно достичь итерационным процессом.

Схему S_2 спектральной k -кластеризации можно представить следующим образом:

Шаг 1 S_2 . Построить матрицу Лапласа L .

Шаг 2 S_2 . Вычислить k первых собственных векторов $\mathbf{v}_1, \dots, \mathbf{v}_k$ матрицы L .

Шаг 3 S_2 . Построить матрицу $M \in \mathbb{R}^{n \times k}$, в качестве столбцов которой выступают векторы $\mathbf{v}_1, \dots, \mathbf{v}_k$.

Шаг 4 S_2 . Построить $y_i \in \mathbb{R}^k (i = 1, \dots, n)$ — вектор, соответствующий i -й строке матрицы M . Применить алгоритм кластеризации (например, k -средних) к точкам y_i (соответствующим вершинам) в пространстве \mathbb{R}^k для получения кластеров C_1, \dots, C_k . Подобная схема использовалась, например, в работах [12, 13].

2. Нормализованная кластеризация. Нормализация разреза является одним из способов предотвратить тенденцию деления графа на небольшие изолированные кластеры. Существуют несколько приемов нормализации, например, в работе [14] представлено отношение $\frac{Cut(C_1, C_2)}{\min(|C_1|, |C_2|)}$, в работе [15] — отношение $\frac{Cut(C_1, C_2)}{|C_1| \times |C_2|}$. В работе [16] предложен нормализованный разрез:

$$Ncut(C_1, C_2) = \left(\frac{1}{vol(C_1)} + \frac{1}{vol(C_2)} \right) Cut(C_1, C_2),$$

где $vol(C) = \sum_{i \in C} deg_i, C \subset V$. Разрез $Ncut$ оценивает доли размера разреза относительно всех связей каждого кластера.

Кластеризация производится на основании второго минимального вектора, являющегося решением обобщенного уравнения:

$$L\mathbf{v} = \lambda D\mathbf{v}. \quad (10)$$

Уравнение (10) приводится к стандартному виду $D^{-1/2}LD^{-1/2}\mathbf{v} = \lambda\mathbf{v}$, матрица $L^{sym} = D^{-1/2}LD^{-1/2}$ называется симметрично-нормализованной матрицей Лапласа. В работе [16] утверждается, что если λ^L — второе минимальное (8) собственное значение L^{sym} , а \mathbf{v}^L — соответствующий ему собственный вектор и существует разбиение $C^2 = \{C_1, C_2\}$, такое что

$$v_i^L = \begin{cases} \alpha, & \text{если } i \in C_1, \\ \beta, & \text{если } i \in C_2, \end{cases}$$

то разбиение оптимальное и $Ncut(C_1, C_2) = \lambda^L$. Таким образом, схема нормализованной бикластеризации совпадает со схемой S_1 за тем исключением, что вместо матрицы L используется матрица L^{sym} .

Схема S_3 нормализованной k -кластеризации аналогична схеме S_2 и отличается тем, что k минимальных собственных векторов являются решением обобщенного уравнения (10).

В работе [17] рассматривается интерпретация нормализованной спектральной кластеризации с точки зрения дискретного случайного блуждания. Если нормализовать взвешенную матрицу A по строкам, то получим стохастическую матрицу:

$$P = D^{-1}A, \quad (11)$$

которая характеризует Марковский процесс случайного блуждания, P_{ij} — вероятность перехода из вершины i в вершину j за один шаг. Предполагается, что все вершины имеют ненулевые взвешенные степени. Пусть

$$\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n \quad (12)$$

упорядоченные по невозрастанию собственные значения матрицы P , а $\mathbf{v}_1 = \mathbf{1}, \mathbf{v}_2, \dots, \mathbf{v}_n$ — соответствующие им собственные векторы, т. е. решения уравнения

$$P\mathbf{v} = \lambda\mathbf{v}. \quad (13)$$

В работе [17] показано, что если λ, \mathbf{v} являются решением (13) и $P = D^{-1}A$, то пара $((1-\lambda), \mathbf{v})$ является решением (10). Другими словами, спектральная проблема, сформулированная алгоритмом *Ncut*, и проблема собственных значений/векторов стохастической матрицы P эквивалентны. При этом минимальный вектор для (10) соответствует максимальному вектору (13). Там же приведена схема S_4 нормализованной k -кластеризации на основе матрицы случайного блуждания:

Шаг 1 S_4 . Построить матрицу P .

Шаг 2 S_4 . Вычислить k наибольших собственных векторов $\mathbf{v}_1, \dots, \mathbf{v}_k$ матрицы P .

Шаг 3 S_4 . Построить матрицу $M \in \mathbb{R}^{n \times k}$, в качестве столбцов которой выступают векторы $\mathbf{v}_1, \dots, \mathbf{v}_k$.

Шаг 4 S_4 . Построить $\mathbf{y}_i \in \mathbb{R}^k$ ($i = 1, \dots, n$) — вектор, соответствующий i -й строке матрицы M .

Шаг 5 S_4 . Применить алгоритм кластеризации, например k -средних, к \mathbf{y}_i для получения кластеров C_1, \dots, C_k .

Спектральная кластеризация способна группировать вершины согласно подобию вероятностей переходов между подмножествами вершин. Т.е. если множество V разделено на две части, то случайное блуждание, начавшееся в одной из частей, имеет тенденцию остаться в ней. Связь стационарного распределения с кластеризацией используется в работе [18]. Стационарное распределение рассматривается как показатель частоты посещения каждой вершины графа.

Матрица P (11) используется также в механизме k -кластеризации, предложенном в статье [19]. Определяется расстояние между вершинами графа:

$$r_{ij}(t) = \sqrt{\sum_{l=1}^n \frac{(P_{il}^t - P_{jl}^t)^2}{deg_l}},$$

где P_{il}^t — вероятность перехода из вершины i в вершину l за t шагов, deg_l — степень вершины l . Показано, что если собственные значения матрицы P упорядочены согласно (12), а $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n$ — соответствующие им собственные векторы, то для достаточно больших t имеет место равенство

$$r_{ij}(t) = \sum_{l=2}^n \lambda_l^{2t} (v_l(i) - v_l(j))^2,$$

где $v_l(i)$ — элемент i вектора \mathbf{v}_l . Затем определяется r_{iC} — расстояние между вершиной i и кластером C . Строится иерархический объединяющий алгоритм, основанный на методе Уорда [20], рассматривающем в качестве кандидатов на объединение $C_3 = C_1 \cup C_2$ “близко” расположенные кластеры. В данном случае на каждом шаге в качестве таких кандидатов рассматриваются пары смежных кластеров. Минимизируется выражение

$$\frac{1}{n} \left(\sum_{i \in C_3} r_{iC_3}^2 - \sum_{i \in C_1} r_{iC_1}^2 - \sum_{i \in C_2} r_{iC_2}^2 \right).$$

Алгоритм *WTR*, применяемый в нашей работе, использует эту технику.

3. Кластеризация на основе матрицы модульности. В работе [21] задача оптимального разбиения графа рассматривается как максимизация числа ребер внутри кластеров. Понятие “модульности” базируется на предположении, что структура изучаемого графа, как правило, отличается от структуры случайного графа. Определяется функция

$$Q = N_1 - N_2, \quad (14)$$

где N_1 — число ребер внутри кластеров, N_2 — ожидаемое число таких ребер. Функция Q названа “модульностью”, большие значения указывают на тесную взаимосвязь внутри кластеров. Вычисление модульности де-факто является способом проверки качества разбиения вершин графа, см. [19, 22–24].

Ожидаемое число ребер N_2 вычисляется согласно “нуль-модели” графа, который имеет то же число вершин и может быть разделен на то же число кластеров, что и анализируемый граф. Вероятность наличия ребра (i, j) в модели обозначим P_{ij} . Формула (14) может быть представлена в виде:

$$Q = \frac{1}{2m} \sum_{i,j} (A_{ij} - P_{ij}) \delta(C_i, C_j), \quad (15)$$

где C_i — кластер, в который попадает вершина i , $\delta(r, s) = 1$, если $r = s$. Выбор P_{ij} ограничен по следующим причинам: во-первых, поскольку рассматривается неориентированный граф, то $P_{ij} = P_{ji}$; во-вторых, предполагается, что $Q = 0$, если все вершины попадают в один кластер, т. е.

$$\sum_{i,j} (A_{ij} - P_{ij}) = 0.$$

Отсюда

$$\sum_{i,j} P_{ij} = \sum_{i,j} A_{ij} = 2m. \quad (16)$$

Ожидаемая степень вершины i задается выражением $\sum_j P_{ij}$; в предположении, что ожидаемая степень вершины приближается к реальной, имеем:

$$\sum_j P_{ij} = \text{deg}_i.$$

Далее в качестве нуль-модели рассматривается модель случайного графа с фиксированной степенной последовательностью (см. [25–27]):

$$P_{ij} = \frac{\text{deg}_i \text{deg}_j}{2m}.$$

Рассмотрим задачу бикластеризации. Согласно (2), (3) имеем:

$$\delta(C_i, C_j) = \frac{1}{2}(s_i s_j + 1).$$

Теперь (15) можно представить в виде:

$$Q = \frac{1}{4m} \sum_{i,j} (A_{ij} - P_{ij})(s_i s_j + 1) = \frac{1}{4m} \sum_{i,j} (A_{ij} - P_{ij}) s_i s_j.$$

Второе равенство вытекает из (16). В матричной форме:

$$Q = \frac{1}{4m} \mathbf{s}^\top B \mathbf{s}, \quad (17)$$

где B — симметричная матрица с элементами:

$$B_{ij} = A_{ij} - P_{ij}.$$

Матрица B называется матрицей модульности. В работе [10] представлен метод спектральной кластеризации на основе матрицы B , которая выполняет ту же роль, что и матрица Лапласа: уравнение (17) соответствует (4). Показано, что на основе собственного вектора, соответствующего наибольшему положительному собственному значению матрицы B , можно произвести бикластеризацию согласно знакам элементов собственного вектора.

4. Вычислительный эксперимент. Выше была представлена методика спектрального анализа, послужившая основой для алгоритмов кластеризации WTR и LEV . Цель эксперимента состояла в сравнении результатов работы указанных алгоритмов по выявлению сообществ в множестве научных журналов, размещенных в одной библиографической базе данных.

В качестве исходных данных выступает сеть цитирования журналов, построенная на основе информации о цитировании, извлеченной из БД *RePEc* [28]. Сеть представляется взвешенным оргграфом без кратных ребер и петель. Анализируется главная компонента $G = (V, E)$, $|V| = 1729$, $|E| = 135702$, с взвешенной матрицей смежности A . Поскольку алгоритмы предназначены для неориентированных графов, преобразуем оргграф G в неориентированный тремя способами (см. [29]). В результате были получены три неориентированных графа, которые были использованы в эксперименте:

а) граф G^U , представлен матрицей смежности $A + A^\top$, $|E(G^U)| = 116190$ (пары разнонаправленных дуг заменяются одним ребром с суммарным весом);

б) граф G^{bib} , соответствующий “сети библиографического сочетания”, представлен нормализованной матрицей AA^\top , без учета одиночных вершин $|V(G^{bib})| = 1432$, $|E(G^{bib})| = 844476$;

в) граф G^{coc} , соответствующий “сети коцитирования”, представлен нормализованной матрицей $A^\top A$, без учета одиночных вершин $|V(G^{coc})| = 1582$, $|E(G^{coc})| = 820982$.

Алгоритм WTR реализует метод, предложенный в статье [19]. Работа начинается с разбиения $R_1 = \{\{v\}, v \in V\}$, каждая вершина является кластером. Вычисляются все расстояния между всеми смежными вершинами. На шаге k :

1. Выбираются два кластера C_1, C_2 , слияние которых приводит к минимальному увеличению целевой функции (27);

2. Образуется новый кластер $C_3 = C_1 \cup C_2$ и новое разбиение $R_{k+1} = (R_k \setminus \{C_1, C_2\}) \cup \{C_3\}$;

3. Обновляются расстояния между смежными кластерами.

На шаге $n - 1$ алгоритм заканчивает работу, при этом $R_n = \{V\}$. На каждом шаге вычисляется Q (15). Лучшим считается разбиение с максимальным значением Q . Сложность вычислений оценивается как $\mathcal{O}(|E||V|^2)$, для разреженного графа — $\mathcal{O}(|V|^2 \log |V|)$.

Таблица 1

Размеры кластеров графа G^U

WTR		LEV	
#Cl	#J	#Cl	#J
1	692	1	1549
1	280	1	180
1	270		
1	183		
1	76		
1	62		
1	44		
2	8		
1	4		
1	2		

Таблица 2

Тематика кластеров графа G^U , алгоритм WTR

#J	Тематика	Коды Jel
183	Сельское хозяйство, ресурсы, экология	Q
76	Здравоохранение, соц. обеспечение	I
62	Транспортная экономика, администрирование	R
44	Эконометрические и статистические методы	C1
8	Образование	A2, I2

Алгоритм *LEV* является реализацией метода, предложенного в статье [10]. Это итерационный процесс деления множества вершин на две части. На шаге k :

1. Для графа/выбранного подграфа строится матрица модульности B ;
2. Вычисляется вектор, соответствующий наибольшему по абсолютной величине собственному значению; если это значение β_n положительно, то собственный вектор искомый; если отрицательно, то повторяем вычисление для матрицы $B - \beta_n I$ (где I — единичная матрица);
3. Кластер делится на две части соответственно знакам элементов найденного собственного вектора.

4. Проверяется, увеличилось ли значение модульности исходного графа; если да, то считаем разделение правомерным и переходим к шагу 1.

Алгоритм *LEV* заканчивает работу, если значение Q не увеличивается или ни один кластер невозможно разделить на две части. Кластер невозможно разделить на части, если все собственные значения, кроме нулевого, отрицательны. Сложность *LEV* оценивается как $\mathcal{O}(|E| + |V|^{2 \times \text{steps}})$, где *steps* — число шагов деления на два сообщества. Алгоритмы *WTR* и *LEV* реализованы с помощью библиотеки *C* пакета *igraph* [30].

4.1. *Кластеризация G^U* . В табл. 1 приведены размеры кластеров, полученных в результате исполнения алгоритмов *WTR* и *LEV* в применении к графу G^U . Здесь и далее одновершинные кластеры не включаются в таблицы, так, при кластеризации *WTR* число одновершинных кластеров равно ста; $\#Cl$ — число кластеров размера $\#J$. Для сравнения результатов кластеризации использовались три индекса согласованности: *NMI* [31], *RAND* [32] и *ADJUSTED_RAND (ARI)* [33]. Индексы согласованности выглядят так: *NMI* = 0,12; *RAND* = 0,38; *ARI* = 0,07. Очевидно, что сходство минимальное.

При кластеризации с помощью алгоритма *LEV* 89,6 % журналов (1549) попали в один кластер, что не позволяет интерпретировать результаты. Однако с помощью алгоритма *WTR* были выделены тематические области (табл. 2.) на основе ключевых слов в названиях журналов и классификатора *Jel* [34]. Интерпретируемым результатом будем считать

Таблица 3

Размеры кластеров графа G^{bib}

WTR		LEV	
#Cl	#J	#Cl	#J
1	718	1	486
1	369	1	339
1	288	1	316
1	30	1	291
1	25		

Таблица 4

Тематика кластеров графа G^{bib} , алгоритм WTR

#J	Тематика	Коды Jel
369	Математические методы, бизнес-экономика, экономическое развитие, ресурсы, транспортная экономика	С, М, О, Q, R
288	Общая экономика, математические методы, финансовая экономика	А, С, G
30	Математические методы	С
25	Математические методы, организация производства	С, L

кластеризацию, которая позволяет соотнести кластер с 1–2 тематическими областями. Несмотря на различные результаты, 69% журналов, входящих в WTR кластер, имеющий размер 183 (см. табл. 1, столбец 2, строка 4) входят в LEV кластер, имеющий размер 180 (см. табл. 1, столбец 4, строка 2). Если сравнивать алгоритмы по размеру разреза графа G^U , то в результате разбиения LEV размер разреза меньше. Такое соотношение сохраняется для нормализованных разрезов, определенных согласно [14, 16]. Нормализация согласно [15] дает меньшее значение для разбиения WTR.

4.2. *Кластеризация G^{bib}* . Размеры кластеров, полученных в результате кластеризации G^{bib} алгоритмами WTR и LEV, представлены в табл. 3. Индексы согласованности имеют вид: $NMI = 0,77$; $RAND = 0,84$; $ARI = 0,53$. Если сравнивать алгоритмы WTR и LEV по размеру разреза графа G^{bib} , то в результате разбиения WTR размер разреза меньше. Такое соотношение сохраняется для нормализованного разреза, определенного согласно [15]. Нормализация согласно [14] и [16] дает меньшее значение для разбиения LEV. Тематика WTR кластеров представлена в табл. 4.

4.3. *Кластеризация G^{coc}* . Размеры кластеров, полученных в результате кластеризации G^{coc} представлены в табл. 5. Индексы согласованности имеют вид: $NMI = 0,57$; $RAND = 0,74$; $ARI = 0,3$. Согласованность алгоритмов ниже, чем при кластеризации графа G^{bib} .

Заключение. Цель работы состояла в сравнении спектральных методов кластеризации коллекции научных журналов БД RePEc, связанных отношением цитирования. Задача кластеризации представлена как задача минимизации сбалансированного разреза соответствующего графа. Проанализирована связь определения целевой функции и матрицы, на основе собственных векторов которой достигается оптимизация, а именно, обоснованность применения матриц Лапласа, случайного блуждания и модульности.

Реализованы алгоритмы спектральной кластеризации WTR и LEV. Исследованы результаты их работы для графов G^U , G^{bib} и G^{coc} . Проведенное исследование позволяет заключить, что для рассматриваемой сети цитирования журналов больше подходит спектральная кластеризация алгоритмом WTR на основе матрицы случайного блуждания. Приемлемую согласованность рассматриваемые алгоритмы достигли при кластеризации графа G^{bib} . В свою очередь, граф G^U лучше поддается кластеризации, чем два других варианта преобразования исходного орграфа. Анализ показал, что кластеры большого

Таблица 5

Размеры кластеров графа G^{coc}

WTR		LEV	
#Cl	#J	#Cl	#J
1	517	1	675
1	350	1	459
1	255	1	448
1	235		
1	118		
1	32		
1	30		
1	9		
9	< 9		

Таблица 6

Тематика кластеров графа G^{coc} , алгоритм WTR

#J	Тематика	Коды Jel
350	Финансовая экономика, бизнес-экономика, транспортная экономика	Е, G, М, R
255	Транспортная экономика, бизнес-экономика	R, M
235	Математические методы, организация производства, экономическое развитие, ресурсы	С, L, O, Q
118	Восточная Европа	А, M
32	Румыния	
30	Математические методы	С
9	Восточная Европа	

размера ($|C_i| > 200$) не удастся классифицировать согласно 1–2 тематикам. При кластеризации графа коцитирования алгоритмом WTR, кроме тематических сообществ, выделены сообщества журналов, связанных по территориальному признаку издательств. Отметим, что большинство кластеров, кроме журналов основных тематик, содержат журналы, относящиеся к “математическим методам” (Jel код С). Выявлена зависимость результатов кластеризации от применяемого алгоритма, способа преобразования орграфа в неориентированный граф и способа нормализации размера разреза.

Список литературы

1. Brandes U., Gaertler M., Wagner D. Experiments on graph clustering algorithms // Proc. of the 11th Annual European Symposium on Algorithms (ESA'03). 2003. P. 568–579.
2. Alpert C. J., Yao S.-Z. Spectral partitioning: The more eigenvectors, the better // Proc. of the 32nd annual ACM/IEEE Design Automation Conference. 1995. P. 195–200.
3. Garey R. R., Johnson D. S. Computers and intractability: A guide to the theory of NP-completeness. 1990. NY: W. H. Freeman & Co.
4. Gould P. The Geographical Interpretation of Eigenvalues // Institute of British Geographers Transactions. 1967. V. 42. P. 53–85.
5. Donath W. E., Hoffman A. Algorithms for partitioning of graphs and computer logic based on eigenvectors of connection matrices // IBM Technical Disclosure Bulletin. 1972. V. 15, iss. 3. P. 938–944.
6. Barnes E. An Algorithm for Partitioning the Nodes of a Graph // SIAM J. Alg. Disc. Math. 1982. V. 3, iss. 4. P. 541–550.
7. Sarkar S., Boyer K. L. Quantitative measures of change based on feature organization: Eigenvalues and eigenvectors // Computer Vision and Image Understanding. 1998. V. 71, N1. P. 110–136.
8. Hall K. M. An r -dimensional Quadratic Placement Algorithm // Management Science. 1970. V. 17. P. 219–229.
9. Mohar B. The Laplacian Spectrum of Graphs. Graph Theory & Application. Wiley, 1991. P. 871–898.

10. Newman M. E. J. Finding community structure using the eigenvectors of matrices // *Phys. Rev. E* 74, 036104 (2006).
11. West D. B. *Introduction to Graph Theory*. Prentice Hall, 1996.
12. Donetti L., Muñoz A. Detecting network communities: a new systematic and efficient algorithm // *J. of Statistical Mechanics*. 2004. P. 10012.
13. Barnard S., Pothen A., Simon H. A spectral algorithm for envelope reduction of sparse matrices // *Numer. Linear Algebra Appl.* 1995. V. 2. P. 317–334.
14. Guattery S., Miller G. On the quality of spectral separators // *SIAM J. Matrix Anal. Appl.* 1998. V. 19. P. 701–719.
15. Wei Y.-C., Cheng C.-K. Toward efficient hierarchical designs by ratio cut partitioning // *Proc. of the IEEE International Conference on Computer Aided Design*. 1989. P. 298–301.
16. Shi J., Malik J. Normalized cut and image segmentation // *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2000. V. 22, iss. 8. P. 888–905.
17. Meila M., Shi J. A random walks view of spectral segmentation // *Proc. International Workshop on AI and Statistics (AISTAT) 2001*. [Electron. resource]. <https://dblp1.uni-trier.de/db/conf/aistats/aistats2001.html>.
18. Rosvall M., Bergstrom C. T. Maps of random walks on complex networks reveal community structure // *Proc. Natl. Acad. Sci. USA*. 2008. V. 105, N 4. P. 1118–1123.
19. Pons P., Latapy M. Computing communities in large networks using random walks // *J. of Graph Algorithms and Applications*. 2006. V. 10, N 2. P. 191–218.
20. Ward J. H. Hierarchical grouping to optimize an objective function // *J. of the American Statistical Association*. 1963. V. 58, N 301. P. 236–244.
21. Newman M. E. J., Girvan M. Finding and evaluating community structure in networks // *Phys. Rev.* 2004. E 69 (2) 026113.
22. Raghavan U. N., Albert R., Kumara S. Near linear time algorithm to detect community structures in large-scale networks // *Phys. Rev. E* 76, 036106.
23. Brandes U., Delling D., Gaertler M., Gorke R., Hoefer M., Nikoloski Z., Wagner D. On Modularity Clustering // *IEEE Transactions on Knowledge and Data Engineering*. 2008. V. 20, iss. 2. P. 172–188.
24. Blondel V., Guillaume J., Lambiotte J., Lefebvre E. Fast unfolding of communities in large networks // *J. Stat. Mech.* 2008, P10008.
25. Chung F., Lu L. Connected components in random graphs with given degree sequences // *Annals of Combinatorics*. 2002. V. 6. P. 125–145.
26. Luczak T. Sparse random graphs with a given degree sequence // *Proc. of the Symposium on Random Graphs*. Poznac, 1989. NY: John Wiley, 1992. P. 165–182.
27. Molloy M., Reed B. A critical point for random graphs with a given degree sequence // *Random Structures and Algorithms*. 1995. V. 6. P. 161–179.
28. RePec. General principles. [Electron. resource]. <http://repec.org/>.
29. Бредихин С. В., Ляпунов В. М., Щербакова Н. Г. Кластерный анализ сети цитирования журналов // *Проблемы информатики*. 2017. № 3. С. 38–52.
30. igraph – The network analysis package. [Electron. resource]. <http://igraph.org/c/doc/ix01.html>.
31. Fred A. L. N., Jain A. K. Robust data clustering // *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR, Minneapolis (USA), June 16–22, 2003*. P. 128–136.
32. Rand W. M. Objective criteria for the evaluation of clustering methods // *J. Amer. Statistical Association*. 1971. V. 66, N 336. P. 846–850.
33. Hubert L., Arabie P. Comparing partitions // *J. Classification*. 1985. V. 2, iss. 1. P. 193–218.
34. Jel classification system / *EconLit Subject Descriptors*. 2016. [Electron. resource]. <https://www.aeaweb.org/econlit/jelCodes.php?view=jel>.



Бредихин Сергей Всеволодович — канд. техн. наук, ведущий научный сотрудник Ин-та вычислительной математики и математической геофизики СО РАН; e-mail:

bred@nsc.ru;

Сергей Бредихин окончил механико-математический факультет Новосибирского государственного университета в 1968 г. С 1968 г. — сотрудник Института автоматики и электрометрии СОАН СССР. Кандидат технических наук с 1983 г. В период 1988–2017 гг. руководил лабораторией ИВМ и МГ СО РАН. Исполнял обязанности технического директора проекта „Сеть Интернет Новосибирского научного центра“. Лауреат государственной премии РФ по науке и технике 2012 г. Сфера научных интересов: анализ и измерение распределенных информационных сетей. Автор и соавтор более ста научных работ и двух монографий: „Методы библиометрии и рынок электронной научной периодики“, „Анализ цитирования в библиометрии“.

Sergey Bredikhin — Ph.D. of Engineering Sciences, Leading Researcher, Institute of Computational Mathematics and Mathematical Geophysics SB RAS, e-mail: bred@nsc.ru.

Sergey Bredikhin graduated from Novosibirsk State University in 1968, faculty of Mechanics and Mathematics, and became an employee of Institute of Automation and Electrometry SB RAS. In 1983 he received Ph.D degree in Engineering Science. Since 1988 he was the head of the laboratory of Computing Center (now ICM&MG) SB RAS. He was the technical manager of „Akademgorodok Internet Project“. He is the state prize winner in science and engineering RF at 2012. Since 2017 he is the leading researcher of ICM&MG. Sphere of his scientific interests — analysis and measurement of the distributed information networks. He is the author and co-author of more than hundred scientific works and two monographs: "Metody bibliometrii i rynek elektronnoy nauchnoy periodiki" "Analiz tsitirovaniya v bibliometrii".



Ляпунов Виктор Михайлович — ведущий инженер Ин-та вычислительной математики и математической геофизики СО РАН; e-mail: vic@nsc.ru;

Виктор Ляпунов окончил механико-математический факультет Новосибирского государственного университета в 1978 г. В 1978 г. стал сотрудником Вычислительного центра СОАН СССР, а с 1990 г. — сотрудником Института систем информатики СО АН СССР. С 2004 г. — ведущий инженер Института вычислительной математики и математической геофизики СО РАН. Занимается вопросами извлечения информации из баз данных и обработкой больших массивов данных. Соавтор более 10 работ в этой области.

Victor Lyapunov — Leading Software Engineer, Institute of Computational Mathematics and Mathematical Geophysics SB RAS, e-mail: vic@nsc.ru.

Victor Lyapunov graduated from Novosibirsk State University in 1978 (faculty of Mechanics and Mathematics). In 1978, he became an employee of Computing Center of SB AS USSR, since 1990 — an employee of Institute of Informatics Systems SB RAS. Since 2004 he works as software engineer in Institute of Computational Mathematics and Mathematical Geophysics SB RAS. His current research interests include methods of information extracting from databases and processing of large data sets. He is the co-author of more than 10 works in that area.



Щербакова Наталья Григорьевна — старш. науч. сотр. Ин-та вычислительной математики и математической геофизики СО РАН; e-mail: nata@nsc.ru.

Наталья Щербакова окончила Новосибирский государственный университет по специальности Математическая лингвистика в 1967 г. С 1967 г. работала в Институте математики СО АН СССР, затем — в Институте автоматики и электрометрии СО АН СССР в области создания программного обеспечения систем пе-

редачи данных. С 2000 г. — сотрудник Института вычислительной математики и математической геофизики СО РАН, где с 2002 г. занимает должность старшего научного сотрудника. Являлась участником проекта “Сеть Интернет Новосибирского научного центра”, занималась вопросами мониторинга и анализа IP-сетей. Автор и соавтор более 40 работ, соавтор монографии “Анализ цитирования в библиометрии”. Научные интересы лежат в области исследования методов оценки научной деятельности на основе анализа цитирования научной литературы.

Natalia Scherbakova — Senior Researcher, Institute of Computational Mathematics and Mathematical Geophysics SB RAS, e-mail: nata@nsc.ru.

Natalia Shcherbakova graduated from Novosibirsk State University in 1967 (mathematical linguistics). Since 1967 she worked at Institute of Mathematics SB RAS, then at Institute of Automation and Electrometry SBRAS in the field of software design for data transmission systems. In 2000 — the employee of Institute of Computational Mathematics and Mathematical Geophysics SB RAS, since 2002 works as senior researcher. She is a member of “Akademgorodok Internet Project”, dealt with software of monitoring and the analysis of IP networks. She is the author and co-author of more than 40 works, the co-author of the monograph “Ansliz tsitirovaniya v bibliometrii”. The current research interests lie in the field of bibliometrics: methods of measuring of scientific.

Дата поступления — 30.03.2018