

Лосев А. С.
A. S. Losev

КОНТЕКСТНЫЙ ПОДХОД В АНАЛИЗЕ СЕТЕВЫХ СТРУКТУР

CONTEXTUAL APPROACH TO ANALYSIS OF NETWORK STRUCTURES

Лосев Александр Сергеевич – кандидат физико-математических наук, доцент, старший научный сотрудник Института прикладной математики ДВО РАН (Россия, Владивосток); тел. 8(914)328-91-58. E-mail: A.S.Losev@yandex.ru.

Alexander S. Losev – PhD in Physics and Mathematics, Associate Professor, Senior Researcher, Institute of Applied Mathematics Far Eastern Branch of the Russian Academy of Sciences (Russia, Vladivostok); tel. 8(914)328-91-58. E-mail: A.S.Losev@yandex.ru.

Аннотация. В работе предложено решение проблемы первичной обработки сетевых структур большой размерности в условиях неопределённости. Раскрываются особенности использования контекстного подхода в структурном анализе объектов исследования, представимых в виде теоретико-графовых моделей. Получены алгоритмы кластеризации и классификации ориентированного графа, определяющие набор доминантных связей, которые отражают особенности строения графа и позволяют выбрать дальнейшие методы его обработки.

Summary. The paper proposes a solution to the problem of primary processing of high-dimensional network structures under conditions of uncertainty. The features of using the contextual approach in the structural analysis of research objects, represented in the form of graph-theoretic models, are revealed. Algorithms for clustering and classification of an oriented graph are obtained, which determine a set of dominant connections that reflect the structural features of the graph and allow the selection of further methods for its processing.

Ключевые слова: контекст, множество доминант, связность, кластеризация.

Key words: context, set of dominants, connectivity, clustering.

Работа выполнена в рамках государственного задания ИПМ ДВО РАН (№ 075-01290-23-00).

УДК 1519.178

Введение. Порядок измерения объёма передаваемой информации растёт геометрически, суточный интернет-трафик одной рабочей станции, подключённой к сети, соизмерим с годовым трафиком двадцатилетней давности. Естественным образом возникает проблема алгоритмической обработки большого потока информации и её дальнейшего структурирования. Наиболее распространённым представлением по причине наглядности и удобства считается сетевая модель. Она широко используется в различных научных (биологических, химических, информационных, технических) и промышленных (транспорт, экономика, социология) сферах [2].

Отличительными особенностями такого представления информации являются наглядность, компактность, открытость к дополнению и уточнению. Наравне с этим существует множество современных методов обработки сетевых структур с помощью машинного обучения и искусственного интеллекта. Среди них выделяют вероятностные, логические, теоретико-графовые и искусственно интеллектуальные подходы [9]. Каждый из подходов обладает своими особенностями, требованиями к реализации и ограничениями. Но использование каждого из них требует наличия входной информации, которая уже отражает особенности строения исследуемой структуры. Например, в задачах кластеризации с помощью машинного обучения необходимы число кластеров разбиения или признаков [5; 10], плотность рассеивания или взаимное пространственное расположение объектов структуры [4], или прочие характерные свойства [3; 7; 11]. Таким образом, из-

вестные подходы в большей степени ориентированы на уточнение информации о сетевой структуре, которая зачастую является представителем знакомого класса исследуемых объектов.

Следовательно, возникает проблема выбора метода обработки сетевой структуры большой размерности в условиях неопределённости и отсутствия информации об особенностях её функционирования и строения. Подобная проблема возникает в обработке большого объёма выборок в статистических исследованиях, содержащих высокую долю случайных флуктуаций, которая решается методами первичной обработки.

В настоящей работе предлагается использовать контекстный подход в обработке и структурном анализе сетевых моделей, основанный на выделении доминантных связей. В отличие от известных подходов кластеризации и классификации, он не требует уникальной информации об объекте исследования и особенностях взаимного расположения его частей, а опирается только на информацию о наличии связей между его элементами, образующих сетевую структуру.

Основная идея предлагаемого подхода основана на работах академиков П. К. Анохина и А. А. Ухтомского. В работах П. К. Анохина [1], посвящённых исследованию особенностей функционирования систем, отдельное место занимает вопрос синергетических эффектов, возникающих в зависимости от ситуации, приводящих к проявлению тех или иных свойств системы, которые определяют особенности её строения и функционирования в отдельно взятый момент времени. Аналогичные эффекты занимают центральное место в теории доминант А. А. Ухтомского [12], по мнению которого данные эффекты определяются набором отдельных связей (рёбер) в системе (структуре). Им доказано, что набор доминант и их расположение отражают основные характеристики системы (структуры) и определяют особенности её функционирования. Наравне с этим обоими исследователями показано, что перечень этих связей не стационарен и подвержен изменению в зависимости от наблюдаемой ситуации и условий. Сравнительный анализ данных результатов позволил прийти к умозаключению, что контекст рассмотрения объекта исследования задаётся совокупностью доминантных связей, где доминантной связью является отдельно взятое ребро или соединение из системы, ничем не отличающееся по рассматриваемым характеристикам от прочих, но обладающее значительным воздействием на исследуемые характеристики системы с позиции поставленной задачи.

Основываясь на идее доминантных связей в системах, воспользуемся контекстным подходом для структурного анализа сетевых структур в предположении, что они имеют большую размерность и в качестве исходной информации об их строении известны только матрицы смежности, отражающие наличие связей между отдельными объектами сети. В общем виде под сетевой структурой будем понимать совокупность узлов и дуг, характеризующих взаимосвязь однотипных объектов.

В качестве математической модели сетевой структуры возьмём ориентированный граф $\Gamma(U, W)$ без петель с конечным множеством вершин U и рёбер W , в котором $|U| = n$, $|W| = m$. Положим, что две вершины $u, v \in U$ графа Γ циклически эквивалентны и принадлежат одному кластеру $[k]$, если существует цикл, содержащий эти вершины, т. е. ориентированный путь из u в v и обратно. На множестве всех $[U]$ классов эквивалентности, на которые разбивается множество U , введём бинарное отношение $\succsim$ такое, что $[u_i] \succsim [u_j]$, $[u_i], [u_j] \in [U]$, если существует путь из вершины класса $[u_i]$ в вершину класса $[u_j]$.

Непосредственной проверкой можно убедиться, что данное отношение является отношением частичного порядка, т. к. удовлетворяет свойствам:

1. $\forall [u_i] \in [U], [u_i] \succsim [u_i]$.
2. $\forall [u_i], [u_j], [u_l] \in [U], [u_i] \succsim [u_j] \wedge [u_j] \succsim [u_l] \Rightarrow [u_i] \succsim [u_l]$.
3. $\forall [u_i], [u_j] \in [U], [u_i] \succsim [u_j] \wedge [u_j] \succsim [u_i] \Rightarrow [u_i] \succsim [u_j]$.

Ранее в работе [18] представлен алгоритм динамического разбиения сетевой структуры на кластеры по принципу циклической эквивалентности, его работа выстроена таким образом, что он

не требует перезапуска при добавлении новой вершины в случае расширения сетевой структуры, что позволяет его использовать при изучении динамических структур. Естественным образом динамичность структуры не позволяет говорить о стационарном итоговом результате, лишь о его промежуточных позициях, тем самым отражая контекстное рассмотрение структуры на каждом шаге рассмотрения.

Входные данные алгоритма:

– матрица смежности $\|r_{ij}\|_{i,j=1}^n$ графа Γ , в которой, помимо $r_{ij} = 1$ для $(u_i, u_j) \in W$, положим $r_{ii} = 1, 1 \leq i \leq n$, где n – число вершин графа;

– матрица смежности $B = \|b([i], [j])\|_{[u_i], [v_j] \in [U]}$, характеризующая отношение частичного порядка $\succsim$ на множестве $[U]$ классов эквивалентности, в которой

$$b([i], [j]) = \begin{cases} 1, [u_i] \succsim [v_j] \\ 0, \text{ иначе} \end{cases}.$$

Основное тело алгоритма:

Шаг 1. Возьмём первую вершину $u_1 \in U$, образующую первый класс циклической эквивалентности $[1]$, добавим его во множество всех кластеров $K = \{[1]\}$ и положим $b([1], [1]) = 1$.

Шаг 2. Перебираем по очереди следующие $t-1$ вершин, полагаем, что они образуют разбиение множества вершин $\{u_1, \dots, u_{t-1}\}$ на непересекающиеся подмножества $[1], \dots, [t-1]$ и соответствующее множество кластеров K . Здесь же при обработке каждой вершины формируется матрица $B = \|b([i], [j])\|_{[u_i], [v_j] \in [U]}$, где кластер $[j] \in K$ индексируется максимальным номером вершины u_j , входящей в него.

Шаг 3. Рассмотрим вершину u_t , которая имеет связь с вершинами из имеющихся кластеров. Определим множества

$$I = \{i \in \{1, \dots, t-1\} : r_{ti} = 1\}, J = \{j \in \{1, \dots, t-1\} : r_{jt} = 1\}.$$

Здесь же каждой вершине из множеств I, J сопоставим кластер и обозначим $[I], [J]$ множества этих кластеров.

Введём в рассмотрение и определим следующие множества:

$$K_I = \bigcup_{[i] \in [I]} \{[k] \in K : b([i], [k]) = 1\}, K_J = \bigcup_{[j] \in [J]} \{[k] \in K : b([k], [j]) = 1\},$$

$$B_1 = K_I \cap K_J, B_2 = K_I \setminus B_1, B_3 = K_J \setminus B_1, B_4 = K \setminus (B_1 \cup B_2 \cup B_3).$$

Тогда вершина u_t и вершины из множества B_1 образуют новый кластер $[t] = \{u_t\} \cup B_1$.

Следовательно, общее число кластеров уменьшается $K = (K \setminus B_1) \cup \{[t]\}$, как и размерность матрицы B , которая переопределяется по следующим правилам:

$$b([t], [i]) := 1, [i] \in B_2 \cup \{[t]\}, \quad b([i], [j]) := 1, [i] \in B_3, [j] \in B_2 \cup \{[t]\},$$

$$b([i], [j]) := 0, [i] \in B_2, [j] \in B_3 \cup \{[t]\} \cup B_4, \quad b([i], [j]) := 0, [j] \in B_3, [i] \in B_4 \cup \{[t]\},$$

$$b([t], [i]) = b([i], [t]) := 0, [i] \in B_4.$$

Все остальные значения матрицы B остаются без изменения.

Шаг 4. Алгоритм продолжается до полного перебора всех вершин графа, которые в зависимости от наличия или отсутствия связи с вершинами из имеющихся кластеров обрабатываются в соответствии с шагом 2 или шагом 3.

Выходные данные алгоритма:

- множество кластеров $[k] \in K$;
- множество фактор-вершин $[U]$, множество фактор-рёбер $[W]$;
- матрица смежности $B = \|b([i], [j])\|_{[u_i], [v_j] \in [U]}$, построенная на основе фактор-рёбер, которые связывают фактор-вершины.

В содержательном смысле связи, отражённые в матрице смежности фактор-графа, образующие множество фактор-рёбер, считаются доминантными, т. к. отражают особенности строения и иерархического распределения вершин исходного графа при разбиении на компоненты сильной связности.

Необходимо отметить, что по вычислительной сложности алгоритм сравним с известными алгоритмами определения компонентов сильной связности Косарайю и Тарьяна [6; 8], но при этом обладает более высокой скоростью реализации, т. к. основная идея алгоритма, основанная на уточнении информации о сетевой структуре на каждом шаге, не требует повторного обращения к уже обработанным вершинам, что присутствует у известных аналогов.

Последующее уточнение структурных особенностей соединения проводится с помощью построения иерархической структуры каждого кластера или дендрограммы. Определяющим правилом при построении является длина минимального цикла, содержащего пару вершин кластера.

В общем виде на множестве рёбер W графа Γ зададим целочисленную длину $d_w > 0$ каждому ребру $w \in W$. На множестве вершин U введём в рассмотрение матрицу мер близости $M = \|\mu_{ij}\|_{i,j=1}^n$, где $\mu_{ij} = l_{ij} + l_{ji}$ – положительное целое число, характеризующее близость вершин u_i, u_j , составленное из l_{ij} и l_{ji} длин кратчайших путей из вершины u_i в u_j и обратно. Обозначим $\beta = \max_{1 \leq i \neq j \leq n} \mu_{ji}$ и сопоставим каждому целому значению $\lambda, 0 \leq \lambda \leq \beta$ симметричную матрицу смежности $M^{(\lambda)} = \|\mu_{ij}^{(\lambda)}\|_{i,j=1}^n$ графа $\Gamma^{(\lambda)}$, в которой

$$\mu_{ij}^{(\lambda)} = \begin{cases} 1, & \mu_{ij} \leq \lambda \\ 0, & \text{иначе} \end{cases}.$$

Входные данные алгоритма: граф Γ , состоящий из множества кластеров $[k] \in K$, каждый из которых представляет собой сильно связный граф $\Gamma^{(k)}$.

Основное тело алгоритма:

Шаг 1. Используя известный алгоритм Флойда [17], определяем l_{ij} и l_{ji} для всех пар вершин в отдельно взятом кластере $[k_1]$ и строим матрицу $M = \|\mu_{ij}\|_{i,j=1}^m$, где m – число вершин в кластере $[k_1]$.

Шаг 2. Применяя λ_1 раз известный алгоритм в ширину или глубину [6], разбиваем неориентированный граф, образованный матрицей смежности $M^{(\lambda_1)} = \|\mu_{ij}^{(\lambda_1)}\|_{i,j=1}^n, 0 \leq \lambda_1 \leq \beta_1$ на компоненты связности $[[k_1^{(\lambda_1)}]] \in [k_1]$.

Шаг 3. Применяем шаг 1 и шаг 2 по очереди ко всем кластерам $[k] \in K$.

Выходные данные алгоритма: по каждому кластеру $[k_s]$ формируется дендрограмма высотой λ_s , которая начинается с одной вершины-кластера $[[k_s^{(\beta_s)}]]$ уровня β_s , содержит все вершины кластера $[k_s]$ и заканчивается набором вершин-кластеров $[[k_s^{(0)}]_1], [[k_s^{(0)}]_2], \dots, [[k_s^{(0)}]_m]$, уровня 0, каждая из которых содержит по одной вершине из кластера $[k_s]$.

В содержательном смысле связи, отражающие деление кластеров по мере движения к нижнему уровню, считаются доминантными, они отражают особенности строения и иерархическое распределение вершин в каждом компоненте связности по мере изменения параметра близости.

Полученные алгоритмы были успешно апробированы в решении задачи о представлении биоинформации на белковой сети растения Арабидопсис, которая является модельной структурой по изучению жизнеспособности клеток, устойчивости к химическим веществам и стрессам. В ходе исследования всей структуры, которая содержит порядка 35 000 белков, экспертами в области биоинженерии была выделена подсеть вторичного метаболизма сети Арабидопсиса размерностью 2824 вершины и 7570 рёбер. В качестве экспертов выступила научно-исследовательская группа лаборатории биоинженерии, возглавляемая член-корреспондентом РАН доктором биологических наук Виктором Павловичем Булгаковым [13–16].

В результате применения алгоритма динамического разбиения сетевой структуры на кластеры по принципу циклической эквивалентности размерность структуры была уменьшена в 1,78 раза. Обработка данного соединения при равных условиях компьютерной реализации составила порядка 2 часов, что в разы меньше по сравнению с методом расчёта матрицы расстояний, реализация которого потребовала почти 7 дней, и намного информативнее по сравнению с методом Тарьяна [6; 8], реализация которого заняла часы, но не отражала доминантных связей. Последующим этапом по наиболее значимым кластерам (см. рис. 1) были построены соответствующие дендрограммы (см. рис. 2).

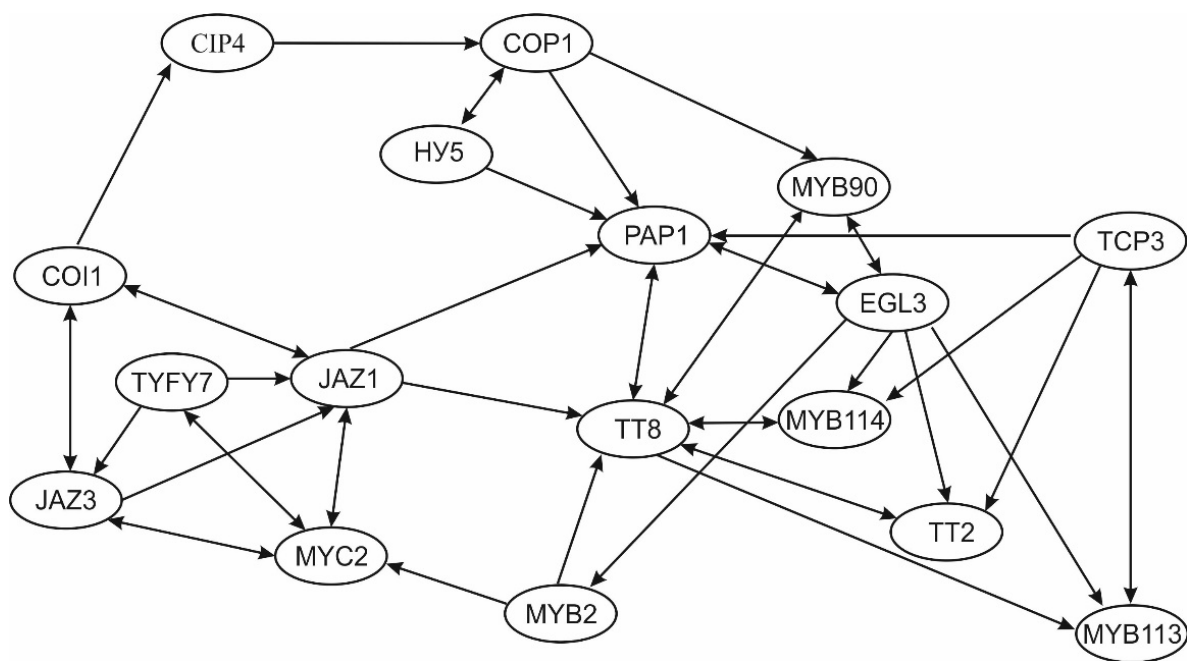


Рис. 1. Пример сильно связанного графа, образующего кластер

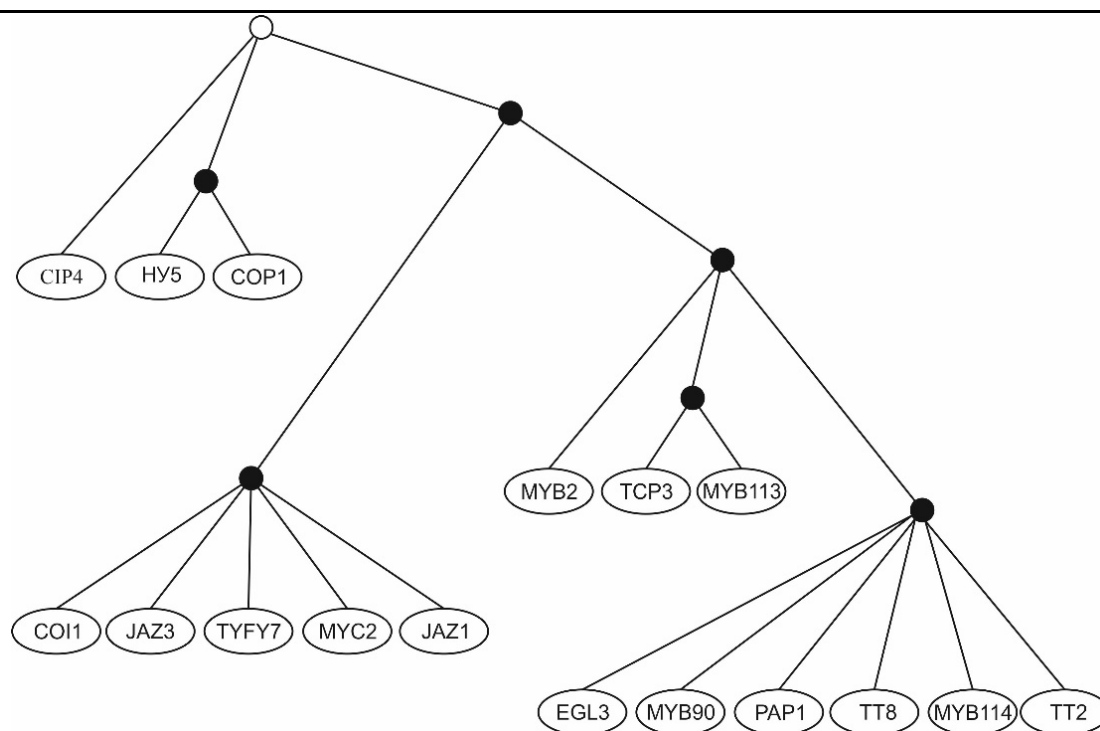


Рис. 2. Пример дендрограммы

Заключение

Применение контекстного подхода является обоснованным и целесообразным в ситуации первичной обработки сетевых структур большой размерности, строение которых сложно охватить общими представлениями. Наравне с этим принцип, основанный на выделении доминантных связей, позволяет наметить дальнейшие шаги исследования. А разработанные алгоритмы, в отличие от известных аналогов выделения компонентов сильной связности, не являются закрытыми, т. е. не требуют повторного перезапуска при любых изменениях в структуре, и отражают доминантные связи, характеризующие иерархическое представление образованных кластеров.

ЛИТЕРАТУРА

1. Анохин, П. К. Избранные труды: Кибернетика функциональных систем / П. К. Анохин; под ред. К. В. Судакова. – М.: Медицина, 1998. – 400 с.
2. Баяк, Д. А. Практическое применение методов кластеризации, классификации и аппроксимации / Д. А. Баяк, Д. В. Березин, В. А. Иванюк. – М.: Прометей, 2020. – 448 с.
3. Белявский, Г. И. Обучение без учителя и робастная оптимизация в задаче о портфеле / Г. И. Белявский, Н. В. Данилова, А. Д. Логунов // Известия вузов. Северо-Кавказский регион. Серия: Естественные науки. – 2020. – № 4 (208). – С. 4-9.
4. Брюхомицкий, Ю. А. Параметрические методы распознавания образов динамической биометрии / Ю. А. Брюхомицкий // Известия ЮФУ. Технические науки. – 2011. – № 12. – С. 170-180.
5. Казаков, М. А. Алгоритм кластеризации на основе разбиения пространства признаков / М. А. Казаков // Вестник КРАУНЦ. Физ.-мат. науки. – 2022. – Т. 39. – № 2. – С. 136-149.
6. Алгоритмы. Построение и анализ / Т. Кормен, Ч. Лейзерсон, Р. Ривест, К. Штайн. – М.: Лаборатория базовых знаний, 2004. – 955 с.
7. Корнеев, В. В. Параллельные вычислительные системы / В. В. Корнеев. – М.: Нолидж, 1999. – 311 с.
8. Кристофидес, Н. Теория графов. Алгоритмический подход / Н. Кристофидес. – М.: Мир, 1978. – 432 с.
9. Олескин, А. В. Сетевые структуры в биосистемах и человеческом социуме / А. В. Олескин. – М.: Издательство URSS, 2012. – 280 с.
10. Петрова, А. Н. Разработка модели для ранжирования объектов в системах больших данных / А. Н. Петрова, Д. О. Фролов // Учёные записки Комсомольского-на-Амуре государственного технического университета. Науки о природе и технике. – 2024. – № III (75). – С. 43-47.

11. Петрова, А. Н. Разработка нейронных сетей для сквозного обучения моделей поиска информации / А. Н. Петрова, Д. О. Фролов // Учёные записки Комсомольского-на-Амуре государственного технического университета. Науки о природе и технике. – 2024. – № III (75). – С. 48-52.
12. Ухтомский, А. А. Доминанта / А. А. Ухтомский. – СПб.: Питер, 2002. – 448 с.
13. Bulgakov, V. P. Critical analysis of protein signaling networks involved in the regulation of plant secondary metabolism: focus on anthocyanins / V. P. Bulgakov, T. V. Avramenko, G. Sh. Tsitsiashvili // Crit. Rev. Biotechnol. – 2017. – Vol. 37. – P. 685-700.
14. Bulgakov, V. P. Linking Brassinosteroid and ABA Signaling in the Context of Stress Acclimation / V. P. Bulgakov, T. V. Avramenko // International Journal of Molecular Sciences. – 2020. – Vol. 21. – № 14. – Article ID 5108.
15. Bulgakov, V. P. Basic Protein Modules Combining Absciscic Acid and Light Signaling in Arabidopsis / V. P. Bulgakov, O. G. Koren // Frontiers in Plant Science. – 2022. – Vol. 12. Article ID 808960.
16. Bulgakov, V. P. Chromatin modifications and memory in regulation of stress-related polyphenols: finding new ways to control flavonoid biosynthesis / V. P. Bulgakov // Critical Reviews in Biotechnology. – 2024. – Article ID 2336529.
17. Floid, R. W. An adaptive algorithm for spatial greyscale / R. W. Floid, L. Steinberg // SID 75 Digest. – 1975. – P. 36-37.
18. Tsitsiashvili G. Sh. Hierarchical Classification of Directed Graph Nodes and Application of Protein networks / V. P. Bulgakov, A. S. Losev, G. Sh. Tsitsiashvili // Biostatistics and Biometrics Open Access Journal. – 2017. – Vol. 1 (4).