OSTIS-2016
(Open Semantic Technologies for Intelligent Systems)

УДК 004.04: 004.822

АНАЛИЗ ПРОЦЕССОВ ИНФОРМАЦИОННОГО ОБМЕНА В НАУКОМЕТРИЧЕСКИХ БАЗАХ ДАННЫХ

Потебня А.В., Погорелый С.Д.

Киевский национальный университет им. Тараса Шевченко, г. Киев, Украина
admin@artem.bz.ua
sdp@univ.net.ua

В статье проведено исследование процессов распространения информации в наукометрических базах данных. На основе разработанной математической модели с применением методов комбинаторной оптимизации определены тенденции кластеризации баз данных и установлены основные способы поддержания их целостности. Выполнено экспериментальное исследование системы *ArXiv*.
Ключевые слова: наукометрическая база данных; задача разбиения графов; функция модулярности.

Введение

Модели сложных систем (*complex system*), которые описываются набором отдельных компонентов и способом взаимодействия между ними, являются чрезвычайно распространенными в современных научных исследованиях. К ним относятся социальные сети, наукометрические базы данных, всемирная паутина WWW, коммуникационные структуры и т.п. При этом, формирование сложных систем связано с появлением ряда новых свойств, в частности, синергизма (*synergy*) и эмерджентности (*emergence*). Явление синергии связано со стремительным ростом эффективности всей системы по сравнению с деятельностью отдельных компонентов. Эмерджентность является проявлением целостности системы и определяет появление качественно новых свойств, которые не являются присущими ни одному из отдельных участников [Newman, 2010].

В настоящее время общепринятой является глобальная интеграция научных исследований, которая осуществляется средствами наукометрических баз данных и социальных сетей. Однако, при этом эффективность научного развития оказывается зависимой от состояния соответствующих сложных систем, которые в связи с активизацией дисинергетических процессов могут войти в фазу рецессии или подвергнуться полному разрушению. Как следствие, важной задачей системного анализа является исследование процессов информационного обмена в таких системах и диагностика их общего состояния [Newman, 2006].

Например, популярная в данный момент база *Scopus* индексирует более 53 млн. научных работ. Кроме того, размерность этой системы растет ежегодно более чем на 2 млн. статей. Подобные тенденции развития стабильно демонстрируют другие наукометрические базы (*Web of Science*, *Google Scholar*, *IEEE Explore*) и академические социальные сети, предназначенные для обмена документами и их обсуждения (*ReseachGate*, *Academia.edu*, *Mendeley*). Вместе с тем, стремительное внедрение баз данных подвергается постоянной критике в связи с закрытостью ведущих систем, платным доступом к публикациям, непрозрачностью механизмов расчета импакт-факторов и ограничением доступа системы *Google Scholar* некоторыми издателями. Следствием стало формирование каталога открытых журналов *DOAJ* и ряда баз регионального назначения (например, *РИНЦ*), что привело к существенному разделению основных научных результатов. Кроме того, после короткой фазы роста, структура каждой из таких систем становится крайне неустойчивой и нуждается в дополнительной внешней поддержке для сохранения жизнеспособности [Fortunato, 2010].

Следует отметить, что целостность наукометрической базы определяется особенностями ее структурной организации и процессами циркулирования информации. В связи с этим, *целью данной работы* является разработка метода диагностики топологии и определения влиятельности узлов, необходимого современным системам для предотвращения катастрофического распада центрального ядра основного фрактального компонента на отдельные «острова», неспособные к обмену информацией. Кроме того, такой анализ

необходим для установления способности системы к проведению сетевой мобилизации, при которой большинство участников, получая оперативную информацию и работая в наиболее перспективном направлении, реализуют эффект синергии.

1. Формирование математической модели наукометрической базы

Чрезвычайно важной для исследования организации сложных сетей является задача обнаружения в них скрытых структур и определения режима их функционирования. Известно, что в процессе эволюции сложные системы склонны к формированию сообществ (модулей) с высокой плотностью сильных внутренних связей при незначительном количестве слабых внешних соединений. С этим связаны основные свойства таких систем – феномены *«малых миров»*, *«клуба богатых»*, *«малого диаметра»*, *«тяжелых хвостов»* и др. Установлено, что для эффективной работы сети наиболее важны слабые межмодульные связи, по которым информация передается между отдельными сообществами. Наличие таких связей сокращает среднее расстояние между любыми научными статьями до 6 – 7 рёбер даже для крупнейших баз данных, а их разрушение неизбежно приводит к распаду системы на отдельные фрагменты. Таким образом, архитекторы наукометрических баз обязательно должны учитывать состояние слабых соединений при поддержке своих разработок.

Проблема выделения сообществ в сложных сетях является воплощением распространенной задачи разбиения графов (*graph partition problem*). Она требует поиска оптимального распределения всех статей по подмножествам, при котором заданная целевая функция (*objective function*) достигает экстремума. Важно, что задача разбиения графов относится к NP-полным задачам комбинаторной оптимизации (*combinatorial optimization problem*), и время ее решения экспоненциально зависит от размерности входных данных. Как следствие, при обработке графов реальных систем применяются эвристические алгоритмы, которые предусматривают более эффективное распределение ресурсов при поиске решений [Potebnia et al., 2015].

Применяя структурный подход, представим наукометрическую базу в виде ориентированного графа $G = (V, E)$, где V – набор публикаций, а E – множество цитирований между ними. Задача разбиения такого графа предусматривает распределение множества статей V на k непересекающихся подмножеств V_i, для которых $V = \bigcup_{i=1}^{k} V_i$ и $V_i \bigcap V_j = \varnothing$ при $i \neq j$. Обозначим через E_i множество внутренних ребер группы V_i, а через E_{ij} – набор внешних соединений между сообществами V_i и V_j. Следует отметить, что в отличие от социальных сетей или мировой паутины, графы наукометрических баз данных являются ациклическими в связи с невозможностью цитирования еще не написанных статей. На рисунке 1 приведен пример упрощенной структуры основного фрактального компонента такого графа.

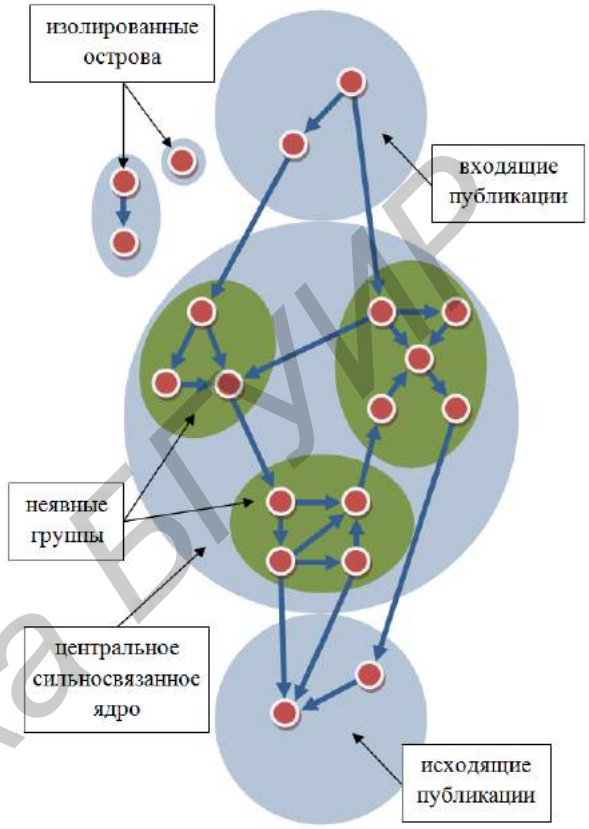

Рисунок 1 – Пример структуры основного фрактального компонента наукометрической базы

Для определения качества разбиения графа принято использовать функцию модулярности (*modularity*) Ньюмана-Гирвана Q, которая описывает соотношение плотности внутренних и межгрупповых связей для выбранного распределения узлов:

$$Q = \frac{1}{2m} \sum_{i,j} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \sigma(c_i, c_j), \qquad (1)$$

Здесь A – матрица смежности графа, k_i – степень вершины i, m – общее количество ребер $|E|$,

$$\sigma(c_i, c_j) = \begin{cases} 1, c_i = c_j \\ 0, c_i \neq c_j \end{cases}, \qquad (2)$$

где c_i – номер класса, к которому принадлежит вершина i. Экспериментально показано, что о реальном наличии сообществ в системе свидетельствуют значения модулярности, превышающие 0,3.

2. Методы оптимального разбиения наукометрических баз и определения влиятельности публикаций

Таким образом, задача оптимального разбиения графа может быть представлена в виде задачи максимизации функции модулярности Q. Однако, ее решение путем простого перебора всех вариантов является невозможным в связи с явлением *«комбинаторного взрыва»*. Поэтому, в работе [Newman, 2006] предложен эффективный жадный алгоритм, основанный на последовательном объединении сообществ V_i и V_j, которые обеспечивают наибольший прирост функции модулярности:

$$\Delta Q(V_i, V_j) = |E_{ij}| - \frac{d(V_i)\,d(V_j)}{2|E|}, \qquad (3)$$

где $|E_{ij}|$ – количество ребер между этими множествами, а $d(V_i)$ – степень множества V_i, которая является суммой степеней вершин $v \in V_i$.

При этом наиболее выгодным является объединение подмножеств с высокой плотностью межгрупповых связей при незначительном количестве внутренних соединений. Важно, что объединение изолированных наборов исключается, поскольку при $|E_{ij}| = 0$ величина прироста $\Delta Q(V_i, V_j)$ не может принимать положительные значения.

В этой работе для исследования реальных наукометрических баз данных применен ускоренный итерационный метод, который состоит из двух стадий [Lancichinetti et al., 2011]. Первый этап предусматривает формирование низкоразмерных сообществ путем оптимизации функции модулярности на локальном уровне. При этом определяется возможность объединения каждого множества V_i с его соседями V_j и рассчитываются соответствующие значения ΔQ. Вычисления на первой стадии заканчиваются получением локального максимума функции модулярности, а второй этап требует проведения агрегации подмножеств, образования кластеров большей мощности и определения веса межкластерных дуг. Итерации алгоритма продолжаются до получения устойчивых множеств, состав которых в дальнейшем остается неизменным.

Для определения относительной влиятельности научных публикаций предлагается метрика центральности узлов графа, которая может быть определена различными способами:

1. Центральность по степени (*degree centrality*) вычисляется как количество связей, инцидентных заданной вершине $C_D(v) = d(v)$. Однако, некоторые статьи при высоком уровне цитирования могут быть связаны с другими кластерами в графе малым количеством ребер. Поэтому применение данной метрики не является исчерпывающим при определении влиятельности публикаций.

2. Центральность по близости (*closeness centrality*) показывает скорость распространения информации между узлами:

$$C_C(v) = \frac{|V| - 1}{\sum_{t \in V \setminus v} d_G(v, t)}, \qquad (4)$$

где $d_G(v,t)$ – кратчайший путь от статьи v к вершине t. Она определяет близость отдельной публикации ко всем другим документам в базе. При этом учитывается не только наличие соседних узлов, но и состояние их связей с другими документами.

3. Центральность по посредничеству (*betweenness centrality*) вычисляется как количество кратчайших путей между всеми парами публикаций, которые проходят через заданный узел, то есть:

$$C_B(v) = \sum_{s \neq t \in V \setminus v} \frac{\sigma_{st}(v)}{\sigma_{st}}, \qquad (5)$$

где σ_{st} – общее количество кратчайших маршрутов между вершинами s и t, а $\sigma_{st}(v)$ – их количество при условии прохождения через точку v. Однако, существенным недостатком этой метрики является ее вычислительная сложность.

4. Центральность по собственному вектору (*eigenvector centrality*) демонстрирует зависимость влиятельности узла $C_E(v)$ от значений влиятельности его соседей $C_E(t)$ при $t \in N(v)$. Вычисление этой метрики связано с решением системы уравнений вида:

$$C_E(v) = \frac{1}{\lambda} \sum_{t \in N(v)} C_E(t), \qquad (5)$$

где λ – некоторые константы.

Кроме того, авторитетность публикаций может быть рассчитана с помощью известных методов ранжирования *PageRank* и HITS (*Hyperlink Induced Topic Search*), которые широко применяются поисковыми системами при обработке веб-страниц. Алгоритм *PageRank* устанавливает значимость статьи путем подсчета влиятельности всех ссылок на нее, а метрика HITS предусматривает расчет для каждого узла оценок авторитетности и посредничества. Важно, что применение алгоритма HITS является чрезвычайно эффективным при анализе графов наукометрических баз, поскольку он позволяет выявить как цитируемые исследовательские работы, так и качественные аналитические обзоры (посредники), которые ссылаются на наиболее авторитетные узлы. Таким

образом, кроме индексов цитирования для определения влиятельности документов в наукометрических базах должны быть применены другие метрики, учитывающие представление статьи в различных неявных сообществах [Holme et al., 2002; Iyer et al., 2011].

3. Исследование процессов распространения информации в наукометрической базе ArXiv

ArXiv является одной из самых известных бесплатных наукометрических баз и содержит более 700 000 публикаций по вопросам физики, математики, компьютерных наука и биологии. В статье для исследования выбран фрагмент этой системы (секция физики высоких энергий), охватывающий 34 546 публикаций и 421 578 ссылок между ними. Связи со статьями, которые не входят в базу *ArXiv*, были изъяты из сформированного графа [Leskovec et al., 2014].

Установлено, что граф наукометрической базы содержит 61 компоненту слабой связанности (*weakly connected component*), которые имеют связь между всеми вершинами по крайней мере в одном направлении. Крупнейший из выделенных подграфов охватывает 99,58% всех статей, а все остальные компоненты являются островами, объединяющими лишь несколько публикаций. Вместе с тем, в графе выделено 21 608 компонент сильной связанности (*strongly connected component*), в пределах которых существует взаимосвязь между любыми узлами. При этом одна из них является гигантской и содержит 36,79% узлов, а размерность остальных островов не превышает десяти публикаций. То есть, большинство статей пытаются цитировать наиболее авторитетные публикации, однако обратную связь имеют лишь треть из них.

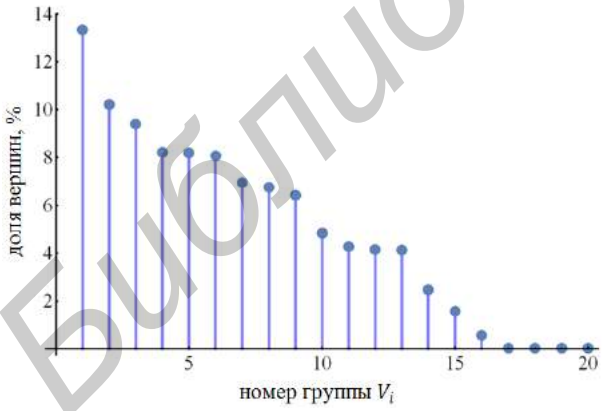

Рисунок 2 – Размерность самых крупных кластеров в наукометрической базе *ArXiv*

Путем оптимального разбиения графа было получено 80 неявных сообществ при значении модулярности $Q = 0{,}727$, что свидетельствует о существенной кластеризации. На рисунке 2 показана размерность 20 крупнейших групп. При этом лишь два из сложившихся множеств охватывают более 10% документов. На рисунке 3 приведена зависимость модулярности от значения порога цитирования публикаций, по которому выполнялась фильтрация узлов в графе. Видно, что статьи с низкими индексами цитирования, как правило, связаны только с узлами своего сообщества и не способствуют увеличению целостности системы.

Однако, изъятие из графа публикаций со средним уровнем цитирования (50 – 120 раз) приводит к стремительному увеличению модулярности, разрушению «мостов» между кластерами и отделению новых островов от гигантской компоненты. При этом наиболее цитируемые работы редко связаны между собой и при высоких уровнях порога отсечки располагаются на отдельных изолированных островах. Таким образом, основную связывающую роль в наукометрической базе выполняют статьи со средним уровнем цитирования.

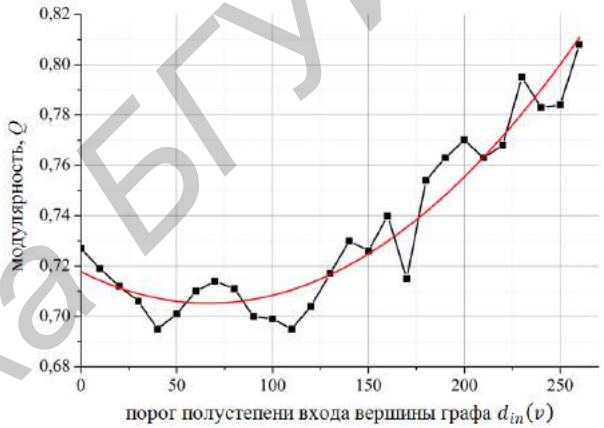

Рисунок 3 – Зависимость целостности базы *ArXiv* от значения порога цитирования публикаций

Рассмотрим самый большой компонент связанности графа системы *ArXiv* при значении порога отсечки, которое составляет 150 цитирований (рисунок 4). Видно, что высокие величины модулярности и значительная автономия сообществ приводят к недостаточности применения полустепеней входа для описания влиятельности публикаций (рисунок 4а). При этом мостовые узлы, изъятие которых приводит к катастрофическому распаду, имеют наименьшие показатели цитирования, а наибольшие значения метрики получают узлы, «погруженные» в глубину своих сообществ. Однако, при использовании в качестве метрики центральности по посредничеству (рисунок 4б) или по близости (рисунок 4в) влиятельность таких документов оказывается более высокой.

Применение алгоритма *PageRank* (рисунок 4г) приводит к получению наибольшего ранга статьями, связи которых, как правило, сосредоточены в пределах одного сообщества. Показательным является использование алгоритма HITS для нахождения лучших посредников, которые также содержатся внутри кластеров рядом с авторитетными исследовательскими работами

(рисунок 4д). Общая структура системы *ArXiv* для статей, цитируемых более 100 раз, показана на рисунке 5.

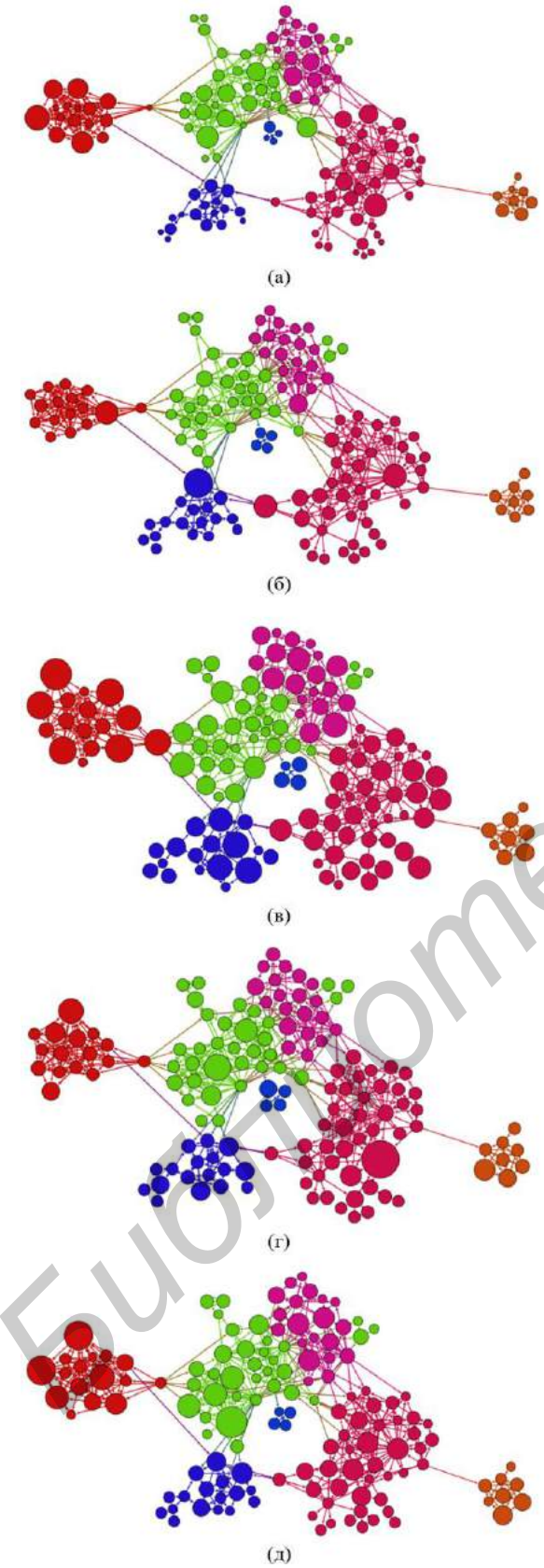

Рисунок 4 – Определение с помощью разных метрик влиятельности публикаций, цитированных в системе *ArXiv* более 150 раз. Цветами выделены разные сообщества в графе. Размер узлов указан пропорционально их авторитетности

В результате проведенных измерений было установлено, что значение модулярности составляет 0,937 для всемирной паутины; 0,763 для блогосферы *LiveJournal*; 0,881 для сети доверия пользователей системы шифрования PGP [Ovelgonne et al., 2010]. Кроме этого, расчитанные значения Q для 10 случайных эго-графов социальных сетей *Facebook*, *Google+* и *Twitter* соответственно равны 0,494, 0,299 и 0,271. Таким образом, по целостности структуры база *ArXiv* уступает эго-графам, но существенно превосходит всемирную паутину. При этом система *ArXiv* и платформа *LiveJournal* имеют подобные тенденции к формированию неявных сообществ и проведению сетевой мобилизации.

Заключение

Таким образом, на примере системы *ArXiv* установлено, что структура ведущих наукометрических баз является распределенной на отдельные неявные сообщества, связи между которыми поддерживаются за счет публикаций со средним уровнем цитирования. Именно они, как правило, одновременно содержат ссылки на разные авторитетные работы и образуют между ними мостовые соединения, которые выполняют важную роль в процессе обмена информацией. Однако, использование индексов цитирования, распространенное в современных системах, не позволяет выявить такие критические узлы. Для этого более целесообразным является применение метрик центральности по посредничеству и по близости.

Библиографический список

[Newman, 2010] Newman, M. Networks: An Introduction / M. Newman // Oxford University Press, 2010. – 720 p.

[Newman, 2006] Newman, M. Modularity and community structure in networks / M. Newman // Proceedings of the National Academy of Sciences. – 2006. – № 103(23). – P. 8577 – 8582.

[Fortunato, 2010] Fortunato, S. Community detection in graph / S. Fortunato // Physics Reports. Elsevier. – 2010. – № 486. – P. 75 – 174. DOI: doi:10.1016/j.physrep.2009.11.002.

[Potebnia et al., 2015] Potebnia, A. Innovative GPU accelerated algorithm for fast minimum convex hulls computation / A. Potebnia, S. Pogorilyy // Proceedings of the 2015 Federated Conference on Computer Science and Information Systems. – 2015. – P. 555 – 561. DOI: 10.15439/2015F305.

[Lancichinetti et al., 2011] Lancichinetti, A. Finding Statistically Significant Communities in Networks / A. Lancichinetti, F. Radicchi, J. Ramasco, S. Fortunato // PLoS ONE. – 2011. – № 6(4): e18961. DOI: 10.1371/journal.pone.0018961.

[Holme et al., 2002] Holme, P. Attack vulnerability of complex networks / P. Holme, B. Kim, C. Yoon, S. Han // Physical Review E 65: 056109. – 2002. DOI: 10.1103/PhysRevE.65.056109.

[Iyer et al., 2011] Iyer, S. Attack Robustness and Centrality of Complex Networks / S. Iyer, T. Killigback, B. Sundaram, Z. Wang // PLoS ONE. – 2013. – № 8(4): e59613. DOI: 10.1371/journal.pone.0059613.

[Leskovec et al., 2014] Leskovec, J. SNAP Datasets: Stanford Large Network Dataset Collection. – June 2014. URL: http://snap.stanford.edu/data.

[Ovelgonne et al., 2010] Ovelgonne, M. Randomized greedy modularity optimization for group detection in huge social networks / M. Ovelgonne, A. Geyer-Schulz, M. Stein // Proceedings of the 4th Workshop on Social Network Mining and Analysis. ACM, New York. – 2010.

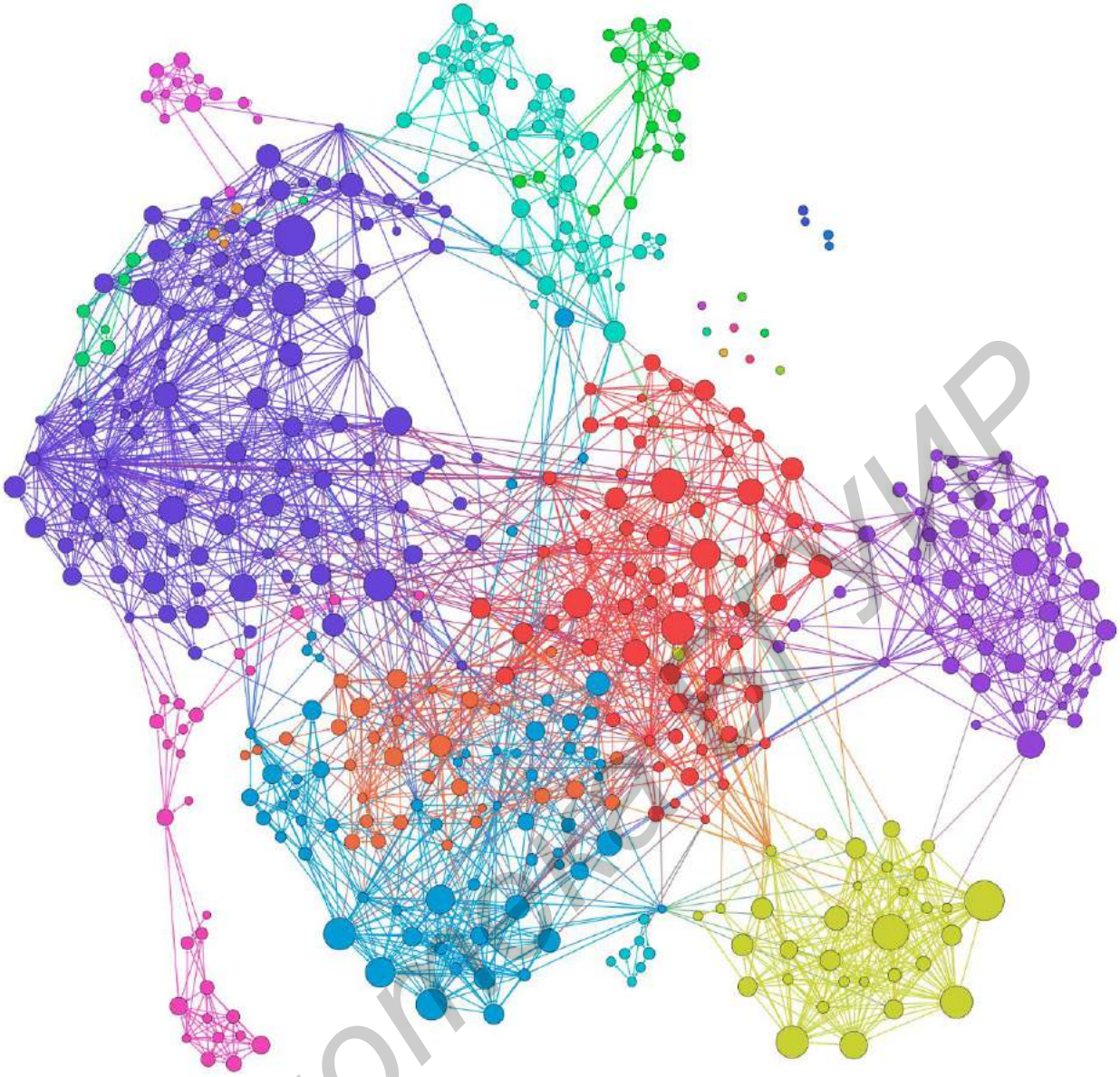

Рисунок 5 – Структура графа наукометрической базы *ArXiv* для публикаций, которые имеют более 100 ссылок. Размер узлов указан пропорционально индексу их цитирования

ANALYSIS OF INFORMATION EXCHANGE PROCESSES IN SCIENTOMETRIC DATABASES

Potebnia A.V., Pogorilyy S.D.

Kyiv National Taras Shevchenko University, Kyiv, Ukraine

admin@artem.bz.ua

sdp@univ.net.ua

This paper presents the investigation of information propagation processes in scientometric databases. On the basis of the developed mathematical model with the application of the combinatorial optimization methods, we have identified the trends of databases clustering. The paper contains the experimental research of the *ArXiv* system. As a result, we have obtained the following important results and conclusions:

- The database structure is divided into a number of implicit document communities, which have a relatively low density of the intergroup connections;
- The primary linking function between these communities is performed by the articles of the average citing level;
- The usage of citation indexes is not enough to identify the critical bridge nodes. For this purpose the calculation of the closeness and betweenness centrality metrics is much more appropriate;
- *ArXiv* system is more vulnerable than the ego-graphs of social networks, but its structural integrity is larger in comparison to the *World Wide Web*;
- In addition, the scientometric base *ArXiv* and blogosphere *LiveJournal* have the similar trends of the implicit communities formation and the network mobilization.

Keywords: scientometric database; graph partition problem; modularity function.