УДК 004.048

Шульдова Светлана Георгиевна
кандидат технических наук,
доцент кафедры программного обеспечения информационных технологий
УО «Белорусский государственный университет информатики и радиоэлектроники,
Республика Беларусь, г. Минск,
shsg@bsuir.by
Shuldava Sviatlana G.
The Belarusian State University of Informatics and Radioelectronics
Belarus, Minsk

Парамонов Антон Иванович
кандидат технических наук,
зав. кафедрой информационных систем и технологий института информационных технологий
УО «Белорусский государственный университет информатики и радиоэлектроники
Республика Беларусь, г. Минск,
a.paramonov@bsuir.by
Paramonov Anton I.
The Belarusian State University of Informatics and Radioelectronics
Belarus, Minsk

Лапицкая Наталья Владимировна
кандидат технических наук,
зав. кафедрой программного обеспечения информационных технологий
УО «Белорусский государственный университет информатики и радиоэлектроники
Республика Беларусь, г. Минск,
lapan@bsuir.by
Lapitskaya Natalia V.
The Belarusian State University of Informatics and Radioelectronics
Belarus, Minsk

ПОДХОД К ОЦЕНКЕ НАУЧНОГО ВЗАИМОДЕЙСТВИЯ

THE APPROACH TO ASSESSING SCIENTIFIC INTERACTION

Аннотация

В статье рассматриваются вопросы оценки научного взаимодействия сотрудников одного подразделения с использованием сетей соавторства и цитирования. Рассмотрены модели представления сетей в виде графов, а также методы кластеризации графов (выявления сообществ), использующих модулярность. Приведены результаты

компьютерного эксперимента по исследованию научных связей профессорско-преподавательского кафедры на основе данных из Google Академии.

Abstract

The article discusses the issues of assessing the scientific interactions between employees of one department using co-authorship and citation networks. Models of representing networks in the form of graphs, as well as methods of graph clustering (identifying communities) using modularity are considered. The results of a computer experiment to study the scientific relations of scientists based on data from the Google Academy are presented.

Ключевые слова: сеть сотрудничества, цитирование, граф соавторства, кластеризация, научные сообщества.

Keywords: collaboration network, citation, co-authorship graph, clustering, scientific communities.

Значительный рост числа научных публикаций обуславливает актуальность задач анализа публикационных показателей, характеризующих активность и значимость автора, научного коллектива и организации, а также анализа научного взаимодействия ученых. Исследование научных связей ученых, научных коллективов и учреждений позволяет оценивать тенденции развития научных направлений, идентифицировать конкретных людей и научные школы [1]. Основой для такого исследования являются сети научного сотрудничества, построенные на основе данных, получаемых из международных научных баз данных, таких как Web of Science (webofknowledge.com), Scopus (www.scopus.com), Google Академия (scholar.google.com), российская библиографическая база данных научного цитирования РИНЦ (elibrary.ru).

Примерами сетей сотрудничества являются сети соавторства и цитирования. Согласно [2], два ученых считаются связанными, если они совместно написали статью. Образование сети цитирования базируется на обязательности ссылок на используемые источники информации в научных публикациях. Библиографические ссылки обеспечивают необходимые информационные связи между публика-

циями, учеными, журналами и организациями. Кроме того, они обеспечивают связь во времени – между предыдущими публикациями и далее, позволяя отслеживать в динамике развитие научной мысли и интереса к ней.

Анализ сетей сотрудничества основывается на их представлении в виде графов и исследовании теоретико-графовыми методами: степени вершин, компоненты связности, кластеризация и выделение сообществ.

В настоящей работе рассматривается подход к оценке научного взаимодействия сотрудников структурного подразделения на основе анализа сетей сотрудничества, результаты применения которого позволят повысить эффективность командной работы, расширят возможности построения связей в междисциплинарных проектах, усилят позиции молодых специалистов, приобретающих опыт академического письма.

Для исследования использовались данные о научных публикациях, полученные из открытых профилей Google Scholar профессорско-преподавательского состава кафедры программного обеспечения информационных технологий учреждения образования «Белорусский государственный университет информатики и радиоэлектроники».

Формальное описание задачи исследования. Публикация $article_i$ представляет собой кортеж:

$$article_i = \langle Title_i, Keys_i, Authors_i, Type, SCI_i, Ref_i, Text_i, Publish, Year_i \rangle, \quad (1)$$

где $Title_i$ – название публикации, $Keys_i$ – множество ключевых слов, $Authors_i$ – множество авторов, SCI_i – цитируемость, Ref_i – аннотация, $Text_i$ – полный текст публикации или доступная часть, $Publish$ – издание, $Year_i$ – год издания и $Type$ – тип (разновидность публикации).

Автор $author_k$ – кортеж из таких характеристик, как фамилия, имя, индекс Хирша, организация:

$$author_k = \langle Surname_k, Name_k, h - index_k, Organization \rangle. \quad (2)$$

Хотя бы один из авторов $Authors_i$ статьи $article_i$ является сотрудником кафедры в год ее публикации, то есть

$$Authors_i = \{author | \exists\, autho \in Employes\}, Employes \subset Authors$$

$$employe_l = \langle Surname_l, Name_l, Degree, Position, Year\_start_l, Year\_End_l \rangle \quad (3)$$

$$Year_i \in [Year\_Start_l,\ Year\_End_l.],$$

В случае, если сотрудник в настоящее время работает на кафедре в указанной должности, значение $Year\_End_l$ равно *null*.

Два автора являются соавторами, если существует статья, множеству авторов которой принадлежат оба автора

$$author_k\ \Re^{au}\ author_l \equiv author_k \text{ соавтор } author_l,$$

$$\exists\, article_i \in Articles,\ author_k \in Authors_i,\ author_l \in Authors_i \quad (4)$$

Граф соавторства представляет собой неориентированный граф $G^{au} = (Au,\ \Re^{au})_k$, без петель, вершины au_k которого соответствуют авторам, дуги – отношениям соавторства $\Re^{au} \subset Au \times Au$, $e_{au} = \langle au_k, au_l \rangle \in \Re^{au}$. Сила связи не исследуется, поэтому вес ребра равен 1.

Граф цитирования автора представляет собой ориентированный граф, ребра которого соединяют цитируемого и цитирующего авторов:

$$G^{ca} = (Au, \Re^{ca})_k$$

Граф цитирования автора строится на основе данных о цитировании статьи без самоцитирования, то есть

$$article_j\ \Re^{ar}\ article_i \equiv article_j \text{ цитирует } article_i, i \neq j,$$

$$author_l\ \Re^{ca}\ author_k \equiv author_l \text{ цитирует } author_k, k \neq l, \quad (5)$$

$$author_k \in Authors_i, author_l \in Authors_j, Authors_i \cap Authors_j = \emptyset.$$

В графе сотрудничества авторов на основе цитирования учитываются публикации, цитирующие статью $article_i$, для которых существует автор, принадлежащий множествам авторов цитируемой и цитирующей статей

$$article_j\ \Re^{ar}\ article_i \equiv article_j \text{ цитирует } article_i, i \neq j,$$

$$\exists\, authors_q \in Authors_j \cap Authors_i. \quad (6)$$

С учетом предложенных обозначений, научное сотрудничество можно определить, как сеть ученых, плотно связанную отношениями соавторства и сотрудничества на основе цитирования, а также общей тематикой статей:

$$SCh = \langle CoAuthors,\ Rate\rangle,\ CoAuthors \subset Authors, Rate \subset Keys. \quad (7)$$

Выявление сообществ в социальных графах. Задача выявления сообществ на графах (или кластеризация графов) определяется как выделение таких непересекающихся подмножеств, в каждом из которых вершины связаны между собой более чем с вершинами вне данного подмножества. Многие методы разбиения графа на подмножества в качестве критерия оптимальности используют модулярность [3]. Для графа $G = (V, E)$, заданного матрицей смежности A, данная мера рассчитается по формуле:

$$Q = \frac{1}{4|E|}\sum (A_{ij} - \frac{d_i d_j}{2|E|})\,\delta_{c_i c_j}, \quad (7)$$

где A_{ij} – элемент матрицы смежности, отражающий наличие соединения между вершинами i и j, $|E|$ – количество ребер графа, $d_{i(j)}$ – степень $i(j)$ – той вершины, $\delta_{c_i c_j}$ – делта-функция, которая обозначает принадлежность вершин i и j к одному сообществу, C_i – номер сообщества (группы, кластера), к которому принадлежит вершина i.

Таким образом, модулярность равна разности между долей рёбер внутри сообщества при данном разбиении и долей рёбер, если бы они были случайно сгенерированы. Поэтому она показывает выраженность сообществ (случайный граф структуры сообществ не имеет). На практике значение ноль или около нуля означает, что разбиение графа случайно. Считается хорошей оценкой разбиение графа, если оно достигает значения хотя бы около 0,7 [4].

К основным методам, которые используют модулярность, относятся Edge Betweenness, FastGreedy, WalkTrap.

Метод WalkTrap основан на кластеризации вершин: между двумя вершинами из разных сообществ расстояние велико, а из

одного – мало. Объект может переместиться из вершины i в вершину j с вероятностью $P_{ij} = A_{ij} / d_i$. То есть на каждом шаге равновероятно выбирается "сосед" вершины i. Таким образом определяется матрица переходов P случайного блуждания. Она примечательна тем, что её степени являются вероятностями перехода из одной вершины в другую за соответствующее число шагов: вероятность перехода из i в j за t шагов равна $(P^t)_{ij}$. Также следует отметить, что $P = D^{-1}A$, где D — матрица со степенями вершин на диагонали. На основе вышеизложенного, вводится метрика на вершинах:

$$r_{i,j} = \sqrt{\sum_{k=1}^{n} \frac{\left(P_{ik}^t - P_{jk}^t\right)^2}{d(k)}} = \left\| D^{-\frac{1}{2}} P_{i\bullet}^t - D^{-\frac{1}{2}} P_{j\bullet}^t \right\|, \quad (8)$$

где $P_i{}^{t\cdot}$ – вектор из вероятностей перехода за t шагов из вершины i во все другие. Авторы [4] советуют брать $3 \leq t \leq 8$. Естественным образом расстояние между вершинами обобщается на расстояние между сообществами:

$$r_{C_1 C_2} = \left\| D^{-\frac{1}{2}} P_{C_1\bullet}^t - D^{-\frac{1}{2}} P_{C_2\bullet}^t \right\| = \sqrt{\sum_{k=1}^{n} \frac{\left(P_{C_1 k}^t - P_{C_2 k}^t\right)^2}{d(k)}}, \quad (9)$$

где $P_{C_j}^t = \frac{1}{|C|} \sum_{i \in C} P_{ij}^t$.

На основе метрики (8) выделяются кластеры в графе. Начальное разбиение – по одной вершине в каждом кластере $\mathcal{P}_1 = \{\{v\}, v \in V\}$. Также для всех пар инцидентных вершине считается расстояние.

Далее для каждого k выполняются такие действия:

1. Выбрать C_1 и C_2 из $\mathcal{P}_k$ согласно некоторому метрическому критерию.

2. Объединить два сообщества в новое $C_3 = C_1 \cup C_2$ и обновить разбиение $\mathcal{P}_{k+1} = (\mathcal{P}_k \setminus \{C_1, C_2\}) \cup C_3$.

3. Обновить расстояния между инцидентными сообществами. После $n-1$ шага получается дендрограмма разбиений, а $\mathcal{P}_n = \{V\}$.

Таким образом, остался неясным только критерий выбора пар сообществ на шаге 1. Выбирается пара сообществ, минимизирующих приращение среднего квадратов расстояний между каждой вершиной и их сообществом при объединении этих сообществ:

$$\Delta\sigma(C_1,C_2) = \frac{1}{n}\left(\sum_{i\in C_3} r_{iC_3}^2 - \sum_{i\in C_1} r_{iC_1}^2 - \sum_{i\in C_2} r_{iC_2}^2\right) \to \min_{c_1 c_2} \qquad (10)$$

Таким образом, выбирается разбиение, на котором модулярность достигает максимума.

Результаты компьютерного эксперимента. Для проведения эксперимента были сформированы два набора данных: первый, содержащий публикации 14 сотрудников одной кафедры: профессора, 10 доцентов и 3 сотрудников без степени, второй набор – публикации, цитирующие публикации первого набора согласно формуле (6). На первом этапе средствами языка программирования R был выполнен препроцессинг данных, который включал приведение фамилий авторов к виду «ио_фамилия», транслитерацию фамилий авторов, поиск нечеткого соответствия строк и их преобразование для исключения различного написания фамилии и инициалов одного и того же человека. Далее для каждого набора был построен граф и выполнена кластеризация методом WalkTrap. Последовательность построения графа следующая: формирование лексического корпуса, построение терм-документной матрицы и ее преобразование в матрицу смежности графа. Кластеризация выполнена с различным числом шагов t, выбор полученного в результате разбиения основывался на максимальном значении модулярности.

Первый набор данных содержал 759 записей о публикациях сотрудников кафедры. Пример одного из полученных графов сообществ показан на рисунке 1. Число вершин графа – 268, ребер – 569. Получено 10 кластеров при числе шагов t = 8, модулярность составила 0,84.

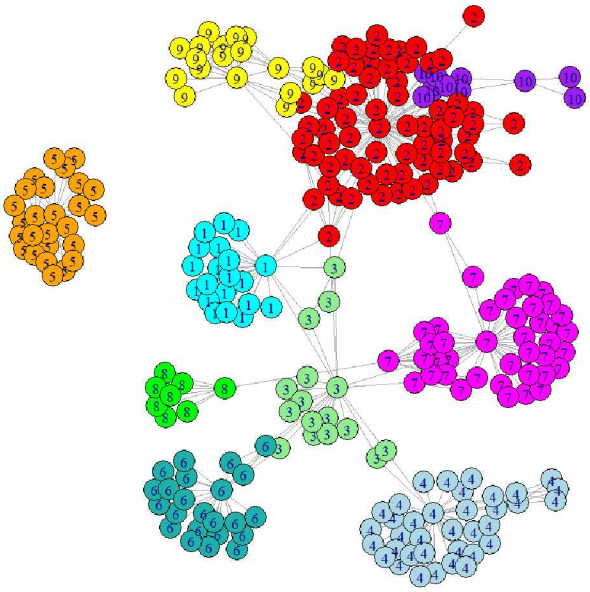

Рис. 1. Граф соавторства с сообществами

Второй набор данных включал 1038 записей о статьях, цитирующих статьи первого набора, один или более авторов которых принадлежат множеству авторов первого набора. Граф сотрудничества представлен на рисунке 2, он содержит 354 вершины и 727 ребер. В результате кластеризации при числе шагов $t = 8$ идентифицировано 12 сообществ, модулярность составила 0,85.

Таким образом, в результате компьютерного эксперимента по исследованию научных связей сотрудников одной кафедры на основе их публикаций в открытых банках данных (на примере Google Академии) получены сообщества ученых, связанных отношениями соавторства и сотрудничества на основе цитирования. Используемый объем исходных данных позволил оценить соответствие состава авторов выявленных сообществ реальным научным связям сотрудников кафедры, что в свою очередь позволяет рекомендовать предложенный подход для решения задачи оценки научного взаимодействия ученых в коллективе любых масштабов. В дальнейшем

исследовании представляется интересным анализ локации и аффилиации авторов, ранжирование публикации по типу и изданию.

Полученные результаты могут быть полезны магистрантам и аспирантам при определении актуальных направлений исследований и ученых, чьи научные интересы им соответствуют, формировании списка публикаций для анализа современного состояния изучаемой проблемы.

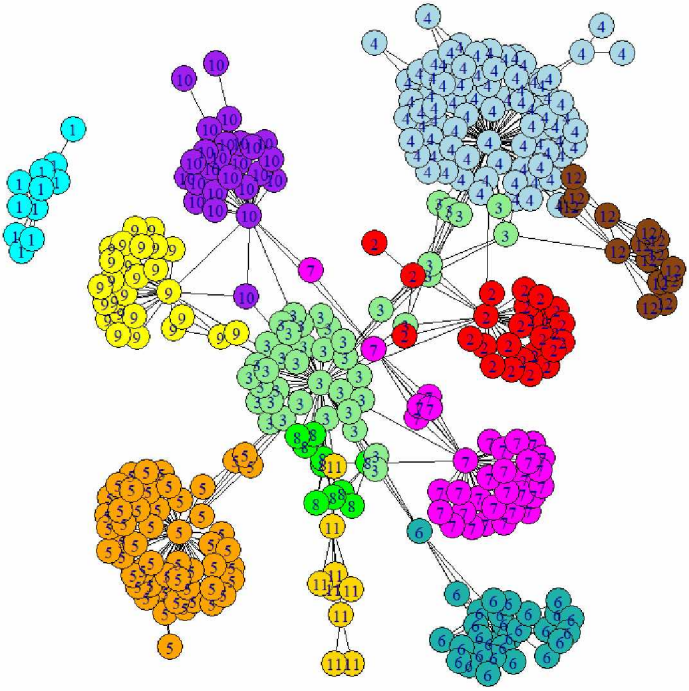

Рис. 2. Граф сотрудничества на основе цитирования с сообществами

СПИСОК ЛИТЕРАТУРЫ

1. Тронин В. Г. Оценка результатов научно-исследовательской работы и наукометрия : учеб. пособие / В. Г. Тронин, А. Р. Сафиуллин. Ульяновск : УлГТУ, 2019. 136 с.

2. Newman, M. Finding and evaluating community structure in networks / M. Newman, M. Girvan // Physical Review E. 2004. Vol. 69(2). P 026113.

3. A Review of Clustering Techniques and Development / Saxena, Amit and Prasad, Mukesh and Gupta, Akshansh and Bharill, Neha and Patel, op and Tiwari, Aruna and Er, Meng and Lin, Chin-Teng // Neurocomputing. 2017. № 267.
4. Pons, P. Computing communities in large networks using random walks / Pascal Pons, Matthieu Latapy // Physical Review E. 2005. Режим доступа: http://arxiv. org/abs/physics/0512106v1 (дата доступа: 20.06.2024).