DOI: 10.51790/2712-9942-2020-1-4-2

ИНСТРУМЕНТ АНАЛИЗА НАУЧНЫХ СООБЩЕСТВ НА ОСНОВЕ МЕТОДА МОДЕЛИРОВАНИЯ ТЕМ И ТЕОРИИ ГРАФОВ

И. Н. Девицын [1], **И. В. Савин** [2]

[1] *Сургутский государственный университет, г. Сургут, Российская Федерация*
ORCID: https://orcid.org/0000-0003-0173-6186, ivnide@gmail.com
[2] *Уральский федеральный университет, г. Екатеринбург, Российская Федерация;*
Институт экологических наук и технологий, Автономный университет Барселоны,
Барселона, Испания, ORCID: https://orcid.org/0000-0002-9469-0510, ivanvsavin@hotmail.com

Аннотация: в статье рассматривается новый инструмент анализа научных сообществ с использованием методов моделирования тем и теории графов. Результаты применения предложенного нами подхода представлены для публикаций авторов, аффилированных с Сургутским государственным университетом в Scopus за период 1995–2021 гг. Разработанный инструмент позволяет определять основные направления научных исследований, выявлять передовые коллективы научных работников по отдельным направлениям, а также анализировать взаимосвязи научных коллективов. Представлены результаты распределения публикаций по времени, девяти основным темам, расчет метрик графов соавторства, построенных на основе исследуемого набора данных. В будущем разработанный подход можно применить для оценки научно-исследовательского потенциала научных организаций, для оперативного определения направлений научных исследований, выявления передовых коллективов и научных работников по перспективным направлениям.

Ключевые слова: научное сообщество, машинное обучение, естественная обработка языка, сетевой анализ, граф соавторства.

Благодарности: И. В. Савин благодарит за финансовую поддержку Российский научный фонд (номер гранта 19-18-00262).

Для цитирования: Девицын И. Н., Савин И. В. Инструмент анализа научных сообществ на основе метода моделирования тем и теории графов. *Успехи кибернетики*. 2020;1(4):13–21. DOI: 10.51790/2712-9942-2020-1-4-2.

RESEARCH COMMUNITY ANALYTIC TOOL BASED ON TOPIC MODELING AND NETWORK ANALYSIS

Ivan N. Devitsyn [1], **Ivan V. Savin** [2]

[1] *Surgut State University, Surgut, Russian Federation*
ORCID: https://orcid.org/0000-0003-0173-6186, ivnide@gmail.com
[2] *Ural Federal University, Yekaterinburg, Russian Federation;*
Institute of Environmental Sciences and Technologies, Autonomous University of Barcelona,
Barcelona, Spain, ORCID: https://orcid.org/0000-0002-9469-0510, ivanvsavin@hotmail.com

Abstract: the study presents a new research community analytical tool based on topic modeling and methods from graph theory. The results of the proposed approach are presented for Scopus-indexed publications by the authors affiliated with Surgut State University in 1995–2021. The tool makes it possible to determine the key research areas, identify the leading research teams in certain areas and analyze the relationships between these teams. The paper includes the distribution of publications over time, nine main areas of publications, and a range of metrics for the co-authorship graphs of the studied dataset. In the future, the tool can be applied to assess the potential of research organizations, select the research areas, and identify the leading research teams and researchers in promising areas.

Keywords: research community, machine learning, natural language processing, network analysis, co-authorship graph.

Acknowledgements: I. Savin expresses his thankfulness to the Russian Science Foundation for its financial support (RSF grant 18-00262).

Cite this article: Devitsyn I. N., Savin I. V. Research Community Analytic Tool Based on Topic Modeling and Network Analysis. *Russian Journal of Cybernetics*. 2020;1(4):13–21. DOI: 10.51790/2712-9942-2020-1-4-2.

Введение

В крупных городах, где имеются высшие учебные заведения, научно-исследовательские институты и наукоемкие предприятия, научными исследованиями и разработками занимаются множество научных коллективов, часто работая над схожими проблемами параллельно, но независимо. Данные о структуре компетенций разрозненны, а процессы, связанные с их актуализацией, требуют большого количества трудозатрат и, как правило, растягиваются во времени. Это особенно актуально сегодня, когда количество публикуемых исследований растет с каждым годом, что приводит к отсутствию единой карты научно-исследовательских коллективов, результатов их работы и взаимосвязей между ними. У руководителей научных организаций и у самих научно-исследовательских коллективов зачастую нет возможности получить оперативную информацию о наличии нужных компетенций под решение новой технологической задачи у других коллективов. Данная работа направлена на получение инструмента оценки научно-исследовательского потенциала авторских коллективов в научных организациях для оперативного определения направлений их научных исследований, выявления передовых коллективов и научных работников по перспективным направлениям, включая междисциплинарные исследования, а также анализа взаимосвязей научных коллективов и сотрудников.

Для решения данной задачи, быстрого и объективного анализа как публикуемых результатов, так и взаимосвязей между авторами, мы применяем метод моделирования тем, разработанный недавно на стыке дисциплин обработки естественного языка и машинного обучения. Он позволяет объективно оценить количество тем, присутствующих в текстах в совокупности и в каждом из текстов по отдельности. Поскольку в дополнение к текстам публикаций мы знаем их авторов, мы также используем теорию графов, чтобы построить сеть (граф) ученых с точки зрения частоты их соавторства и применить к ней популярные метрики из теории графов.

Современные исследования по направлению

С ростом возможностей современных компьютеров и повышением доступности больших объемов данных выбор инструментов для исследования текстовой информации сильно вырос. Если раньше тексты вычитывались и классифицировались вручную, то сегодня в нашем распоряжении так называемый метод моделирования тем (ММТ). ММТ — это метод кластеризации текстовых данных с целью определения в них осмысленных тем, анализа трендов этих тем, реклассификации и аннотирования документов [1, 2]. ММТ несет в себе идею, что тексты содержат скрытые темы в определенной пропорции, а каждая тема — это вероятностное распределение по существующему в текстах списку слов. Тогда как в читаемых нами текстах темы и их распределение скрыты от нас, ММТ открывает темы, скрытые в текстах, и их распределение, которое наилучшим образом объясняет каждый конкретный текст. Преимущества ММТ над простым анализом ключевых слов в том, что слова могут иметь разное значение в зависимости от контекста. Также он полностью определяется на основе данных: не нужно наперед задавать темы.

ММТ широко применяется для текстов разного типа. Например, его активно используют для анализа патентных данных. Он применялся для реклассификации патентов на продуктовые и технологические субклассы, чтобы затем исследовать технологическое развитие и географию инноваций технологий фотовольтаики в США [3]; для определения возникающих тем среди патентов США, Японии и ЕС [4]; для определения ведущих (первых) патентов новых тем [5] и для предсказания трендов развития патентов [6, 7]. Кроме патентов, ММТ широко использовался для других типов текстовых данных. Многие исследования изучали научную литературу, опубликованную в разных реферируемых журналах [2, 8, 9], или все экономические статьи из базы данных вроде JSTOR [10]. Более того, некоторые исследования специально фокусировались на литературе, посвященной таким темам, как информационная безопасность [11] или биоинформатика [12]. Тогда как одни ученые смотрят только на аннотации статей [2], другие изучают полные тексты статей [10].

Другое популярное направление применения ММТ — это новостные статьи. Это могут быть как финансовые новости (например, взятые из Dow Jones Newswires Archive [13] или финансовых

аналитических докладов [14]), публикации в Интернете по проблеме изменения климата [15] и публикации из социальных сетей вроде Твиттера [16]. Наконец, недавно ММТ был применен к открытым вопросам из опросов общественного мнения [17–20]. Все это иллюстрирует универсальность данного метода в применении к текстовым данным очень разного типа с точки зрения как их объема, так и содержания.

Что касается существующих инструментов для сетевого анализа публикаций, наиболее близким к разрабатываемому является VOSviewer — бесплатный программный инструмент для построения и визуализации библиометрических сетей, разработанный сотрудниками Центра исследований науки и технологий (CWTS) Лейденского университета [21]. Он позволяет визуализировать сети по признаку соавторства, отдельным публикациям, ключевым словам и производить кластеризацию полученных данных. Данные для построения графов могут быть загружены из файлов или через API поддерживаемых наукометрических баз, таких как WoS, Scopus, Dimensions, CrossRef, Medline. Например, авторы [22] применили его для визуализации кластеров патентов, в которых встречались аморфные сплавы, чтобы проанализировать тенденции патентного ландшафта в этой области металлургии. В [23] с помощью VOSviewer исследовалось взаимодействие авторов американского сообщества масс-спектрометрии – исследование показало значительные отличия в публикационной активности университетов и лабораторий, а также, что довольно ожидаемо, — более частое соавторство между сотрудниками, работающими в одном учреждении. Авторы [24, 25] — разработчики VOSviewer — производили кластеризацию публикаций на основе числа цитирований с помощью предложенного ими подхода, который по умолчанию и используется в VOSviewer. В [26] предлагается инструмент для автоматизированного сбора данных для наукометрического анализа из eLibrary на основе парсинга страниц публикаций, т.е. без использования платного API, с последующей визуализацией с помощью VOSviewer. Примечательно, что кроме работ [24, 25], выполненных разработчиками VOSviewer, в остальных рассмотренных публикациях авторы проводят предварительную обработку массива данных с помощью других инструментов, т.к. в VOSviewer отсутствуют сложные инструменты отбора или обработки данных и метаданных публикаций, а значит, он самостоятельно не сможет различить слова, написанные с большой или маленькой буквы, или слова в единственном или множественном числе. Более того, VOSviewer ориентируется на ключевые слова, а не на метод моделирования тем, а значит, не способен (как нами указывалось ранее) различать значения слов в зависимости от контекста. Все это говорит о большом спектре возможностей для улучшения существующих инструментов, и наше исследование как раз стремится сделать вклад в этом направлении.

Моделирование тем

Для целей разработки и тестирования был получен набор данных, включающий 798 статей, опубликованных авторами из Сургутского государственного университета на английском языке за период с 1995 по 2021 гг. Данные были получены с помощью API Scopus с использованием библиотеки Pybliometrics. Для целей данного пилотного исследования было принято решение ограничиться небольшой выборкой, в будущем можно добавить в набор данных и труды авторов из других научных организаций.

Прежде чем применять ММТ, необходимо провести предварительную обработку полученных данных. Названия и аннотации статей были объединены для удобства обработки и увеличения размера данных. Заглавные буквы были заменены на строчные, удалены цифры и знаки препинания. Затем была применена лемматизация слов, т.е. приведение слов к их словарной форме, с использованием библиотеки pymystem3 от Яндекс (например, слова «captures» и «capturing» приведены к «capture»). После этого текст разбивался на токены, к которым применялся метод определения и формирования n-грамм (биграмм, триграмм и т.д.). Идея заключается в том, что часто используемые вместе слова образуют устойчивый оборот (например, «vortex structure», «physical education»). И, наконец, полученная база была очищена от стоп-слов и слов, состоящих менее чем из трех букв, с помощью библиотеки NLTK.

Чтобы определить оптимальное количество тем в текстах статей, использовались индикаторы согласованности (coherence score) и ошибки прогноза (perplexity score) сформированной модели ММТ. Тогда как первый индикатор указывает на то, насколько часто слова, отнесенные к каждой из тем, встречаются друг с другом и таким образом формируют более согласованную тему, второй индикатор

показывает, насколько хорошо модель ММТ, построенная на обучающей выборке данных, предсказывает распределение слов на проверочной выборке.

Непосредственно для ММТ был применен метод Latent Dirichlet Allocation (LDA) [27] как наиболее популярная модель ММТ из библиотеки GenSim на количестве от 2 до 50 тем. Как видно из рис. 1, ошибка прогноза модели практически непрерывно снижается с количеством тем, тогда как индикатор согласованности достигает максимума для 12 тем. Тем не менее, поскольку 9 тем обладают меньшей ошибкой прогноза и их легче интерпретировать, в то время как индикатор согласованности лишь чуть ниже, чем для 12 тем, нами было выбрано 9 тем.

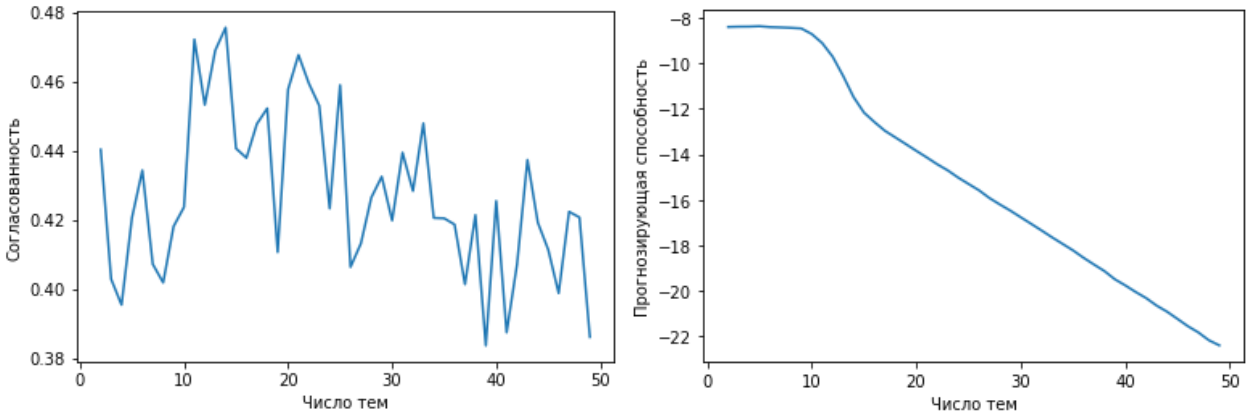

Рис. 1. *Графики индикаторов согласованности (слева) и прогнозирующей способности (справа)*

В целях понимания и интерпретации полученных тем для каждой из них были выгружены наиболее частые слова, т.е. слова, обладающие наиболее высокой вероятностью встретиться в каждой из тем, и иллюстративные примеры текстов — аннотации публикаций, в которых доля заданной темы максимальна. На основе данной информации мы назвали полученные темы следующим образом (см. рис. 2). Также из рис. 2 можно увидеть доли каждой из тем в общем объеме публикаций.

Номер темы	Наименование	Процент публикаций	Ключевые слова
1	Общественные науки	18.92	student, sport, education, university, educational, research, formation, activity, social, environment
2	Теория динамических систем	11.03	parameter, theory, motion, quasiattractor, vortex_structure, evolution, stochastic, chaos, measurement, state
3	Биология	7.39	species, compound, family, flavonoid, isolated, phase, layer, existence, spectra, oxidation
4	Математическое моделирование	22.81	gas, temperature, model, energy, procedure, oil, mathematical, equation, fuel, flow
5	Клиническая медицина	5.64	patient, treatment, syndrome, therapy, diaphragm, clinical, infant, acute, children, diagnosis
6	Теоретическая медицина	17.29	body, group, patient, age, functional, condition, living, region, cardiovascular, population
7	Физическая культура	8.27	physical, children, activity, athlete, training, health, mental, physical_culture, sport, competitive
8	Физика и химия	7.39	structure, copper, grain_boundary, optical, morphological, particles, surface, grain, composition, alloy
9	Лингвистика	1.25	scientists, russian, capital, slums, english, siberia, científicos, infrastructure, ciencia, endemic

Рис. 2. *Темы, их характеристика и доля в наборе данных*

Чтобы визуально оценить качество кластеризации терминов в каждой теме, воспользуемся пакетом pyLDAvis для построения карты распределения тем на плоскости. Хорошая тематическая модель будет давать темы, распределенные по всей плоскости и мало пересекающиеся между собой. Также мы воспользовались алгоритмом t-distributed Stochastic Neighbor Embedding (t-SNE) [28], т.к. он тоже позволяет уменьшить размерность данных, сохранив при этом малое расстояние между схожими точками, чтобы наглядно показать их распределение. Из рис. 3 видно, что темы достаточно хорошо отделены друг от друга, хотя редкие термины из одной из тем могут находиться в другой, что свойственно LDA-моделям.

Также было рассчитано число публикаций, преимущественно посвященных каждой из тем, и доля каждой из тем в наборе данных. Распределение представлено на рис. 4. Еще одна интересная характеристика, которую можно получить на основе имеющихся данных, — изменение доли каждой из полученных тем во времени. Из графика на рис. 4 видно, что они представлены очень неравномерно. В отдельные годы, например, 2016-2020, публикаций очень много, тогда как в другие (1995-2000) их

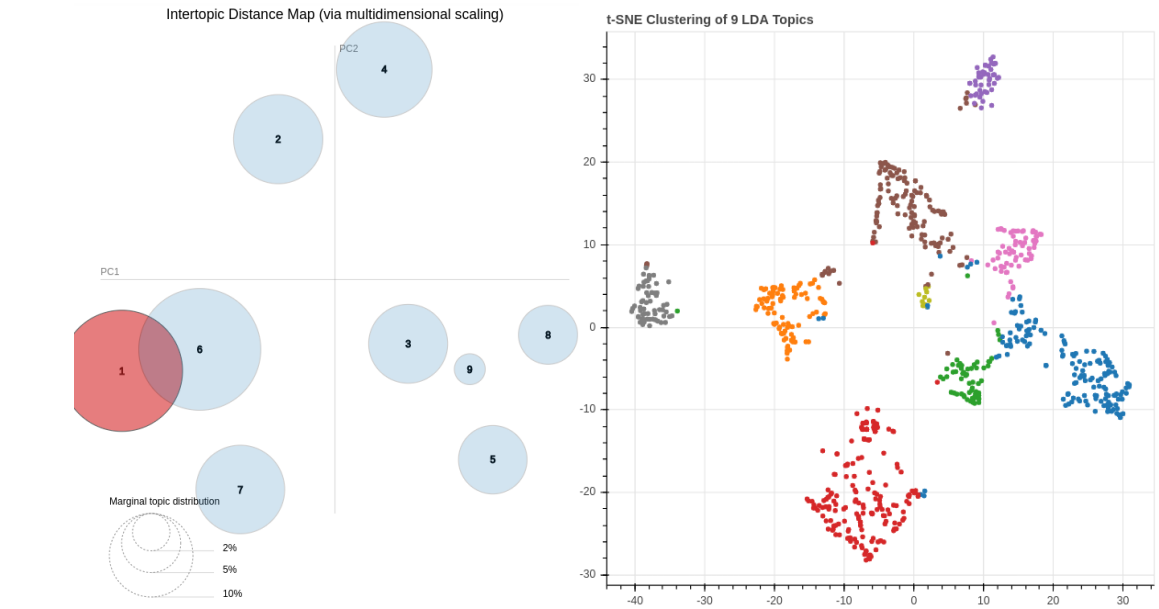

Рис. 3. *График распределения тем (слева) и их t-SNE визуализация (справа)*

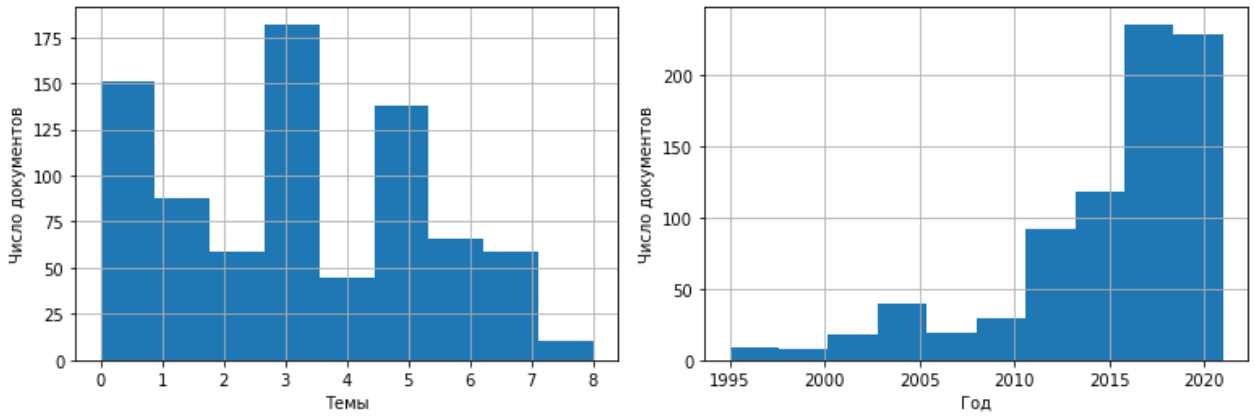

Рис. 4. *Распределение публикаций из набора данных по темам (слева) и по годам (справа)*

крайне мало. Поскольку данные столь несбалансированные, стоит воспринимать результаты динамики тем по времени с осторожностью, так как они сильно зависят от размера выборки.

На рис. 5 представлен график распределения относительной доли тем по годам. Видно, что представленность тем в научных публикациях по времени сильно колебалась. Тема физики и химии, например, сильно теряла в популярности в 2005 году, а также в 2017–2019 гг., а доля темы общественных наук выросла в последние десять лет.

Сетевой анализ

Для того чтобы иметь возможность построения графов научных коллективов, нужно располагать данными об авторах статей. Тот факт, что выборка состоит из статей, полученных через API Scopus, позволяет теперь по идентификатору публикации получить список ее авторов с их Scopus-идентификаторами. По ним были получены данные обо всех авторах, участвующих в наборе данных, включая: фамилию и инициалы, число публикаций, число цитирований, индекс Хирша, организацию, с которой автор аффилирован, включая страну и город ее расположения, и список идентификаторов соавторов (все известные соавторства, имеющиеся в Scopus). Теперь, объединив полученные из Scopus данные об авторах с результатами ММТ, а именно с номерами полученных тем, можно приступать к построению и анализу графов научного сообщества.

Под графом научного сообщества или коллектива будем понимать совокупность вершин, отражающих авторов публикаций, соединенных ребрами, показывающими отношение соавторства между

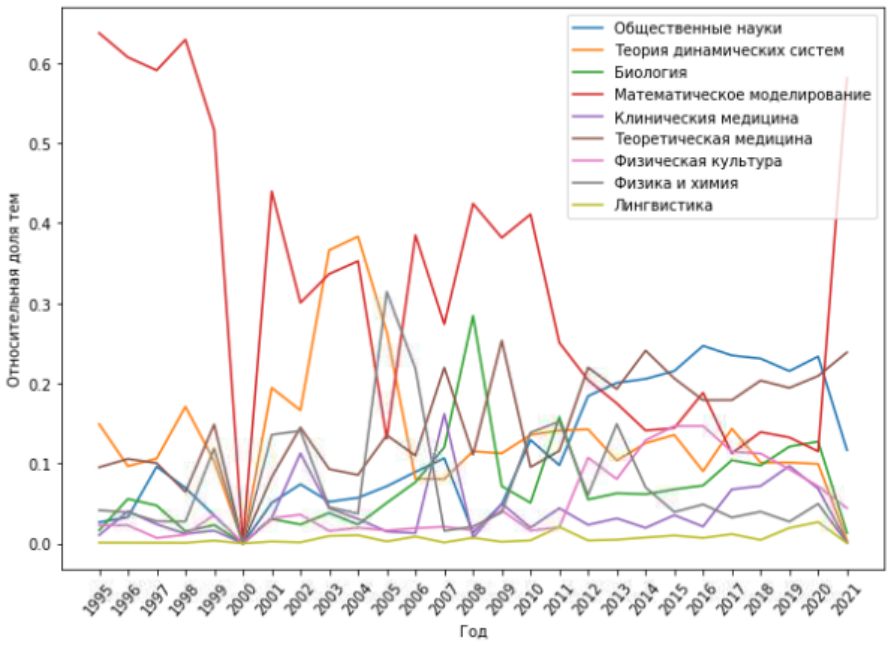

Рис. 5. *Динамика доли публикаций во времени*

ними. Таким образом, граф научного сообщества будет неориентированным (т.е. если автор N связан с M, то и M связан с N) и несвязным (т.е. не все элементы графа связаны между собой и между научными коллективами может не быть никаких общих соавторов). Поскольку каждый автор является соавтором самого себя, для простоты связями с самим собой в таком графе мы будем пренебрегать. По причине того, что граф содержит достаточно большое количество данных об авторах, для его построения было принято решение использовать совокупность библиотек python-igraph — для хранения и манипулирования графами как структурами данных и plotly — для визуализации графов с возможностью интерактивной подсветки узлов и инструментов просмотра, таких как масштабирование, панорамирование, выделение отдельных наборов вершин [29]. Результат такого построения представлен на рис. 6. Цветом узла на графе выделена страна, в которой работает соответствующий автор, а размер узла отражает индекс Хирша автора.

Так как строящиеся графы почти всегда будут содержать большое число узлов, для их распределения на плоскости был выбран силовой алгоритм визуализации Graphopt [30]. Используя данные о темах, полученные на предыдущем этапе, можно строить графы сообществ, работающих в конкретных интересующих нас направлениях, например, на рис. 7 представлен граф темы 2, посвященной динамическим системам. По графу видно, что в центре он содержит выраженный конгломерат соавторов (тесно связанную друг с другом группу соавторов), а также небольшие группы соавторов по краям, не имеющих общих публикаций с остальным коллективом.

Для анализа коллективов была реализована функция построения наибольшей связной компоненты графа, позволяющая быстро выделить из существующего графа подграф с основным массивом соавторов. Наибольшая связная компонента графа соавторов темы 2 представлена на рис. 7. Проанализировав данный граф, можно заметить любопытные особенности: формирование коллективов вокруг тесно взаимодействующего «ядра» с большим числом совместных публикаций, а также взаимосвязь центрального конгломерата соавторов с периферийными через 1-2 связующих соавторов.

Для поиска выдающихся авторов были применены методики анализа социальных сетей, а именно расчет центральностей графа. Пример расчета для наибольшей связной компоненты графа соавторов всего набора данных приведен на рис. 8.

Это позволяет выделить значимых авторов на основе расчета метрик центральности, а именно: степень связности (degree) — число связей, которую можно интерпретировать как «популярность» автора или число его уникальных соавторов; степень посредничества (betweenness) — число кратчайших путей, которые проходят через вершину, что можно интерпретировать как меру того, насколько часто автор выступает связующим звеном между разными коллективами; степень близости (closeness)

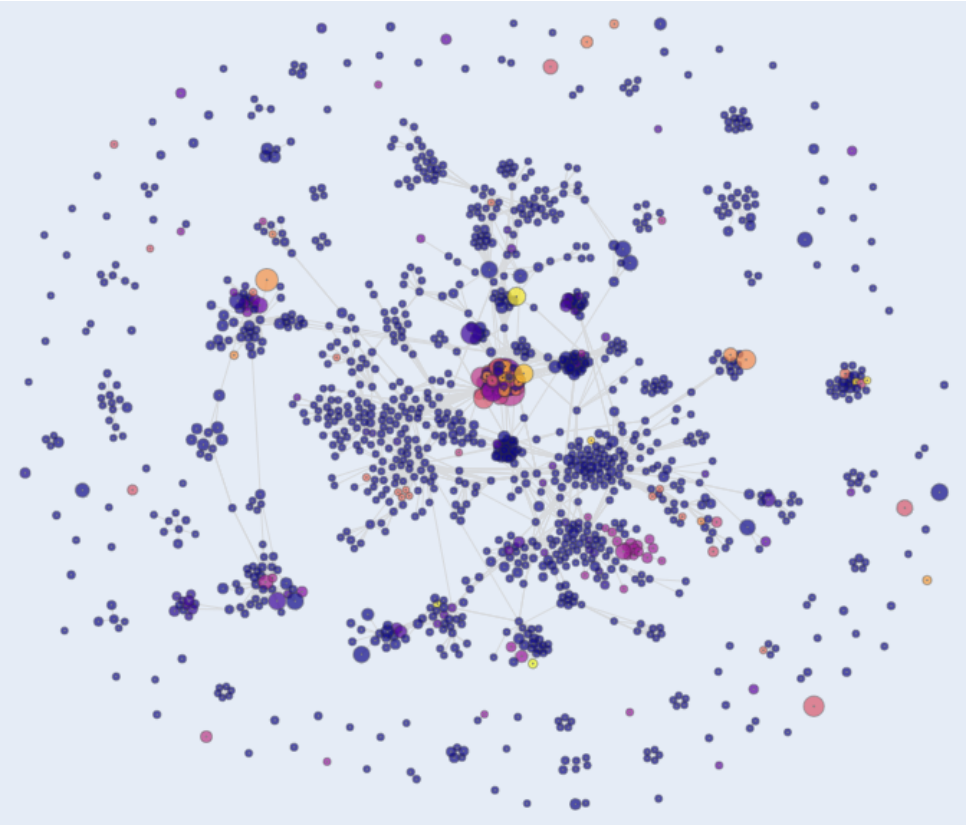

Рис. 6. *Граф авторов всего набора данных*

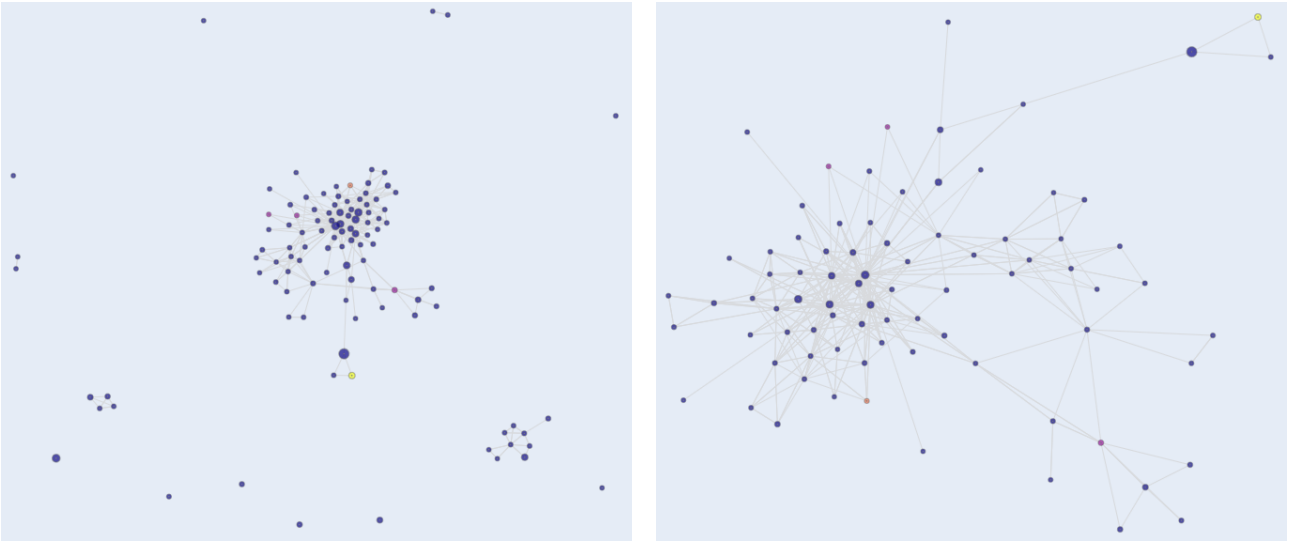

Рис. 7. *Граф авторов темы 2 «Теория динамических систем» (слева) и его наибольшая связная компонента (справа)*

— мера близости вершины ко всем остальным и влиятельности (eigenvector) — которая измеряет влиятельность вершины в зависимости от того, с какими другими вершинами она связана. Также имеется возможность рассчитать коэффициент кластеризации графа, например, для наибольшей связной компоненты всего набора данных он равен 0,81, что является формальным свидетельством наличия тесно связанных групп соавторов (см. рис. 6-7). Для данного графа также было установлено, что он, как и другие социальные структуры, удовлетворяет т.н. свойству «тесного мира», т.е. каждый узел связан с остальными через небольшое количество узлов, при том что этот же самый граф характеризуется высокой кластеризацией, когда соавторы N также пишут вместе статьи и без него, — своеобразный аналог «теории шести рукопожатий» для графов [31].

scopus id	author	degree	betweenness	closeness	eigenvector
57218615562	Kovalenko L. V.	0.039394	45772.445497	0.201342	3.180247e-08
55780169000	Bashkatova Yu V.	0.052525	149635.680709	0.197487	1.566641e-09
6603854205	Botirov É. Kh	0.080808	172920.325679	0.190715	2.747858e-06
6603639422	Eskov Valery M.	0.093939	99683.113075	0.187607	3.907006e-11
6603198324	Karpin V. A.	0.031313	19087.044156	0.183605	1.559942e-09
55399217600	Filatov Mikhail	0.037374	10986.569925	0.181885	3.597863e-11
56625937600	Filatova S. Yu	0.009091	0.000000	0.181485	1.551130e-09
16237928700	Drenin A. A.	0.040404	7719.204400	0.180492	1.783151e-07
57201259024	Eskov V. V.	0.066667	13719.487328	0.180295	3.728627e-11
36496932700	Kul'kov M. G.	0.014141	194849.000000	0.180065	1.361874e-04

Рис. 8. *Метрики центральности графа*

Заключение

Для решения поставленной задачи авторами были применены методы анализа и кластеризации текстовых данных (в частности, модели Latent Dirichlet Allocation как инструмента моделирования тем), получение дополнительной информации о совместных публикациях ученых из ресурсов сети Интернет (в частности, базы данных Scopus) и методы анализа из теории графов.

В результате был разработан инструмент оценки научно-исследовательского потенциала научных коллективов, который был успешно применен на основе данных о публикациях авторов из Сургутского государственного университета. Были определены направления научных исследований, выявлены передовые коллективы научных работников по отдельным направлениям, а также проанализированы взаимосвязи научных коллективов.

Среди прочего было показано, что научные публикации содержат девять основных тем; что динамика популярности этих тем существенно колебалась по времени; что граф соавторств научных исследований имеет выраженную структуру «центр-периферия» и высокий коэффициент кластеризации, указывающий на то, что соавторы соавторов часто являются также соавторами между собой; и, наконец, было показано, что структура графа соавторств научных публикаций имеет сходство со структурой «тесного мира», отличительной чертой которого является высокая кластеризация при коротком пути, соединяющем любые две вершины в графе.

В дальнейшем планируется добавление поддержки других источников данных, помимо API Scopus, адаптация инструментария к работе с другими коллективами авторов вне зависимости от исходных данных (в т.ч. возможность исследовать публикации не только на английском, но и на русском языке), а также расширение функциональности аналитической части инструмента. Рассматривается возможность оформления инструмента в виде самостоятельного программного обеспечения.

ЛИТЕРАТУРА

1. Blei D. M. Probabilistic Topic Models. *Communications of the ACM*. 2012;55(4):77–84.
2. De Battisti F., Ferrara A., Salini S. A Decade of Research in Statistics: a Topic Model Approach. *Scientometrics*. 2015;103(2):413–433.
3. Venugopalan S., Rai V. Topic Based Classification and Pattern Identification in Patents. *Technological Forecasting and Social Change*. 2015;94:236–250.
4. Lee W., Han E., Sohn S. Predicting the Pattern of Technology Convergence Using Big-Data Technology on Large-Scale Triadic Patents. *Technological Forecasting and Social Change*. 2015;100:317–329.
5. Kaplan S., Vakili K. The Double-Edged Sword of Recombination in Breakthrough Innovation. *Strategic Management Journal*. 2015;36:1435–1457.
6. Chen H., Zhang G., Zhu D., Lu J. Topic-Based Technological Forecasting Based on Patent Data: A Case Study of Australian Patents from 2000 to 2014. *Technological Forecasting and Social Change*. 2017;119(C):39–52.

7. Suominen A., Toivanen H., Seppänen M. Firms' Knowledge Profiles: Mapping Patent Data with Unsupervised Learning. *Technological Forecasting and Social Change*. 2017;115(9):131–142.
8. Lüdering J., Winker P. Forward or Backward Looking? The Economic Discourse and the Observed Reality. *Journal of Economics and Statistics*. 2016;236(4):483–515.
9. Griffith T., Steyvers M. Finding Scientific Topics. *Proceedings of the National Academy of Sciences*. 2004;101:5228–5235.
10. Ambrosino A., Cedrini M., Davis J., Fioria S., Guerzoni M., Nuccio M. What Topic Modeling Could Reveal about the Evolution of Economics. *Journal of Economic Methodology*. 2018;25(4):367–377.
11. Chang H.-C. The Synergy of Scientometric Analysis and Knowledge Mapping with Topic Models: Modelling the Development Trajectories of Information Security and Cyber-Security Research. *Journal of Information & Knowledge Management*. 2016;15(4):1650044.
12. Liu L., Tang L., Dong W., Yao S., Zhou W. An Overview of Topic Modeling and its Current Applications in Bioinformatics. *SpringerPlus*. 2016;5(1):1608.
13. Larsen V., Thorsrud L. *Business Cycle Narratives*. CESifo Working Paper Series 7468, CESifo. 2019.
14. Huang A., Lehavy R., Zang A. Y., Zheng R. Analyst Information Discovery and Interpretation Roles: A Topic Modeling Approach. *Management Science*. 2017;64(6):2833–2855.
15. Farrell J. Corporate Funding and Ideological Polarization about Climate Change. *Proceedings of the National Academy of Sciences*. 2016;113(1):92–97.
16. Chae B., Park E. Corporate Social Responsibility (CSR): A Survey of Topics and Trends Using Twitter Data and Topic Modeling. *Sustainability*. 2018;10(7):1–20.
17. Roberts M. E., Stewart B., Tingley D., Lucas C., Leder-Luis J., Gadarian S., Albertson B., Rand D. Structural Topic Models for Open-Ended Survey Responses. *American Journal of Political Science*. 2014;58(4):1064–1082.
18. Tvinnereim E., Fløttum K. Explaining Topic Prevalence in Answers to Open-Ended Survey Questions about Climate Change. *Nature Climate Change*. 2015;5:744-747.
19. Savin I., Drews S., Maestre-Andres S., van den Bergh J. Public Views on Carbon Taxation and Its Fairness: A Computational-Linguistics Analysis. *Climatic Change*. 2020;162:2107–2138.
20. Savin I., Drews S., van den Bergh J. Free Associations of Citizens and Scientists with Economic and Green Growth: A Computational Linguistics Analysis. *Ecological Economics*. 2021;180:106878.
21. *VOSviewer – Visualizing Scientific Landscapes*. Available at: https://www.vosviewer.com/.
22. De Oliveira B. S., Milanez D. H., Leiva D. R., de Faria L. I. L., Botta W. J., Kiminami C. S. Thermal Spraying Processes and Amorphous Alloys: Macro-Indicators of Patent Activity. *Materials Research*. 2018;20(Suppl. 1):89–95.
23. Palmblad M., van Eck N. J. Bibliometric Analyses Reveal Patterns of Collaboration between ASMS Members. *Journal of The American Society for Mass Spectrometry*. 2018;29(3):447–454.
24. Waltman L., van Eck N. J., Noyons Ed. A Unified Approach to Mapping and Clustering of Bibliometric Networks. *Journal of Informetrics*. 2010;4:629–635. 10.1016/j.joi.2010.07.002.
25. Van Eck N. J., Waltman L. Citation-Based Clustering of Publications Using CitNetExplorer and VOSviewer. *Scientometrics*. 2017;111:1053–1070. 10.1007/s11192-017-2300-7.
26. Низомутдинов Б. А., Тропников А. С. Автоматизированный сбор данных для наукометрического анализа. *Научный сервис в сети Интернет: труды XXI Всероссийской научной конференции (23-28 сентября 2019 г., г. Новороссийск)*. М.: ИПМ им. М. В. Келдыша; 2019;523–531.
27. Blei D. M., Ng A. Y., Jordan M. I. Latent Dirichlet Allocation. *Journal of Machine Learning Research*. 2003;3(4–5):993–1022.
28. Van der Maaten L. J. P., Hinton G. E. Visualizing Data Using t-SNE. *Journal of Machine Learning Research*. 2008;9:2579–2605.
29. Ognyanova K. *Network Visualization with R*. 2019. Available at: https://kateto.net/network-visualization.
30. Hu Y. Algorithms for Visualizing Large Networks. *Combinatorial Scientific Computing*. 2011;5. 10.1201/b11644-20.
31. Watts D. J., Strogatz S. Collective Dynamics of 'Small-World' Networks. *Nature*. 1998;393(6684):440–442.