# Социология:

## методология, методы, математическое моделирование

*Специальный выпуск:*
*Сетевой анализ*

Научный журнал
Российской
академии наук

Основан в 1991 году

№ 56
2023

# Sociology:

## Methodology, Methods, Mathematical Modeling

*Special Issue: Network Analysis*

EDN: BTNCSJ

Федеральное государственное
бюджетное учреждение науки

Федеральный
научно-исследовательский
социологический центр
Российской академии наук

Журнал издается при финансовой поддержке
научно-исследовательского центра
«Демоскоп»

***Адрес редакции***: 117218, Москва, ул. Кржижановского, д. 24/35, корп. 5, комн. 214
**Тел.**: (499) 391-02-80. **E-mail**: sociology.4m@gmail.com

**Главный редактор – И.Ф. Девятко**
*НИУ ВШЭ; Институт социологии ФНИСЦ РАН* (*Москва*)

**Редактор спецвыпуска – Д.В. Мальцева**
*НИУ ВШЭ* (*Москва*)

**Редакционный совет**

**О.Б. Божков** *Социологический институт РАН – филиал ФНИСЦ РАН* (*Санкт-Петербург*)
**Е.Е. Горяченко** *ИЭОПП СО РАН; Новосибирский национальный исследовательский государственный университет* (*Новосибирск*)
**Ю.Н. Гаврилец** *Центральный экономико-математический институт РАН* (*Москва*)
**А.С. Готлиб** *Самарский государственный университет* (*Самара*)
**П.М. Козырева** *Федеральный научно-исследовательский социологический центр РАН* (*Москва*)
**М.С. Косолапов** *Институт социологии ФНИСЦ РАН* (*Москва*)
**В.А. Мансуров** *Институт социологии ФНИСЦ РАН* (*Москва*)
**А.Ю. Мягков** *Ивановский государственный энергетический университет* (*Иваново*)
**А.И. Орлов** *МГТУ им. Н.Э. Баумана* (*Москва*)
**А.П. Петров** *Институт прикладной математики им. М.В. Келдыша РАН*
**Г.И. Саганенко** *Социологический институт РАН – филиал ФНИСЦ РАН* (*Санкт-Петербург*)
**Г.А. Сатаров** *Фонд ИНДЕМ* (*Москва*)
**Г.Г. Татарова** *Институт социологии ФНИСЦ РАН* (*Москва*)
**Ю.Н. Толстова** *НИУ ВШЭ; Институт социологии ФНИСЦ РАН* (*Москва*) – зам. гл. редактора
**Т.Ю. Черкашина** *ИЭОПП СО РАН; Новосибирский национальный исследовательский государственный университет* (*Новосибирск*)
**В.А. Шведовский** *МГУ им. М.В. Ломоносова* (*Москва*)

Ответственный редактор – *К.А. Гаврилов*
Редактор – *В.В. Камышан*
Компьютерная верстка – *Н.К. Орлова*

**Editor-in-Chief – Inna F. Deviatko**
*NRU HSE; Institute of Sociology FCTAS RAS* (*Moscow*)

**Special Issue Editor – Daria V. Maltseva**
*NRU HSE* (*Moscow*)

**Editorial Board**

| | |
|---|---|
| **Oleg B. Bozhkov** | *Sociological Institute of the RAS – FCTAS RAS* (*Saint Petersburg*) |
| **Elizaveta E. Goryachenko** | *Institute of Economics and Industrial Engineering SB RAS; Novosibirsk State University* (*Novosibirsk*) |
| **Yuriy N. Gavrilets** | *Central Economics and Mathematics Institute RAS* (*Moscow*) |
| **Anna S. Gotlib** | *Samara State University* (*Samara*) |
| **Polina M. Kozyreva** | *Federal Center of Theoretical and Applied Sociology of the RAS* (*Moscow*) |
| **Mikhail S. Kosolapov** | *Institute of Sociology FCTAS RAS* (*Moscow*) |
| **Valeriy A. Mansurov** | *Institute of Sociology FCTAS RAS* (*Moscow*) |
| **Alexander Yu. Myagkov** | *Ivanovo State Power Engineering University* (*Ivanovo*) |
| **Alexander I. Orlov** | *Bauman University* (*Moscow*) |
| **Alexander P. Petrov** | *Keldysh Institute of Applied Mathematics RAS* |
| **Galina I. Saganenko** | *Sociological Institute of the RAS – FCTAS RAS* (*Saint Petersburg*) |
| **Georgy A. Satarov** | *Foundation for Information on Democracy* (*Moscow*) |
| **Galina G. Tatarova** | *Institute of Sociology FCTAS RAS* (*Moscow*) |
| **Yuliana N. Tolstova** | *NRU HSE; Institute of Sociology FCTAS RAS* (*Moscow*) – deputy editor |
| **Tatyana Yu. Cherkashina** | *Institute of Economics and Industrial Engineering SB RAS; Novosibirsk State University* (*Novosibirsk*) |
| **Vyacheslav A. Shvedovsky** | *Lomonosov Moscow State University* (*Moscow*) |

Managing Editor – *Kirill Gavrilov*
Copy Editor – *Victoria Kamyshan*
Layout Design – *Natalia Orlova*

# СОДЕРЖАНИЕ

# CONTENTS



© Социология: 4М. 2023. № 56

## *ОБЩИЕ ВОПРОСЫ МЕТОДОЛОГИИ СЕТЕВОГО АНАЛИЗА*

DOI: 10.19181/4m.2023.32.1.1
EDN: ZBAAGN

Д.В. Мальцева, И.А. Павлова,
Л.В. Капустина, В.А. Ващенко
(*Москва*)
Д. Фиала
(*Чехия*)

# СРАВНИТЕЛЬНЫЙ АНАЛИЗ ВОЗМОЖНОСТЕЙ WOS И ELIBRARY ДЛЯ АНАЛИЗА БИБЛИОГРАФИЧЕСКИХ СЕТЕЙ[1]

**Дарья Васильевна Мальцева** – кандидат социологических наук, заведующая Международной лабораторией прикладного сетевого анализа, Национальный исследовательский университет «Высшая школа экономики», Москва, Россия. Email: dmaltseva@hse.ru.
**Ирина Анатольевна Павлова** – кандидат экономических наук, старший научный сотрудник Международной лаборатории прикладного сетевого анализа, Национальный исследовательский университет «Высшая школа экономики», Москва, Россия. Email: iapavlova@hse.ru.
**Лика Владимировна Капустина** – стажер-исследователь Международной лаборатории прикладного сетевого анализа, Национальный исследовательский университет «Высшая школа экономики», Москва, Россия. Email: lkapustina@hse.ru
**Василиса Андреевна Ващенко** – стажер-исследователь Международной лаборатории прикладного сетевого анализа, Национальный исследовательский университет «Высшая школа экономики», Москва, Россия. Email: vvashchenko@hse.ru.
**Далибор Фиала** – доцент Факультета прикладных наук Департамента компьютерных наук и инженерии, Западночешский университет, Чехия, Пльзень. Email: dalfia@kiv.zcu.cz.

[1] Статья подготовлена в ходе проведения исследования в рамках Программы фундаментальных исследований Национального исследовательского университета «Высшая школа экономики» (НИУ ВШЭ).

В статье проводится сравнительный анализ баз данных научных публикаций Web of Science Core Collection и eLibrary с целью выделения их особенностей и описания возможностей анализа при изучении библиографических сетей российских авторов. Актуальность исследования определяется необходимостью адаптации и разработки подходов и инструментов для сбора, предобработки и анализа библиографических данных на русском языке. Анализ проводится на основе сравнения массивов данных публикаций в научных журналах в области социологии, выгруженных за период 2010–2021 гг. Выделяются основания для сопоставления двух баз, характеризующие получение доступа к данным, организацию данных в базах, количественные и содержательные характеристики данных. Анализ отобранных параметров позволяет найти пересечения между массивами данных и содержательными результатами анализа. Делаются выводы о соотношении двух баз, их возможностях и ограничениях по использованию в качестве основного (единственного) источника информации, даются рекомендации об их использовании для изучения отечественной науки.

*Ключевые слова:* сетевой анализ, сравнительный анализ, библиографические базы, библиографические сети, eLibrary, Web of Science

## Введение

Анализ библиографических сетей – частный случай применения методологии анализа социальных сетей. Он основан на построении и анализе сетей соавторства и коллаборации, цитирования и социтирования, библиографического сочленения, соприсутствия библиометрических единиц анализа. Направление способно показать закономерности развития взаимодействия в научном сообществе, определить его структуру, динамику, направления исследований [1; 2; 3]. Основные этапы исследования с применением анализа библиографических сетей подразумевают использование технологических решений для 1) формирования базы библиографических данных, 2) ее предобработки и постро-

ения различных видов библиографических сетей и 3) последующего изучения с применением методов сетевого анализа (social network analysis).

Как и в любом исследовании, выбор источника информации является определяющим для качества анализа – по принципу GIGO[1], получение достоверных результатов напрямую зависит от стратегии поиска и полноты используемой базы данных. Выбор баз данных для исследователя является достаточно широким – помимо часто используемых для учета эффективности работы ученых баз научного цитирования Web of Science (WoS) и Scopus, большую популярность приобрели бесплатные базы данных, агрегирующие библиографическую информацию, такие как «универсальные» Google Scholar и OpenAlex, включающие информацию о патентах Digital Science Dimensions и Lens, базы медицинских исследований PubMed и Cochrane, научные социальные медиа SciFinder, Mendeley и др. Значительное количество баз данных научных публикаций привело к появлению исследований, посвященных сравнительному анализу различных площадок, где они сравниваются по различным характеристикам.

Международные базы данных могут выступать источниками информации и при изучении научного производства российских авторов. Данные из WoS и Scopus до недавнего времени[2] активно использовались для оценки научной продуктивности ученых в рамках ряда государственных программ финансирования университетов (например, «Проект 5–100»). Однако механика формирования этих баз предполагает неполное покрытие всей

[1] GIGO – аббревиатура для используемой в информатике фразы “garbage in, garbage out”, означающей, что при неверных входящих данных будут получены неверные результаты, даже если сам по себе алгоритм правилен. В русском языке аналогом является пословица «что посеешь, то и пожнешь».

[2] До момента прекращения работы данных организаций в России в связи с началом проведения СВО в 2022 г.

научной продукции отдельной страны: представленность индексируемых в базах журналов (и издательств) является выборочной и основана на наукометрических показателях. В связи с этим при использовании международных баз данных в качестве источника объем производимой в России научной литературы оказывается недооцененным.

В поле российской науки имеется несколько баз данных, аккумулирующих информацию о научных публикациях отечественных исследователей. Крупнейшей в России базой научных публикаций является научная электронная библиотека eLibrary, которая интегрирована с Российским индексом научного цитирования (РИНЦ) – созданной по заказу Минобрнауки РФ общедоступной базой данных российских научных публикаций. В качестве альтернативы можно отметить научную библиотеку КиберЛенинка, предоставляющую доступ к публикациям на основе принципов открытой науки. Эти базы данных аккумулируют значительное число публикаций, производимых российскими авторами (в том числе и через интеграцию с международными базами научного цитирования и получение информации о публикациях в международных журналах[1]), и могут рассматриваться как источники данных для изучения российской науки.

В поле библиометрического анализа разработаны специальные инструменты, которые позволяют производить анализ библиографических сетей на основе данных из международных наукометрических баз: производить предобработку полученных данных, строить различные виды сетей и затем анализировать их с помощью методологии сетевого анализа. Пакеты Bibliometrix для R и Python и их веб-приложение Biblioshiny[2] позволяют

[1] Авторы не имеют данных о том, продолжается ли эта интеграция с 2022 г. по настоящее время.

[2] Bibliometrix – An R-tool for comprehensive science mapping analysis [site]. URL: https://www.bibliometrix.org/home/ (date of access: 08.04.2024).

работать с базами данных Scopus, WoS, PubMed, Digital Science Dimensions, Cochrane, Lens и OpenAlex для анализа социтирования, библиографического сочленения библиографических единиц анализа, соавторства и соприсутствия ключевых слов. Программа VOSviewer[1], помимо перечисленных, позволяет работать с такими базами, как Crossref, Europe PMC, Semantic Scholar, OpenCitations и WikiData через их API-сервисы, запрашиваемые в интерактивном режиме в самой программе (осуществляя таким образом и сбор данных). Программа CitNetExplorer[2], предназначенная для анализа цитирований научной литературы, импортирует данные из WoS. На использование данных WoS ориентирован и методологический подход, разработанный В. Батагелем, А. Ферлигой и П. Дореаном (подробнее о подходе см.: [4]), применявшийся для анализа некоторых зарубежных научных дисциплин и описанный в отечественной литературе [5, 6], который использует программу WoS2Pajek для создания из данных WoS сетевых файлов для работы в программе для анализа и визуализации больших сетей Pajek[3].

В связи с тем, что указанные инструменты используют англоязычные коллекции словарей для предобработки и нормализации данных (дизамбигуации имен, лемматизации и токенизации слов), они могут применяться для анализа только части работ российских авторов, опубликованных в международных базах на английском языке. Если же речь идет о публикациях на русском языке, то использование этих инструментов затруднено и должно сопровождаться их адаптацией. Единственным инструментом, позволяющим измерять публикационную активность ученых и ор-

[1] VOSviewer. Visualizing scientific landscapes [site]. URL: https://www.vosviewer.com/ (date of access: 08.04.2024).

[2] CitNetExplorer. Analyzing citation patterns in scientific literature [site]. URL: https://www.citnetexplorer.nl/ (date of access: 08.04.2024).

[3] Pajek. Analysis and visualization of very large networks [site]. URL: http://mrvar.fdv.uni-lj.si/pajek/ (date of access: 08.04.2024).

ганизаций в русскоязычном научном пространстве, является информационно-аналитическая надстройка Science Index, реализованная на платформе eLibrary и основанная на анализе внесенных в базу публикаций. Однако ни функции сбора данных, ни инструменты для сетевого библиометрического анализа на площадке не представлены. Таким образом, в отечественном научном пространстве отсутствует полноценная методология по сбору, предобработке и анализу библиографических данных на русском языке.

В текущей ситуации, связанной со сложностями применения привычных инструментов оценки публикационной продуктивности российских авторов, актуализируется задача по изучению различных баз данных с точки зрения их возможностей для наукометрического анализа. В данной статье мы фокусируемся на сравнительном анализе двух научных баз – международной базы WoS и российской базы eLibrary – для изучения возможностей, которые они предоставляют для анализа библиографических сетей российских авторов. Тогда как использование базы WoS как источника данных позволяет использовать ряд предложенных методологических решений по обработке и анализу данных, выбор базы eLibrary в качестве источника данных предполагает необходимость адаптации существующих и разработки новых технологических решений. Сравнение двух баз проводится посредством сравнения их контента – формата, полноты представления и объема библиографических данных по публикациям российских авторов (насколько похожи данные), а также содержательных характеристик, получаемых при сетевом библиографическом анализе (насколько похожи результаты анализа). В качестве примера взяты все публикации в области социологии за 2010–2021 гг. на обеих площадках – массивы данных из 3995 публикаций в WoS и 75 232 публикаций в eLibrary.

В результате литературного обзора выделяются параметры для сравнения двух баз данных, которые затем анализируются посредством описательного, статистического и сетевого анализа.

Статистический и сетевой анализ основных библиометрических единиц (публикаций, авторов, соавторов, журналов, ключевых слов) и базовых производных сетей для обоих массивов позволяет сравнивать распределения и рейтинги изучаемых библиографических единиц и делать выводы о содержательных различиях между массивами данных. Анализ позволяет делать выводы о соотношении двух баз данных, их возможностях и ограничениях по использованию в качестве основного (единственного) источника информации и давать рекомендации об их использовании для изучения отечественной науки.

## Сравнение баз данных научных публикаций

Платформа WoS компании Clarivate Analytics является первой базой научного цитирования, построенной на основе Индекса цитирования научных статей (Science Citation Index), разработанного в 1960-е гг. одним из основателей наукометрии Ю. Гарфилдом. На основе анализа цитирований статей Гарфилд разработал подход к рейтингованию научных журналов, составляющих «ядро» научных дисциплин (Core Collection – CC), в котором позже появились и другие коллекции научных публикаций[1]. Долгое время WoS имела монополию на предоставление информации о научной литературе, однако с появлением в 2004 г. платформ Scopus и Google Scholar ситуация изменилась. Scopus, как и WoS, стал предоставлять информацию о цитировании, получаемую в виде метаданных от производителей научной литературы, однако существенно расширил покрытие научных журналов. Google Scholar расширил диапазон источников до материалов конфе-

[1] Индексы цитирования социальных наук (Social Sciences Citation Index – SSCI), искусств и гуманитарных наук (Arts and Humanities Citation Index – AHCI), новых источников (Emerging Sources Citation Index – ESCI), конференционных публикаций и книг.

ренций, книг, диссертаций, отчетов и других типов публикаций с сайтов издателей и конференций, используя автоматические методы извлечения информации из электронных файлов научных публикаций. С 2004 г. появилось много других агрегаторов научной информации, таких как OpenAlex (универсальная база), Digital Science Dimensions и Lens (начинались как патентные базы), PubMed и Cochrane (медицинские исследования), SciFinder, Mendeley, ResearchGate (научные социальные медиа) и др.

В обзорах развития наукометрических исследований [1; 2] приводится масса ссылок на исследования, сравнивающие базы данных WoS, Scopus и Google Scholar друг с другом, а также с более новыми базами. Рассмотреть все эти публикации не представляется возможным ввиду их большого количества, однако отметим ниже некоторые выводы этих исследований, важные для нашего анализа, и проиллюстрируем их примерами.

Наличие существенных различий между WoS, Scopus и Google Scholar по охвату научных дисциплин было подтверждено во многих исследованиях (см., напр.: [7], см. также: [1; 2]); особенно «проседающими» для первых двух баз являются области социальных и гуманитарных наук. Исследователи делают выводы, что новые базы Microsoft Academic и Dimensions способны выступать альтернативой WoS и Scopus с точки зрения публикационного охвата [8; 9; 10], однако при этом WoS и Scopus по-прежнему остаются самыми популярными источниками информации для наукометрических исследований [11].

Для целей настоящего исследования важно остановиться на исследовательском дизайне работ, посвященных сравнению различных баз данных, и определить, какие параметры обычно выступают основаниями для сравнения. Наиболее прямым методом сравнения охвата документов из разных источников данных является получение полных списков всех документов, их сопоставление и оценка размера совпадений, что сложно осуществимо в связи с объемом данных и приводит к необходимости формиро-

вания выборок [9]. С точки зрения формирования массивов данных для анализа единицами анализа могут выступать *публикации*, выборки которых формируются на основе идентичных поисковых запросов по ключевым словам или через сплошной сбор по отобранным журналам, научным дисциплинам, областям наук, группам авторов, университетам или странам (но иногда отбор происходит и в случайном порядке), а также *журналы*, индексируемые в базах, выборки которых формируются экспертным образом. Как правило, для анализа задается определенный период времени; начальной точкой часто выступает год запуска самой новой площадки, участвующей в сравнении. В связи с этим выборки часто «гетерогенно разнообразны» [9] и ограничены по объему, а сравнение осуществляется по ограниченному числу научных баз (чаще всего – двум или трем площадкам). Однако есть и довольно обширные исследования – например, сравнение WoS, Scopus и Google Scholar по 37 научным направлениям в динамике [7], систематическое сравнение трех упомянутых баз и Microsoft Academic, Dimensions и OpenCitations' COCI по 252 категориям [9] или сравнение 12 академических поисковых систем и библиографических баз данных [12]. Еще один возможный вариант формирования выборки – использование канонического набора публикаций как исходной выборки документов ("seed sample"), для которого во всех анализируемых базах находятся все цитирующие их документы, которые и становятся выборками по каждой базе [9]. Полученные различными способами массивы данных могут сравниваться для изучения покрытия баз данных по следующим параметрам.

Характеристики базы [12]:

- *тип базы и доступа* – библиографическая база, поисковая система, агрегатор; платный/бесплатный доступ; открытый /закрытый / доступный по запросу контент; владелец;
- *функциональные возможности* поиска, анализа и экспорта данных, транспарентности алгоритмов (например, обработки

запросов и ранжирования документов на странице результатов);
– *охват* – типы индексируемых документов, скорость их индексации, предметный охват, годы покрытия, доступные поля для поиска (метаданные), качество метаданных (например, соотнесение разделения по предметным областям в разных базах [8]).

Объем:
– *по публикациям* (количество отобранных документов) – например, распределение числа публикаций, найденных по запросу или через исходную выборку документов, в том числе во временной перспективе [7; 9; 10; 12], и подсчет ежегодных темпов роста, в том числе по российским публикациям [13];
– *по журналам* (количество отобранных журналов) – количество журналов, входящих в основные списки журналов (master lists) в анализируемых базах [8].

Распределение публикаций:
– *по журналам* – абсолютное количество и доли публикаций в анализируемых журналах в анализируемых базах [10; 11];
– *по типам документов* – абсолютное количество и доли публикаций типа «статья», «ревю», «глава в книге» и др. в анализируемых базах [8; 10; 13];
– *по предметным областям* – абсолютное количество, доли, среднее число, темпы роста публикаций по основным научным областям / предметным категориям [7; 9; 11] в анализируемых базах, в том числе с учетом распределения по 20 отобранным для анализа странам [8] и используемым языкам [14], а также для российских публикаций [13];
– *по странам* – абсолютное количество публикаций в анализируемых базах, их доли в общем объеме научных исследований в мире, годовые темпы роста для 20 отобранных стан [8] или стран, лидирующих по этим показателям [11; 13];
– *по языку* – доли публикаций на разных языках в общем объеме публикаций в анализируемых базах с учетом их исследовательских направлений или во временной перспективе [14], в том числе русскоязычных и нерусскоязычных [13];
– *по количеству полученных внутри базы цитирований*;
– *по типу коллаборации* – долям публикаций, написанных в коллаборации, в том числе во временной перспективе [13].

Распределение авторов:

– *по полученным внутри базы цитированиям* (и вариациям этой метрики) – среднее число полученных авторами цитирований по различным дисциплинам в анализируемых базах [7];
– *по специализированным индексам оценки исследователей, таким как Индекс Хирша (и вариациям этой метрики),* – например, средние h- и hIa-индексы для исследователей, сгруппированных по научным дисциплинам в анализируемых базах [7];
– *по странам*.

Распределение журналов:

– *по полученным внутри базы цитированиям* – количество цитирований журналов и доля от максимального числа цитирований в анализируемых базах [10];
– *по дисциплинам*.

Пересечение между массивами:

– *по публикациям* – доли пересечений между массивами публикаций, найденные через сопоставление по DOI или названию и авторам (например, с помощью расстояния Дамерау-Левенштайна) [9];
– *по журналам* – доли пересечений между списками журналов через сопоставление по названиям / аббревиатурам / ID журналов [8].

В контексте изучения национальных корпусов научной литературы важно остановиться на вопросе представленности неанглоязычных авторов в международных базах. WoS CC включает национальные индексы цитирований для Китая, Латинской Америки и Южной Африки, Кореи, России и арабского региона, которые формируются в кооперации с представителями этих стран и потенциально должны приводить к большей представленности национальных корпусов. Большое значение для представленности работ имеют официальные языки публикаций, принятые на площадках. Интересно, что начало использования русского языка как публикационного в Scopus привело к росту доли работ российских авторов с 4,8 до 14,8% в 2006–2016 гг., что во многом объясняет экспоненциальный рост числа отечественных публикаций, наблюдаемый в эти годы [13]. Вместе с тем исследования, посвящен-

ные изучению вопросов покрытия WoS и Scopus по различным странам и языкам публикации (см., напр., обзор в работе: [14]), показали чрезмерную представленность в этих базах журналов на английском языке и публикаций из англоязычных стран (более 92% в 2018 г.). Вторым языком в 2018 г. был назван китайский, третьим – испанский. Однако даже в случае представленности языка в базе библиографические данные неанглоязычных авторов могут содержать ошибки. Так, например, исследователи обнаружили, что для испанского языка около 50% имен авторов имеют несколько вариаций написания даже внутри одной и той же международной базы [15]. Если же речь идет о разных базах, то представленные в них библиографические описания одних и тех же статей тоже не всегда консистентны (например, это показывает анализ публикаций ученых Южной Африки [16]).

Крупнейшей базой научных публикаций в России, а также научной периодики на русском языке в мире является научная электронная библиотека eLibrary. Как было сказано выше, эта платформа интегрирована с Российским индексом научного цитирования (базой РИНЦ), куда входят публикации в индексируемых (соответствующих определенным критериям качества) российских научных журналах и непериодических изданиях. С 2016 г. часть публикаций из РИНЦ (порядка 600–700 лучших российских журналов по всем научным направлениям за последние 10 лет) индексируются в WoS в виде отдельной базы Russian Science Citation Index (RSCI). Вместе с публикациями российских авторов, индексируемыми в Scopus и WoS, база RSCI составляет «ядро РИНЦ» – более узкую базу по отношению к РИНЦ. Сама площадка eLibrary при этом шире, чем РИНЦ, так как содержит неиндексируемые в этой базе публикации из изданий, имеющих заключенный с площадкой договор.

Несмотря на наличие отдельной национальной базы научных публикаций, eLibrary в какой-либо из трех вариаций (от максимально полной до ядра РИНЦ) редко выступает источником

для сравнения с другими площадками. Например, в исследовании коллекций публикаций российских авторов [13] сравниваются коллекции публикаций из баз WoS CC и Scopus, но отдельно подчеркивается, что базы RSCI и РИНЦ в анализе не участвуют. Сравнение RSCI с международными базами проводилось аналитиками eLibrary [17] посредством анализа массивов публикаций российских авторов в «квартильных» журналах WoS CC и Scopus, в базе новых источников ESCI (Emerging Sources Citation Index) из ядра WoS и базы RSCI. По всем массивам было подсчитано число публикаций; на основании информации об индексации журналов в различных базах были найдены пересечения между массивами; для RSCI и ESCI и для выделенных в WoS и Scopus групп по квартилям были подсчитаны средние показатели цитирования внутри базы. Проведенный анализ позволил говорить только о частичном пересечении коллекций журналов в трех базах и об оригинальности значительной части контента базы RSCI и большом вкладе в ядро РИНЦ.

Отдельно базы РИНЦ и RSCI (также встречается название RSCI-C – от названия компании Clarivate [13]) выступают источниками данных в библиометрических исследованиях, ориентированных на изучение российского научного поля (см., напр.: [18; 19]). Однако методология сбора данных для баз РИНЦ и RSCI детально не описывается. Отдельным методологическим затруднением может быть то, что на английский язык «Российский индекс научного цитирования» переводят не только как “Russian Index of Science Citation” (RISC), но и как “Russian Science Citation Index” (RSCI), что повторяет название в базе WoS – хотя, как видно из обзора, это хоть и смежные, но разные базы.

Рассмотренные обзорные работы, сравнивающие разные базы на уровне стран (см.: [8]), показывают, что использование информации из разных баз данных может привести к различным результатам библиометрической оценки деятельности национальных научных коллективов. Так, исследователи пришли к выводу,

что и база WoS, и особенно Scopus должны с осторожностью применяться в качестве единственных инструментов измерения результативности неанглоязычных исследователей, в том числе российских [13]. В связи с различиями в покрытии исследователи рекомендуют использовать *несколько* баз данных для формирования наиболее полного массива исследования. В случае изучения развития науки в конкретных странах рекомендуется формировать массив данных из публикаций в международных и национальных базах научного цитирования. Первым шагом для проведения таких исследований является сравнение международных и национальных баз друг с другом. Такой анализ и предпринимается в данной статье.

## Методология и данные исследования

*Методология*

На основе рассмотренных параметров сравнения баз данных научных публикаций, а также предварительного анализа рассматриваемых массивов были сформулированы основания и параметры для сравнения баз WoS и eLibrary (табл. 1), которые касаются: 1) получения доступа к данным, 2) организации данных в базах, 3) количественных характеристик (объем, динамика и т.д.) и 4) содержательных характеристик указанных данных. Для описания особенностей доступа к данным и организации данных в базах использовался описательный анализ доступной информации. Согласно стратегиям исследований, рассмотренных в обзоре, дизайн исследования для более детального анализа подразумевал сравнение аналогичных массивов данных, выгруженных из каждой базы. Два массива рассматривались по определенным параметрам с помощью статистического анализа, затем на основе данных из массивов было построено несколько базовых библиометрических сетей, которые также рассматривались в соотношении друг с другом по ряду рассчитанных параметров.

Анализ указанных параметров позволяет найти пересечения между массивами с точки зрения присутствующих в них единиц анализа – публикаций, журналов, авторов, ключевых слов – и оценить размер множеств, находящихся на пересечении и

*Таблица 1*

ОСНОВАНИЯ, ПАРАМЕТРЫ И СПОСОБЫ АНАЛИЗА ДЛЯ СРАВНЕНИЯ БАЗ ДАННЫХ

| Основания | Параметры | Способы анализа |
|---|---|---|
| Получение доступа к данным | Особенности и возможности сбора данных исследователем | Описательный анализ |
| Организация данных в базах | Формат и структура получаемых массивов, в том числе количество используемых в библиографических описаниях метаданных.<br>Возможности предобработки данных существующим программным обеспечением.<br>Возможности построения файлов для сетевого библиометрического анализа | Описательный анализ |
| Количественные характеристики данных | Объем массивов (число публикаций).<br>Динамика числа публикаций во времени.<br>Объем пропущенных значений по метаданным в библиографических описаниях публикаций.<br>Количество уникальных библиометрических единиц (публикаций, авторов, журналов, ключевых слов) | Статистический анализ |

*Окончание табл. 1*

| Основания | Параметры | Способы анализа |
|---|---|---|
| Содержательные характеристики данных | Распределение уникальных авторов и ключевых слов по частоте встречаемости в публикациях в массиве.<br>Распределение соавторов по авторам в массиве.<br>Наиболее частотные журналы, ключевые слова.<br>Наиболее популярные авторы по числу работ и числу соавторов | Статистический и сетевой анализ двумодальных сетей работ и авторов WA, работ и ключевых слов WK, одномодальной сети коллабораций Co. |

при объединении массивов. Помимо этого, сравнение результатов анализа с точки зрения содержания также позволяет сделать выводы о том, насколько похожие результаты дает использование двух баз данных.

Методология анализа данных для количественной оценки по выделенным основаниям подразумевала использование различных инструментов статистического и сетевого анализа. Процедура анализа и использованное программное обеспечение представлены в тексте ниже при сравнении показателей.

С точки зрения используемых инструментов анализа подсчет общей статистики по набору данных WoS проводился с помощью пакета Bibliometrix в R и его приложения Biblioshiny.

*Данные*

В статье сравниваются публикации российских социологов на площадках WoS Core Collection и eLibrary. Единицами анализа являются научные статьи в научных журналах в области социологии. Оба массива данных включают публикации за 2010–2021 гг. Для сбора данных на двух площадках использовались идентичные

стратегии. Ниже представлена информация о деталях сбора, размере массивов и формате итоговых данных.

**Стратегия сбора данных и размеры массивов.** Анализируемый массив данных собран в рамках проекта, выполняемого в рамках гранта РНФ[1]. При сборе данных из eLibrary работа проводилась совместно с сотрудниками ООО «Научная электронная библиотека», осуществляющими поддержку этой базы. Поиск работ проводился по всем научным журналам, представленным на сайте eLibrary (имеющим заключенный договор). Из всех работ, относящихся по ГРНТИ к рубрике «Социология», были отфильтрованы публикации типа «научная статья», где по крайней мере одним из авторов является российский ученый (в поле «страна» указано «Россия»). По данному запросу был составлен список из 75 232 уникальных идентификаторов публикаций, по которым затем была собрана полная библиографическая информация. В сравниваемый массив данных eLibrary вошло 75 232 публикации.

Анализируемый массив данных WoS CC является подмножеством из набора данных, собранных в рамках проекта по изучению российской науки на основе всех российских публикаций, представленных в WoS CC[2] (1 383 996 библиографических записей о российских публикациях по всем наукам за период с 1992 г. до мая 2022 г.). Стратегия сбора данных этого массива подразумевала использование базы Core Collection. Были отобраны и выгружены все публикации российских авторов (поле CU = “Russia”). Затем на основе категории (“research area”), к которой относится публикация, было выделено подмножество по социологии (поле

[1] Проект «Паттерны коллаборации в российском социологическом сообществе: структура научных школ и возможные точки роста» выполнен в рамках гранта Российского научного фонда в 2021–2023 гг. под руководством Д.В. Мальцевой.

[2] Проект осуществляется совместно Д. Фиалой, отвечающим за сбор и исследовательский анализ данных, и коллективом МЛ ПСА под руководством Д.В. Мальцевой, отвечающим за сетевой анализ массива.

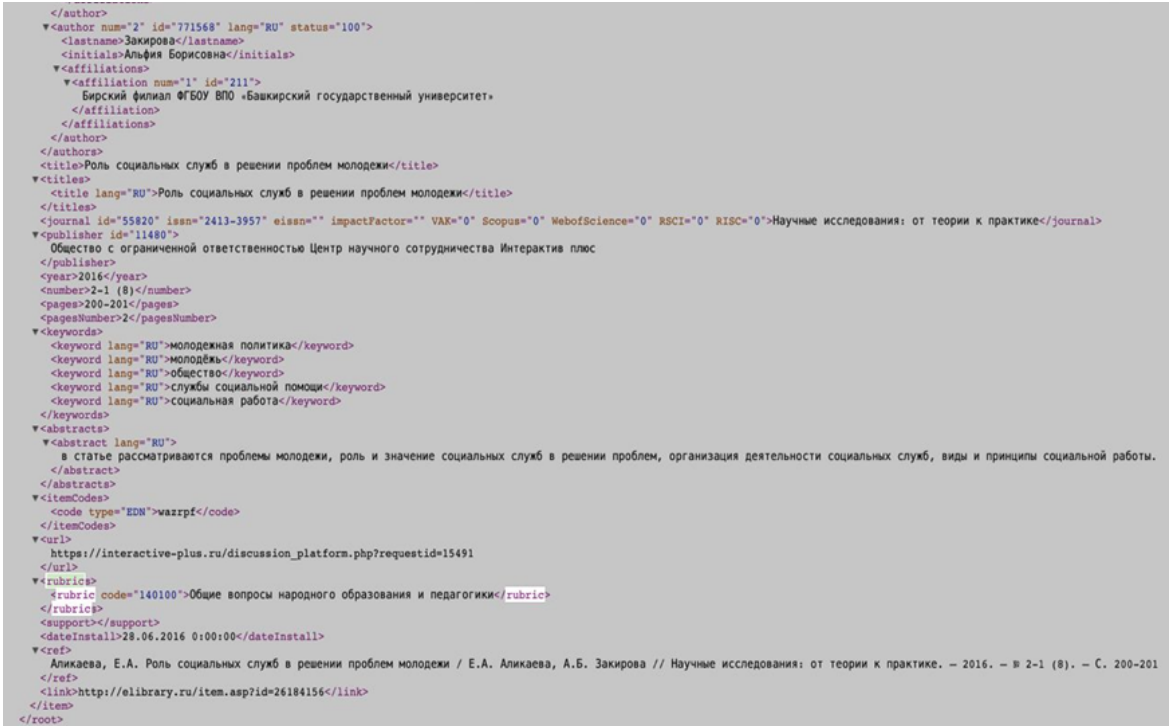
SC = “Sociology”). Первоначально массив состоял из 7915 публикаций, но для целей настоящего анализа он был ограничен по типу публикаций (поле DT = “Article”) и временному периоду (2010–2021 гг.), что дало 3559 научных публикаций, которые и вошли в сравниваемый массив данных WoS CC.

**Выгрузка данных.** База данных eLibrary не предоставляет функциональных возможностей для выгрузки библиографических описаний публикаций с сайта или публичного адреса API-сервиса для автоматизированного парсинга (сбора и структурирования информации) данных. Доступ к закрытому API-сервису возможен при наличии договора с ООО «НЭБ», который был заключен в рамках исследования. Для сбора данных, а также их предобработки и построения сетевых файлов использовался разработанный авторами статьи методологический подход [20], реализованный в программе Bib-eLib[1]. Используя полученный список из идентификаторов публикаций через соответствующий сервис API с помощью специально написанного парсера, делались запросы на информацию по каждой публикации. Выдача данных представляет собой XML-страницу структурированного вида с идентифицированными полями (рис. 1).

Выгрузка данных осуществлялась из нужных полей, а затем записывалась в единый файл формата csv. Данные собраны в октябре 2022 г.

Платформа WoS дает возможность выгрузки библиографических описаний отобранных работ в различных форматах (RIS, Excel, обычный текстовый файл – plain text) для всех зарегистрированных пользователей, чей доступ осуществляется через

[1] Программа для ЭВМ Bib-eLib для сбора и обработки библиографических данных на русском языке из электронной библиотеки eLibrary на языке программирования Python зарегистрирована в виде РИД (свидетельство о государственной регистрации программы для ЭВМ № 2023684182, регистрация в реестре программ для ЭВМ 14.11.2023) и доступна в репозитории платформы GitHub по ссылке: https://github.com/Daria-Maltseva/Collaboration (дата обращения: 08.04.2024).



```
    </author>
  ▼<author num="2" id="771568" lang="RU" status="100">
      <lastname>Закирова</lastname>
      <initials>Альфия Борисовна</initials>
    ▼<affiliations>
      ▼<affiliation num="1" id="211">
          Бирский филиал ФГБОУ ВПО «Башкирский государственный университет»
        </affiliation>
      </affiliations>
    </author>
  </authors>
  <title>Роль социальных служб в решении проблем молодежи</title>
▼<titles>
    <title lang="RU">Роль социальных служб в решении проблем молодежи</title>
  </titles>
  <journal id="55820" issn="2413-3957" eissn="" impactFactor="" VAK="0" Scopus="0" WebofScience="0" RSCI="0" RISC="0">Научные исследования: от теории к практике</journal>
▼<publisher id="11480">
    Общество с ограниченной ответственностью Центр научного сотрудничества Интерактив плюс
  </publisher>
  <year>2016</year>
  <number>2-1 (8)</number>
  <pages>200-201</pages>
  <pagesNumber>2</pagesNumber>
▼<keywords>
    <keyword lang="RU">молодежная политика</keyword>
    <keyword lang="RU">молодёжь</keyword>
    <keyword lang="RU">общество</keyword>
    <keyword lang="RU">службы социальной помощи</keyword>
    <keyword lang="RU">социальная работа</keyword>
  </keywords>
▼<abstracts>
  ▼<abstract lang="RU">
      в статье рассматриваются проблемы молодежи, роль и значение социальных служб в решении проблем, организация деятельности социальных служб, виды и принципы социальной работы.
    </abstract>
  </abstracts>
▼<itemCodes>
    <code type="EDN">wazrpf</code>
  </itemCodes>
▼<url>
    https://interactive-plus.ru/discussion_platform.php?requestid=15491
  </url>
▼<rubrics>
    <rubric code="140100">Общие вопросы народного образования и педагогики</rubric>
  </rubrics>
  <support></support>
  <dateInstall>28.06.2016 0:00:00</dateInstall>
▼<ref>
    Аликаева, Е.А. Роль социальных служб в решении проблем молодежи / Е.А. Аликаева, А.Б. Закирова // Научные исследования: от теории к практике. – 2016. – № 2-1 (8). – С. 200-201
  </ref>
  <link>http://elibrary.ru/item.asp?id=26184156</link>
</item>
</root>
```

***Рис. 1. Пример библиографического описания публикаций в eLibrary (XML-страница)***

Document1 * ×

```
PT J
AU Toshchenko, ZT
AF Toshchenko, ZT
TI Sociology of life as a concept to study social reality
SO SOTSIOLOGICHESKIE ISSLEDOVANIYA
LA Russian
DT Article
AB "Sociology of life as a concept to study social reality" begins with an analysis of reasons for absence of sociological schools in contemp
C1 Russian Acad Sci, Moscow, Russia.
RP Toshchenko, ZT (corresponding author), Russian Acad Sci, Moscow, Russia.
CR ANDRUSHENKO VP, 1996, SOTSIOLOGIYA NAUKA O
   BARAZGOVA ES, 1997, SOTSIS
   ...
   YADOV VA, 1998, STRATEGIYA SOTSIOLOG
NR 23
TC 5
Z9 12
U1 0
U2 3
PU IZDATELSTVO NAUKA
PI MOSCOW
PA PROFSOYUZNAYA UL 90, MOSCOW, 117864, RUSSIA
SN 0132-1625
J9 SOTSIOL ISSLED+
JI Sotsiologicheskie Issled.
PY 2000
IS 2
BP 3
EP +
PG 11
WC Sociology
WE Social Science Citation Index (SSCI)
SC Sociology
GA 292NH
UT WOS:000085798200001
DA 2022-05-09
ER
```

***Рис. 2. Пример библиографического описания публикаций в WoS (текстовый файл)***

организационную подписку. Используемый формат представляет собой текстовый файл структурированного вида с идентифицированными полями (рис. 2).

В зависимости от того, собирается или нет информация о цитируемой литературе (поле “CR”), за одну итерацию можно загрузить до 500 или 1000 библиографических описаний в едином файле. Для сбора данных был написан программный код на Python, который позволял итеративно обращаться к базе и последовательно собирать файлы с описаниями в формате .txt, которые затем были автоматически собраны в единый файл. Данные собраны в мае 2022 г.

**Формат и структура данных.** Собранный массив данных eLibrary представлен в виде таблицы в формате .csv, в которой приведена информация по всем полям, доступным к выгрузке, для всех 75 232 публикаций. Данные WoS CC представлены в виде единого текстового файла в формате .txt с полным библиографическим описанием 3559 публикаций, включая пристатейные списки литературы. В обоих массивах содержится информация по таким библиографическим единицам как публикация, автор(ы) и журнал. Набор метаданных, которыми описываются библиографические единицы в каждом массиве, приведен в табл. 2.

*Таблица 2*

МЕТАДАННЫЕ БИБЛИОГРАФИЧЕСКИХ ОПИСАНИЙ
В WoS И eLibrary

| № | Параметр | WoS | eLibrary |
|---|---|---|---|
| Публикация | | | |
| 1 | ID публикации | + | + |
| 2 | Название на русском языке | - | + |
| 3 | Название на английском языке | + | + |
| 4 | DOI публикации | + | + |
| 5 | Дата публикации (год) | + | + |
| 6 | Предметная область | + | + |

*Окончание табл. 2*

| № | Параметр | WoS | eLibrary |
|---|---|---|---|
| 7 | Язык публикации | + | - |
| 8 | Тип публикации | + | - |
| 9 | Количество страниц (начальная и конечная страницы) | + | + |
| 10 | Число цитирований | + | + |
| 11 | Число использований (доступ, скачивание) | + | - |
| 12 | Аннотация на русском языке | - | + |
| 13 | Аннотация на английском языке | + | + |
| 14 | Ключевые слова на русском языке | - | + |
| 15 | Ключевые слова на английском языке | + | + |
| 16 | Информация о финансировании | - | + |
| 17 | Гиперссылка на статью в базе | - | + |
| 18 | Библиографическое описание | - | + |
| 19 | Список цитируемой литературы | + | |
| 20 | Количество процитированной литературы | + | |
| Автор(ы) | | | |
| 1 | Фамилия и инициалы на русском языке | - | + |
| 2 | Фамилия и инициалы на английском языке | + | + |
| 3 | ID автора | - | + |
| 4 | Название аффилиации автора | + | + |
| 5 | ID аффилиации автора | - | + |
| 6 | Местоположение (страна, город) | + | - |
| Журнал | | | |
| 1 | Название журнала | + | + |
| 2 | ID журнала | - | + |
| 3 | ISSN/e-ISSN | + | + |
| 4 | Импакт-фактор | - | + |
| 5 | Включенность в другие базы данных | +- | + |
| 6 | Выпуск | + | + |
| 7 | Номер | + | + |
| 8 | Название издательства | + | + |
| 9 | ID издательства | - | + |
| 10 | Адрес издательства (город и почтовый адрес) | + | - |

В целом можно видеть, что базы похожи по используемым ими метаданным. Очевидное отличие базы WoS CC заключается в том, что в ней не представлена информация на русском языке (название, аннотация, ключевые слова, имя автора). В этой базе также отсутствует информация об ID авторов и их организаций, ID журналов и издательств – хотя она может дать важные идентифицирующие признаки при решении проблемы дизамбигуации единиц анализа (но нужно отметить, что для журналов WoS помимо полного названия приводит и два вида его стандартизированной аббревиатуры, что может быть использовано для обозначенной цели). В этом смысле наличие ID в описаниях eLibrary выгодно отличает эту базу и предоставляет исследователям дополнительные аналитические возможности. В данных eLibrary также указываются базы, в которые входит публикация (РИНЦ, RSCI, WoS, Scopus и ВАК); в WoS предоставляется информация только о базе внутри ядра CC, к которой принадлежит публикация.

Главным выгодным отличием базы WoS CC является наличие информации о цитируемой литературе (поле “CR”), что позволяет проводить определенные виды анализа – изучать сети цитирований, социтирований, библиографического сочленения между различными библиографическими единицами (авторами, публикациями, журналами и т.д.). В обеих базах подсчитывается также число цитирований, полученных внутри данной базы и других баз. Помимо цитирования, в WoS есть также показатель по использованию публикации другими авторами за определенные периоды времени (доступу к тексту и загрузке), что также показывает внешний интерес к научной работе.

**Предобработка данных.** Поскольку библиографические описания публикаций не всегда содержат полную информацию или приведенная информация может содержать ошибки [20], для повышения качества дальнейшего анализа перед построением сетевых файлов возникает задача по предобработке данных, т.е. устранению пропущенных значений и приведению единиц

анализа к единому виду. Предложенная в авторской методологии [18] логика предобработки данных массива eLibrary отчасти следует логике работы с данными WoS, поэтому вначале рассмотрим предобработку данных, реализованную для массива данных WoS.

Предварительный исследовательский анализ данных массива WoS в программе Biblioshiny (табл. 3) показал, что значительная часть данных отсутствует в полях аннотаций и ключевых слов (поле “DE”, Keywords – около 21%) и в полях с дополнительными ключевыми словами (поле “ID”, Keywords Plus и DOI – 50 и 79% соответственно).

*Таблица 3*

ОЦЕНКА ПОЛНОТЫ БИБЛИОГРАФИЧЕСКИХ МЕТАДАННЫХ В МАССИВАХ WoS И eLibrary: ПРОПУЩЕННЫЕ ЗНАЧЕНИЯ

| Пропущенные элементы метаданных | WoS | | eLibrary | |
|---|---|---|---|---|
| | число | доля, % | число | доля, % |
| Публикация | | | | |
| Название – английский язык | 0 | 0 | 67 048 | 89,1 |
| Название – русский язык | | | 0 | 0 |
| Аннотация – английский язык | 737 | 20,71 | 27 739 | 36,9 |
| Аннотация – русский язык | | | 27 751 | 36,9 |
| Ключевые слова – английский язык | 774 | 21,75 | 27 600 | 36,7 |
| Ключевые слова – русский язык | | | 11 568 | 15,4 |
| Дополнительные ключевые слова (поле Keywords Plus) | 2806 | 78,84 | | |
| DOI | 1780 | 50,01 | 62 247 | 82,7 |
| Год публикации | 0 | 0 | 0 | 0 |
| Количество цитирований работы | 0 | 0 | 0 | 0 |
| Количество цитируемых статей | 0 | 0 | | |
| Цитируемые источники | 235 | 6,60 | | |
| Число страниц | 0 | 0 | 0 | 0 |
| Информация о финансировании | 2951 | 82,9 | 73 447 | 97,6 |

*Окончание табл. 3*

| Пропущенные элементы метаданных | WoS | | eLibrary | |
|---|---|---|---|---|
| | число | доля, % | число | доля, % |
| Автор(ы) | | | | |
| Автор (фамилия и имя) – английский язык | 0 | 0 | 35 | 0,05 |
| Автор (фамилия и имя) – русский язык | | | 1 539 | 2,05 |
| Название аффилиации автора на русском | - | - | 4609 | 6,1 |
| Название аффилиации автора на английском | 2 | 0,06 | 37 891 | 50,4 |
| Журнал | | | | |
| Журнал (название) | 0 | 0 | 0 | 0 |
| Журнал (ID) | | | 0 | 0 |
| Информация об издателе | 0 | 0 | 1223 | 1,6 |

*Примечание*. Знак "-" означает, что данное поле в базе не представлено. В ключевых словах на английском языке для WoS указаны данные из поля "DE". Более темным цветом выделены ячейки с более высокими долями пропущенных данных.

Предобработка данных для массива WoS проводилась с помощью программы WoS2Pajek [21], используемой для предобработки сетевых данных и построения сетевых файлов. С помощью этой программы была произведена чистка исходного файла с библиографическими описаниями – автоматическая идентификация и удаление дублей публикаций и лишних символов в файле. Для формирования массива ключевых слов программа берет информацию из полей "DE" – Keywords, "ID" – Keywords Plus, а также из названий статей и аннотаций – полей "TI" – Title и "AB" – Abstract, что решает проблему неполного покрытия некоторых из этих полей (обозначенную программой Biblioshiny). Программа проводит нормализацию и приводит к единому виду ключевые слова, используя словари для английского языка.

Поскольку программа WoS2Pajek ориентирована на использование информации о цитировании работ (кратких описаний цитируемых публикаций в поле “CR”), фокус делается на обработке библиографических описаний. Работы, указанные в поле “CR”, записаны в формате: AU + ’, ’ + PY + ’, ’ + SO[:20] + ’, V’ + VL + ’, P’ + BP (автор, год публикации, до 20 символов источника публикации / журнала, выпуск, начальная страница), например: TOSHCHENKO ZT, 2000, SOTSIOL ISSLED, V23, P123. Изначально такой подход использовался для повышения точности данных при внесении информации по единому формату. Но так как по факту одна работа может иметь отличающиеся наименования, для повышения точности программа WoS2Pajek использует короткие имена, записываемые в формате: LastNm[:8] + ’ ’ + FirstNm[0] + ’(’ + PY + ’)’ + VL + ’:’ + BP (8 символов фамилии, первая буква имени, год публикации, выпуск журнала, начальная страница), например: TOSHCHEN_Z(2000)23:123. Та же самая процедура создания коротких имен осуществляется и для работ, имеющих полные библиографические описания («хитов»). Для решения проблемы дизамбигуации имена авторов записываются по форме: LastNm[:8] + ’ ’ + FirstNm[0] (8 символов фамилии, первая буква имени), например: TOSHCHEN_Z. Безусловно, при таком подходе могут возникать проблемы «склейки» имен авторов, однако эти проблемы разрешаются путем проверки результатов, получаемых для наиболее важных единиц анализа[1]. Чистка данных, как правило, осуществляется итеративно – при нахождении проблем правки либо вносятся в исходный файл, который снова проходит через программу WoS2Pajek, либо устраняются алгоритмическим образом в программе Pajek.

Предварительный исследовательский анализ массива данных eLibrary показал, что некоторые важные для анализа параметры

[1] Методология следует так называемому статистическому подходу, согласно которому даже при некоторой неконсистентности в данных общие тренды и важные единицы анализа могут проявиться при анализе.

метаданных имеют отсутствующие значения (0 или “none”): из 37 302 уникальных авторов в начальном массиве только у 19 739 авторов имелись РИНЦ ID (хотя наличие именно этой информации рассматривалось как преимущество базы). В ходе предобработки данных eLibrary с целью дизамбигуации имен авторов собранная база трансформировалась: вначале была проведена нормализация имен авторов и их аффилиаций, а затем созданы универсальные ID для авторов в формате: eLibrary_ID + FirstNm[:2] + LastNm[:8] + Affiliation_ID (ID автора в РИНЦ, инициалы, 8 символов фамилии, ID организации автора), например: 1382_ZHT_Toschenk_5350 (подробнее см.: [20]). В результате предобработки количество уникальных названий аффилиаций сократилось на 39,5% за счет нормализации и приведения к единому виду описаний аффилиаций и создания универсальных описаний для аффилиаций с единым ID; количество уникальных ID организаций сократилось на 1,5% – за счет удаления некорректно заполненных ID, а количество уникальных авторов увеличилось на 95% – до 37 790 – за счет идентификации авторов, не имеющих РИНЦ ID [20].

**Построение сетевых файлов.** Используемый авторами подход к работе с данными WoS CC подразумевает использование программы WoS2Pajek [21] для трансформации массива в коллекцию связанных сетей, в частности (используемых для анализа) двумодальных сетей «Работа – Автор» (“Work – Author”) **WA**, «Работа – Ключевое слово» (“Work – Keyword”) **WK**, «Работа – Журнал» (“Work – Journal”) **WJ** (где в первом наборе указаны все публикации, во втором – авторы, ключевые слова или журналы, а далее фиксируются связи между ними). Также создается файл с информацией о годах публикаций работ (Year.clu), на основании которого сети можно разделить на периоды для изучения в динамике и файл с разделением работ – на источники с полным библиографическим описанием («хиты») и только цитируемую литературу (DC.clu).

По этой же логике после формирования массива в eLibrary из соответствующих полей с помощью специально написанного на Python программного кода были выгружены данные для построения двумодальных сетей «Работа – Автор» **WA**, «Работа – Ключевое слово» **WK** и «Работа – Журнал» (“Work – Journal”) **WJ**. Файлы сохранены в формате .net и доступны для дальнейшего анализа в программе Pajek. Был сформирован файл с информацией о годах публикаций отобранных работ[1].

Хотя напрямую сравнить два использованных подхода и инструмента для предобработки и построения сетевых файлов затруднительно, можно сделать выводы о времени, необходимом для работы в каждом случае. Ввиду автоматизированности процесса процедура предобработки данных и построения сетей с помощью программы WoS2Pajek занимает считанные минуты (построение сетевых файлов из начального подмассива заняло 2 мин. 34 сек.). Использование разработанного в рамках проекта программного кода для предобработки и подготовки сетевых файлов из массива eLibrary, зарегистрированного в виде ЭВМ, на данный момент требует гораздо больше времени ввиду необходимости ручной проверки данных на некоторых этапах.

## *Сравнительный анализ массивов данных*

**Полнота библиографических метаданных.** Безусловно, не все библиографические описания в базах WoS и eLibrary содержат все возможные метаданные, обозначенные в табл. 2. Вместе с тем полнота представления метаданных библиографических описаний оказывает значительное влияние на качество анализа, поэтому важно оценить объем пропущенных данных в массивах. Табл. 3

[1] Файлы создавались для проводимого анализа, но могут быть построены и другие двумодальные сети и дополнительные файлы с атрибутами узлов (количество страниц, цитирование), которые можно использовать для решения разных исследовательских вопросов.

представляет данные для оценки полноты библиометрических описаний публикаций в двух рассматриваемых массивах – количество и долю пропущенных значений по отобранным метаданным[1] (уже после проведенной предобработки данных). Для удобства метаданные сгруппированы по типам библиометрических единиц анализа, к которым они относятся (как в табл. 2). Обратим внимание, что данные подсчитаны для публикаций: отсутствующие значения по именам авторов показывают количество статей, в которых не имеется хотя бы одного имени автора на русском и английском языках; данные в названиях аффилиаций подсчитывают число случаев, когда при наличии авторов хотя бы у одного из них отсутствует название аффилиации[2].

Согласно оценке программы Biblioshiny, доли пропущенных значений от 20 до 50% говорят о слабой, а более 50% – о критической представленности данных в поле библиографического описания и не рекомендуются программой для использования в анализе. Как видим, для массива WoS это (от наибольшей доли пропущенных значений к наименьшей) информация о финансировании, дополнительным ключевым словам, DOI (критично), ключевым словам и аннотации публикации на английском языке (слабо). Для массива eLibrary это информация о финансировании, названию на английском языке, DOI, названию аффилиации автора на русском (критично), аннотации на русском и английском, ключевым словам на английском языке (слабо). Ключевые слова на русском языке отсутствуют в 15% публикаций, однако этот показатель считается «проходным».

**Основная информация по массивам.** В табл. 4 собрана основная информация по числу различных единиц анализа в рассматриваемых массивах.

---

[1] За основу структуры взята таблица, формируемая в программе Biblioshiny при загрузке массива данных.

[2] Если автора нет (none и по столбцу с английской фамилией, и по столбцу с русской), наличие аффилиации не проверяется.

*Таблица 4*

ЧИСЛО ЕДИНИЦ АНАЛИЗА В МАССИВАХ WoS И eLibrary

| Единица анализа / Количество | eLibrary | WoS | Доля WoS к РИНЦ, % |
|---|---|---|---|
| Публикации | 75 232 | **3559** | 4,7 |
| Журналы | 3910 | **109** | 2,8 |
| Авторы | 37 790 | 3238 | 8,6 |
| Ключевые слова на английском | 91 109 | 6750 | 7,4 |

*Примечание.* Расчет по WoS приведен по программе WoS2Pajek.

Обратим внимание, что для сравнения в eLibrary взяты данные по числу ключевых слов на английском языке (аналогичный показатель на русском языке составляет 100 594); ключевые слова были взяты в формате, приведенном авторами, и не подвергались обработке (чем можно объяснить их большое количество – из-за наличия множества уникальных слов). По массиву WoS подсчет приведен по данным программы WoS2Pajek[1]. Для информации подсчитано соотношение единиц из массива WoS к массиву eLibrary, показывающее значительное превосходство данных eLibrary по объему.

Динамика количества **публикаций** в обеих базах за рассматриваемый период показана на рис. 3.

Распределение абсолютного числа публикаций (рис. 3) в eLibrary показывает плавный рост и достижение максимума в 2016 г. и следующее за ним снижение. Количество российских социологических публикаций в WoS достигает максимума в 2019 г., однако далее снижается незначительно. Чтобы увидеть

[1] В связи с имплементированными алгоритмами предобработки Biblioshiny выделяет несколько иное число авторов (3554) и ключевых слов (12 215). Данные по ключевым словам, полученные WoS2Pajek, следует рассматривать как более валидные – ввиду более точного подсчета (для Biblioshiny число рассчитано как сумма по полям ID и DE, нет возможности учета пересечений) и заложенных в программу алгоритмов нормализации ключевых слов.

общие тренды, данные были подсчитаны кумулятивно и затем нормированы в диапазоне от 0 до 1 (значение за каждый год разделено на сумму публикаций) (рис. 4).

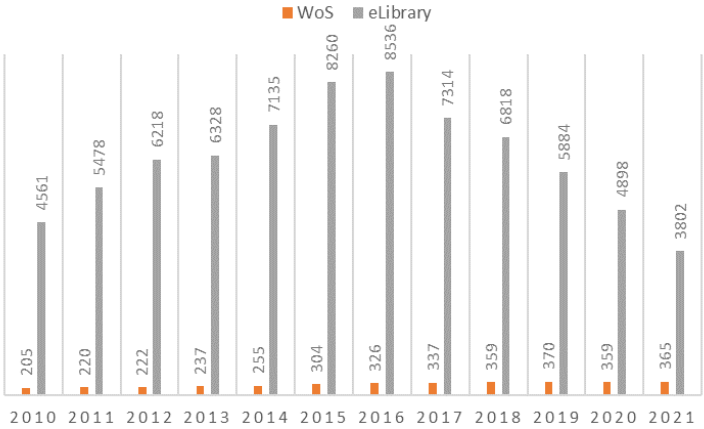



***Рис 3. Динамика количества публикаций в базах WoS и eLibrary: абсолютное число публикаций***

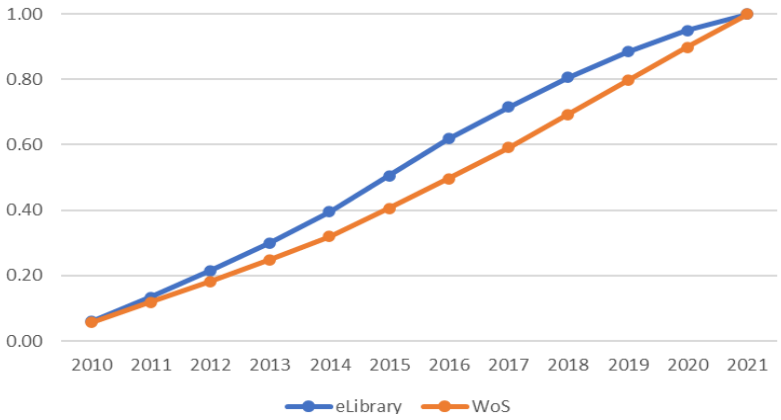



***Рис 4. Динамика количества публикаций в базах WoS и eLibrary: число публикаций, нормированное на кумулятивной шкале***

В такой репрезентации лучше видно, что относительные доли числа публикаций в eLibrary с 2014–2015 гг. были выше, чем в WoS. Однако если средний годовой прирост публикаций[1] в eLibrary на 2021 г. составляет -0,6%, аналогичный показатель для WoS составляет 5,5%, что говорит о более динамичном увеличении числа публикаций в этой базе.

## Анализ пересечений между базами на основе статистики

Интересной находкой исследования стало то, что для каждой публикации в массиве eLibrary содержится информация о том, в каких наукометрических базах (РИНЦ, RSCI, WoS, Scopus, ВАК) она проиндексирована, что дает возможность посмотреть на распределение и пересечение публикаций в разных базах на основе информации от eLibrary. Эта информация напрямую не относится к предмету данной статьи, однако приведена в Приложении. Проведенный анализ подтвердил, что не все статьи, вошедшие в массив eLibrary (75 232 публикации), входят в базу РИНЦ. Также анализ показывает, что база РИНЦ в значительной степени пересекается с другими, меньшими по размеру базами библиографических данных, включая базу WoS. Это дает основания к проверке пересечений между двумя базами, проводимой ниже. В данном разделе сравниваются массивы данных WoS и eLibrary по публикациям и авторам, включенным в два анализируемых массива данных.

*Сопоставление публикаций.* Для сопоставления публикаций, входящих в базы WoS (3559) и eLibrary (75 232), было использовано несколько подходов: мы последовательно сопоставляли мас-

[1] Рассчитанный путем деления разницы между количеством публикаций за каждую пару лет ($n$, $n + 1$) на значение первого года ($n$), и последующий расчет среднего всех полученных значений за каждый год в анализируемый период времени (2010–2021 гг.).

сивы данных по: 1) DOI; 2) названию публикаций на английском языке; 3) сгенерированной комбинации из последовательности авторов и года написания статьи.

Поиск совпадающих DOI статей позволил сопоставить 655 публикаций. Ситуацию осложняло отсутствие DOI у 50% статей в WoS и у 83% в eLibrary. Далее сопоставление статей было продолжено путем сравнения их названий на английском для оставшихся неидентифицированными публикаций (обратим внимание на высокую долю пропущенных значений). Названия были предварительно предобработаны (убраны знаки препинания и цифры, все символы приведены к нижнему регистру, убраны стоп-слова: of, in, at, is; артикли) и по точным совпадениям удалось получить еще 32 совпавшие статьи. Далее мы приступили к поиску совпадений по комбинации авторов и года написания статьи для оставшихся неидентифицированными массивов данных. Например, если Михаил Соколов (*SOKOLOV MM*) и Кирилл Титаев (*TITAEV KD*) написали статью в 2014 г., их статье была присвоена строка "S*OKOLOV MM;TITAEV KD_2014*". Такие «простые ID» мы сгенерировали для всех статей в WoS и eLibrary и искали совпадения по ним. Отметим, что для повышения точности оценки и избежания ложноположительных совпадений поиск совпадений проводился только для статей, чье «простоеID» встречалось в базе данных только один раз. В противном случае совпадение будет неточным: один и тот же социолог или группа исследователей могут написать несколько статей в один и тот же год, и мы не сможем точно сопоставить, например, одну публикацию Ж.Т. Тощенко в 2012 г. в WoS с какой-то из шести публикаций Тощенко в 2012 г. в eLibrary. По результатам этого поиска нашлось еще 358 статей.

Несмотря на то, что комбинация выбранных способов поиска идентичных статей неидеальна, она позволяет получить примерную оценку совпадения двух баз данных без разработки специфических технологических решений. В теории они могли бы включать поиск совпадающих статей по разным вариациям

автоматического перевода названия статьи с русского на английский, поиска совпадающих статей по комбинации авторов с поиском возможных расхождений в один-два символа в фамилиях и пр., однако эта разработка может стать темой отдельного исследовательского проекта. Итоговое значение совпавших статей в базе данных elibrary и WoS составляет 1013 статей – 28,5% от всех публикаций в базе данных WoS или 1,4% от всех публикаций в базе данных elibrary.

*Сопоставление авторов.* Для сравнения авторов, присутствующих в базах WoS и eLibrary, мы выбрали подход, в котором искали совпадения по фамилиям и инициалам авторов; таким образом, ограничением для следующих оценок стало предположение о том, что одна комбинация фамилии и инициалов принадлежит одному автору. Так, хотя в предварительно обработанных данных eLibrary были созданы новые универсальные ID, позволяющие идентифицировать разных авторов (основываясь на ID РИНЦ, инициалах, фамилии и ID аффилиациии), они бы не совпадали с потенциальными ID, которые можно было бы сконструировать на основе данных WoS. WoS содержит информацию о ResearcherID и ORCID-ID исследователей, но эти поля часто не заполнены. В нашей базе в 54,5% статей нет информации о ResearcherID ни для одного из авторов статьи; для ORCID-ID аналогичная оценка составляет 61,4%.

По этим причинам в базе данных статей российских социологов из eLibrary мы создали столбец с перечислением всех авторов, аналогичный по формату записям в базе WoS, где авторы записываются в столбце “AU” следующим образом: “*KOLESNIK NV;SHOPULATOV AN;SINYUTIN MV*”. Отметим, что предобработка фамилий (например, приведение отдельных написаний фамилии к наиболее популярному виду) не производилась, однако такие процедуры можно было бы провести, тем самым повысив точность оценки. Число имен авторов (табл. 5), сформированных таким образом, не полностью совпадает с числом имен авторов

после дизамбигуации для двух массивов (табл. 4), однако такой подход позволяет дать некоторую количественную оценку имеющимся пересечениям авторов в базе и показать, какие авторы присутствуют только в одной или в обеих базах. По полученным результатам (табл. 5), число авторов в обеих базах составило 1180 авторов, что составляет 33% от всех авторов в массиве WoS, но только 3,3% от всех авторов.

*Таблица 5*

ПОКАЗАТЕЛИ СОПОСТАВЛЕНИЯ АВТОРОВ
В БАЗАХ WoS И eLibrary

| Показатель | Значение |
|---|---|
| Число авторов в eLibrary | 35 462 |
| Число авторов в WoS | 3554 |
| Число авторов в обеих базах | 1180 |
| Доля авторов в обеих базах относительно числа уникальных авторов в данных WoS | 33% |
| Доля авторов в обеих базах относительно числа уникальных авторов в данных eLibrary | 3,3% |

*Примечание*. Число авторов в eLibrary подсчитано так же, как в WoS (первые 8 букв имени и инициалы после разделителя), поэтому число авторов не совпадает с числом авторов из табл. 4, к которым был применен подход по дизамбигуации имен.

## *Анализ пересечений между базами на основе содержательных результатов*

В данном разделе проводится опосредованное, непрямое сравнение того, насколько похожими являются массивы данных WoS и eLibrary с точки зрения получаемых результатов при анализе массива и производных сетей.

**Работы и авторы.** Входящая центральность в двумодальной сети **WA** показывает количество работ у авторов. Распределение этого показателя для двух массивов приведено на рис. 5.

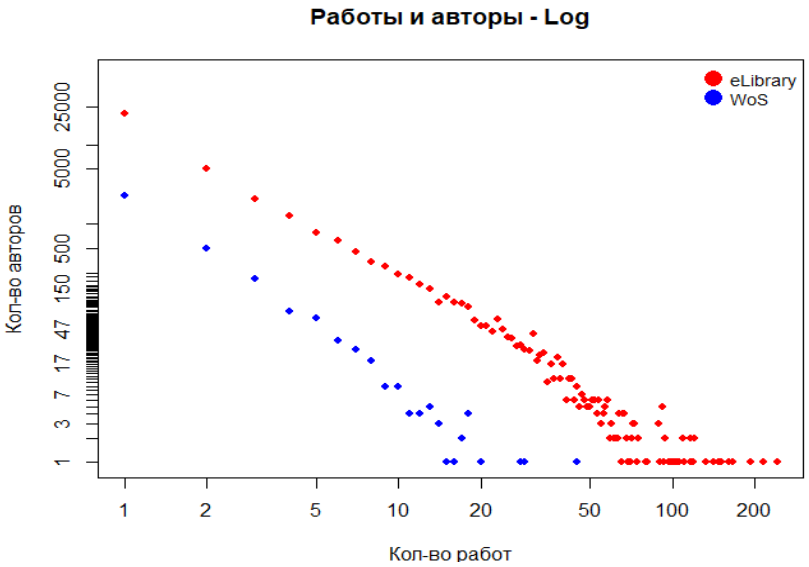



***Рис. 5. Распределение количества работ по авторам в двух массивах данных (логарифмическая шкала)***

Массивы данных значительно различаются по размеру – число публикаций в массиве eLibrary примерно в 20 раз больше числа публикаций в WoS – поэтому различия наблюдаются и в числе авторов. Вместе с тем распределения на рис. 5 похожи по тренду и могут следовать степенному закону, или закону Лотки, описывающему распределение продуктивности ученых[1]. Тогда как 66% авторов в массиве eLibrary и 64% в массиве WoS имеют только одну публикацию, еще 13 и 14% – две, а по 6% – три, некоторые авторы в базах являются суперпродуктивными, имея 241, 2015 и 192 публикации в базе eLibrary и 45, 29 и 28 публикаций в базе WoS.

[1] Согласно закону Лотки, число авторов, опубликовавших в течение определенного периода *n* статей, обратно пропорционально квадрату *n*. Этот закон можно проверить математической функцией.

Наиболее продуктивные авторы с наибольшим количеством работ по двум массивам приведены в табл. 6. Четверо из выделенных топ-20 авторов присутствуют в обеих базах, однако авторы из массива eLibrary, имеющие наибольшее количество публикаций, в базе WoS имеют 1,1 и 5 публикаций.

*Таблица 6*

АВТОРЫ С НАИБОЛЬШИМ КОЛИЧЕСТВОМ ПУБЛИКАЦИЙ, ПО ДВУМ МАССИВАМ WOS И ELIBRARY

| Ранг | Массив WoS | | Массив eLibrary | |
|---|---|---|---|---|
| | ID автора | кол-во публикаций | ID автора | кол-во публикаций |
| 1 | **TROTSUK_I** | 45 | 429210_SI_Samygin_14461 | 241 |
| 2 | PUZANOVA_Z | 29 | 74486_SG_Maksimov_258_7082 | 215 |
| 3 | KRAVCHEN_S | 28 | 767943_TK_Rostovsk_924_1432_1488_4812_5350_13701 | 192 |
| 4 | **TOSHCHEN_Z** | 20 | **137655_GE_Zborovsk_290_1255_7366_14141** | 166 |
| 5 | NARBUT_N | 18 | 75266_NV_Dulina_306_1000 | 160 |
| 6 | **ZBOROVSK_G** | 18 | 145046_OE_Nojanzin_258_7082 | 150 |
| 7 | SOROKIN_P | 18 | 72232_JUG_Volkov_322_1432_3455_14829 | 147 |
| 8 | SOKOLOV_M | 18 | 129623_VA_Il'in_815 | 142 |
| 9 | **GORSHKOV_M** | 17 | 287431_AV_Verescha_322_1432_14461 | 133 |
| 10 | YANITSKI_O | 17 | 504328_MV_Morev_815 | 120 |
| 11 | ROMANOVS_N | 16 | **251886_IV_Trotsuk_421_425** | 120 |
| 12 | TESLYA_A | 15 | 495445_DA_Omel'che_258 | 119 |
| 13 | KOZYREVA_P | 14 | **1382_ZHT_Toschenk_5_5350** | 117 |

*Окончание табл. 6*

| Ранг | Массив WoS | | Массив eLibrary | |
|---|---|---|---|---|
| | ID автора | кол-во публикаций | ID автора | кол-во публикаций |
| 14 | SMIRNOV_A | 14 | 73979_HV_Dzutsev_1432_4812 | 116 |
| 15 | OBRAZTSO_I | 14 | 442046_JUV_Stavropo_259_808 | 116 |
| 16 | TIKHONOV_N | 13 | 674856_NH_Gafiatul_322_761 | 111 |
| 17 | GASPARIS_A | 13 | 265785_VP_Babintse_340_1279_6227 | 110 |
| 18 | RYBAKOVS_L | 13 | 331427_SA_Il'inyh_1068 | 109 |
| 19 | LAPIN_N | 13 | **72610_MK_Gorshkov_1432_14554** | 109 |
| 20 | LARINA_T | 13 | 259120_PA_Ambarova_290 | 106 |

*Примечание*. Жирным шрифтом выделены фамилии авторов, встречающихся в топ-20 по обоим массивам; «ID автора» приведены согласно тому, как авторы указаны в соответствующем массиве данных.

Показатель исходящей центральности в сети **WA** показывает количество авторов в работах (табл. 7).

*Таблица 7*

КОЛИЧЕСТВО АВТОРОВ В ПУБЛИКАЦИЯХВ ДВУХ МАССИВАХ

| WoS | | | eLibrary | | |
|---|---|---|---|---|---|
| число авторов | *N* | доля от всех авторов, % | число авторов | *N* | доля от всех авторов, % |
| 1 | 2217 | 62,29 | 1 | 49 973 | 66,43 |
| 2 | 844 | 23,71 | 2 | 18 473 | 24,55 |
| 3 | 327 | 9,19 | 3 | 5044 | 6,7 |
| 4 | 120 | 3,37 | 4 | 1144 | 1,52 |

*Окончание табл. 7*

| WoS | | | eLibrary | | |
|---|---|---|---|---|---|
| число авторов | *N* | доля от всех авторов, % | число авторов | *N* | доля от всех авторов, % |
| 5 | 34 | 0,96 | 5 | 361 | 0,48 |
| 6 | 5 | 0,14 | 6 | 113 | 0,15 |
| 7 | 6 | 0,17 | 7 | 63 | 0,08 |
| 8 | 2 | 0,06 | 8 | 61 | 0,08 |
| 9 | 1 | 0,03 | | | |
| 12 | 1 | 0,03 | | | |
| 14 | 1 | 0,03 | | | |
| 15 | 1 | 0,03 | | | |

Максимальное число авторов в массиве WoS составляет 15; для массива eLibrary при сборе данных было установлено ограничение в 8 авторов. Как видно, доли публикаций статей с единственным автором в двух массивах являются практически идентичными – 62% в WoS и 66% в eLibrary. Это подтверждает обозначенную гипотезу о распространенности практики публикаций с единственным автором как части публикационной культуры в области социальных наук. Следующий самый часто встречающийся в публикациях формат – подготовка публикаций парами авторов – встречается в 24 и 25% статей в WoS и eLibrary соответственно; за ним следуют публикации, сделанные тремя (9% для WoS и 7% для eLibrary) и четырьмя (3% и 1.5%) авторами. Статьи с относительно большим количеством авторов встречаются в массивах в единичном виде.

**Коллаборации авторов.** На основе сети WA путем ее перемножения может быть построена базовая ненормализованная сеть соавторства Co, где сила связей рассчитывается исходя из количества публикаций, написанных авторами совместно, а петля обозначает общее количество работ у авторов, написанных в соавторстве и самостоятельно [4]. Доли авторов, не имеющих хотя бы одного

соавтора, для массивов eLibrary и WoS составляют 35,8 и 27,8% соответственно. Распределение по числу соавторов у авторов показывает, что большинство из них имеют одного (31% в eLibrary и 27% в WoS), двух (14% и 18,5%) или трех (6,5% и 12,5%) соавторов (рис. 6).

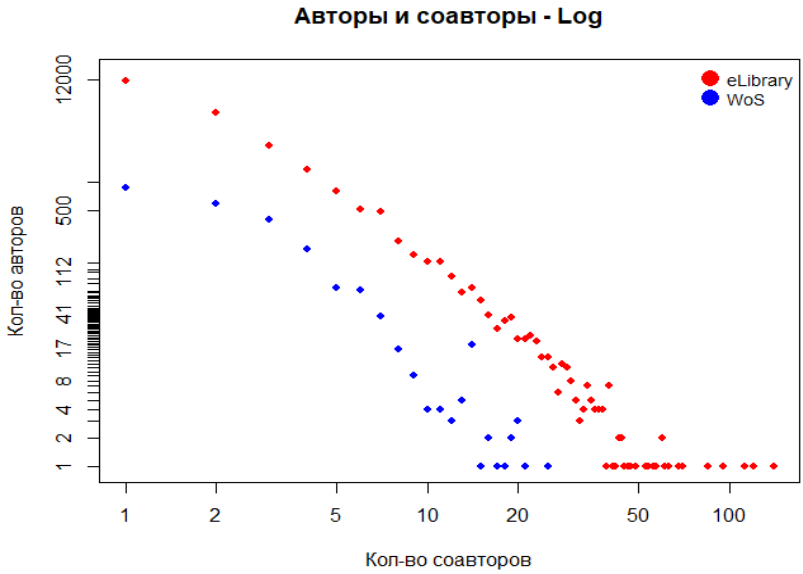



***Рис. 6. Распределение количества соавторов по авторам в двух массивах данных (логарифмическая шкала)***

Однако также выделяются авторы со значительным количеством соавторов, например: в массиве WoS – Н.Е. Покровский (25 соавторов), В.В. Щербина (21) и Ж.Т. Тощенко, Н.В. Романовский и А.Б. Гофман (20), в массиве eLibrary доля авторов с числом соавторов более 20 составляет 0,57%, или 212 авторов, среди которых лидирует С.И. Самыгин со 140 соавторами.

**Работы и журналы.** В табл. 8 приведены топ-25 журналов по количеству публикаций в двух массивах.

*Таблица 8*

ТОП-15 ЖУРНАЛОВ В ДВУХ МАССИВАХ ДАННЫХ ПО ЧИСЛУ ПУБЛИКАЦИЙ

| Ранг | WoS | | | eLibrary | | |
|---|---|---|---|---|---|---|
| | журнал | *N* | доля от всех публикаций, % | журнал | *N* | доля от всех публикаций, % |
| 1 | SOTSIOL ISSLED+ | 1923 | 54,0 | Социологические исследования | 1905 | 2,5 |
| 2 | RUDN J SOCIOL | 546 | 15,3 | Экономика и социум | 1759 | 2,3 |
| 3 | SOCIOL OBOZR | 309 | 8,7 | Теория и практика общественного развития | 1078 | 1,4 |
| 4 | J ECON SOCIOL | 245 | 6,9 | Гуманитарные, социально-экономические и общественные науки | 911 | 1,2 |
| 5 | SOCIOL NAUK TEHNOL | 167 | 4,7 | Социально-гуманитарные знания | 863 | 1,1 |
| 6 | CHANG SOC PERSONAL | 51 | 1,4 | Социология | 737 | 1,0 |
| 7 | INT J SOCIOL SOC POL | 37 | 1,0 | Мониторинг общественного мнения: экономические и социальные перемены | 731 | 1,0 |
| 8 | COMP SOCIOL | 23 | 0,6 | Социология в современном мире: наука, образование, творчество | 670 | 0,9 |

*Окончание табл. 8*

| Ранг | WoS | | | eLibrary | | |
|---|---|---|---|---|---|---|
| | журнал | *N* | доля от всех публикаций, % | журнал | *N* | доля от всех публикаций, % |
| 9 | INT J INTERCULT REL | 22 | 0,6 | Социальная политика и социология | 630 | 0,8 |
| 10 | SOC INDIC RES | 20 | 0,6 | Социальные и гуманитарные науки. Отечественная и зарубежная литература. Серия 11: Социология | 598 | 0,8 |
| 11 | FILOS-SOCIOL | 18 | 0,5 | Гуманитарий Юга России | 582 | 0,8 |
| 12 | SPORT SOC | 8 | 0,2 | Власть | 528 | 0,7 |
| 13 | POETICS | 8 | 0,2 | Журнал социологии и социальной антропологии | 523 | 0,7 |
| 14 | CORVINUS J SOCIOL PO | 5 | 0,1 | Общество: социология, психология, педагогика | 494 | 0,7 |
| 15 | CURR SOCIOL | 5 | 0,1 | Известия Саратовского университета. Новая серия. Серия: Социология. Политология | 482 | 0,6 |

Лидером в обеих базах выступает журнал «Социологические исследования» – абсолютное количество публикаций в нем в WoS и в eLibrary примерно одинаково (1923 и 1905 соответственно). Если же посмотреть на вклад журнала в общее количество публикаций, то значение этого источника для базы WoS становится еще важнее – публикации в нем составляют 54% от всего массива данных. В eLibrary вклад «Социса» растворяется в связи с большим количеством журналов; несколько других журналов с большим вкладом идут с довольно небольшим отставанием. На основе распределения журналов из массива WoS видно, что вклад российских авторов в эту площадку (зарубежную «витрину») в основном делается через публикации в российских журналах, индексируемых в WoS (первые пять российских журналов составляют 90% публикаций). Однако при оценке вклада журналов нужно учитывать эффект периодичности (частоты выхода публикаций) и количества публикаций в каждом номере (которое обычно велико для недобросовестных журналов, которые могут присутствовать в базах).

**Работы и ключевые слова.** Показатель исходящей центральности в сети WK показывает количество ключевых слов в работе. Для работ из массива WoS этот показатель варьируется от 1 до 40, а из массива eLibrary – от 1 до 51 (при этом в 36,7% случаев значения пропущены). Показатель входящей центральности в сети WK показывает частоту использования различных ключевых слов в работах. Как показывает распределение этих значений для двух массивов (рис. 7), 77% ключевых слов в массиве eLibrary и 50,5% в WoS использованы только один раз, еще 10 и 14% соответственно – два раза, 3,6 и 7% – три раза и т.д.

В табл. 9 приведены топ-20 слов, наиболее часто используемых в обоих массивах. Повторяющиеся слова из двух массивов выделены цветом.

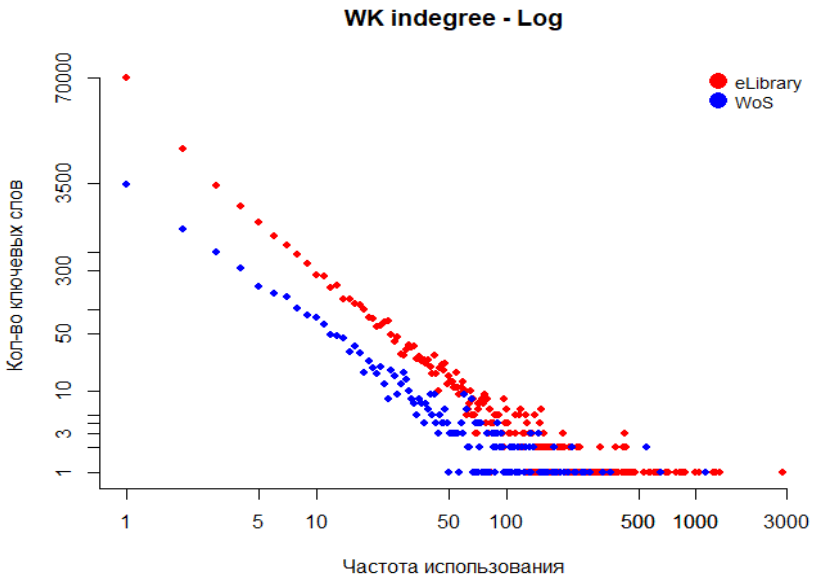



***Рис. 7. Частота использования ключевых слов в работах в двух массивах данных (логарифмическая шкала)***

*Таблица 9*

ТОП-20 КЛЮЧЕВЫХ СЛОВ ДЛЯ ОБОИХ МАССИВОВ

| Ранг | eLibrary | | WoS | |
|---|---|---|---|---|
| | слово | значение | слово | значение |
| 1 | youth | 2849 | social | 1119 |
| 2 | **society** | 1343 | russian | 649 |
| 3 | family | 1269 | **sociology** | 547 |
| 4 | values | 1225 | **russia** | 547 |
| 5 | **culture** | 1035 | **society** | 353 |
| 6 | **education** | 990 | sociological | 322 |
| 7 | globalization | 876 | analysis | 279 |
| 8 | students | 855 | theory | 261 |
| 9 | migration | 839 | study | 253 |
| 10 | socialization | 820 | **education** | 233 |

*Окончание табл. 9*

| Ранг | eLibrary | | WoS | |
|---|---|---|---|---|
| | слово | значение | слово | значение |
| 11 | identity | 791 | **state** | 232 |
| 12 | civil society | 716 | political | 231 |
| 13 | modernization | 696 | research | 230 |
| 14 | communication | 631 | development | 223 |
| 15 | **state** | 611 | science | 223 |
| 16 | management | 601 | life | 212 |
| 17 | **russia** | 580 | **cultural** | 202 |
| 18 | region | 579 | value | 193 |
| 19 | internet | 571 | practice | 185 |
| 20 | **sociology** | 534 | public | 182 |

*Примечание*. Полужирным шрифтом выделены слова, полностью повторяющиеся в двух списках, подчеркнуты слова, имеющие общую часть.

## Выводы и обсуждение

Проведенный литературный обзор исследований по сравнению баз данных научных публикаций подтверждает, что даже при наличии альтернатив WoS является одним из самых популярных источников информации для наукометрических исследований. Безусловным плюсом работы с этой базой является наличие инструментов для выгрузки, предобработки и статистического и сетевого анализа публикаций, но, в случае с данными российских авторов, минусом – ограниченная представленность публикаций. В eLibrary, напротив, отечественные публикации представлены максимально полно (и не ограничиваются только публикациями в научных журналах и главами в монографиях); проблема заключается в отсутствии широко доступных сервисов по обработке и анализу данных для этой базы. В этой ситуации у исследователя, нацеленного на изучение современного состояния развития российской науки, возникает ряд вопросов: можно ли взять только одну базу в качестве источника информации, или необходимо комбинировать данные

из нескольких баз? в случае использования одной базы, насколько валидными будут полученные результаты? если данные должны комбинироваться, то как именно это нужно делать?

В нашем исследовании проводится сравнение двух баз через описательный анализ их возможностей по работе с данными и сопоставление двух массивов по одной и той же предметной области – социологии, – что является распространенной практикой в дизайне аналогичных исследований. Полученные массивы данных сравниваются по своей структуре, размеру, полноте метаданных, а также посредством анализа производных базовых сетей. Это важно не только с наукометрической точки зрения, но и с позиции изучения ориентаций ученых на международные и локальные научные сообщества, если думать о двух площадках как о двух возможных направлениях позиционирования ученых.

Несмотря на похожий набор метаданных, базы WoS и eLibrary имеют некоторые различия. Одним из преимуществ WoS является наличие списков литературы, что важно, если предполагается использовать анализ цитирований. В eLibrary авторы и организации имеют ID, однако по факту в большом количестве случаев эта информация отсутствует, что приводит к необходимости предварительной обработки массивов данных. В отличие от данных WoS, работа с которыми может осуществляться в нескольких программах, работа с данными eLibrary как по предобработке, так и по построению сетевых файлов для дальнейшего библиометрического анализа является гораздо более трудозатратой.

Рассмотренный массив eLibrary является гораздо более крупным, т.к. аккумулирует информацию из различных российских журналов и некоторых иностранных баз данных. WoS Core Collection имеет строгие критерии индексации журналов (что сокращает возможное число национальных журналов, которые могут быть представлены на этой площадке) и является только одной из баз, информация из которой должна включаться в eLibrary. Несмотря на то, что по логике формирования базы данных eLibrary рассматриваемые нами для примера массивы данных должны

в значительной степени пересекаться (массив WoS должен быть включен в массив eLibrary), пересечение между рассматриваемыми массивами является далеко не полным (около 30% массива WoS входят в массив eLibrary). Это может объясняться как несовершенством реализованной процедуры поиска идентичных публикаций и авторов, так и тем, что в массивах содержатся уникальные публикации и авторы. Эта часть анализа в настоящий момент может рассматриваться как экспериментальная и заслуживает дальнейшей проработки для уточнения пересечения массивов. Отметим, что наличие DOI у всех статей и ResearcherID/ORCID-ID у авторов могло бы существенно упростить эту задачу.

Динамика количества публикаций в двух массивах показывает, что база WoS прирастает более активно. Однако данные в обоих массивах распределяются похожим образом, что говорит о том, что они следуют похожим библиометрическим трендам и законам. Схожим образом в обоих массивах разделяются доли числа работ у авторов (две трети авторов с одной статей), авторов у работ (две трети работ наблюдаемых в социальных науках «авторов-одиночек», четверть работ, написанных в парах), соавторов у авторов (около трети авторов без соавторов и столько же – с одним соавтором); доля статей с одним ключевым словом в массиве eLibrary выше, чем в WoS (77% против 50%). Выделенные топ-единицы анализа при этом пересекаются только частично, что говорит о наличии своих особенностей в каждом массиве – самых продуктивных авторов для каждой площадки, наиболее используемых журналов и уникальных ключевых слов, характеризующих исследования. Более подробный анализ наблюдаемых пересечений и отличий может помочь ответить на различные содержательные вопросы о специфике исследований, ориентированных на разные аудитории (хотя возникает вопрос, насколько «ориентированными» на зарубежные исследовательские группы являются публикации в WoS, изначально вышедшие в российских журналах и внесенные в базу благодаря их индексации).

На основе проделанного сравнения и изучения логики формирования рассмотренных баз данных научных публикаций можно сделать некоторые выводы и рекомендации по поводу выбора базы данных для анализа публикаций российских авторов. Выяснилось, что множество статей в eLibrary не идентично множеству статей РИНЦ – в последнем индексируется меньше журналов и иных типов публикаций ввиду применения более строгих правил отбора, что делает данные в базе РИНЦ более надежными для анализа развития научной продукции. Обращение в качестве источника данных к базе RSCI, применяющей строгие критерии для отбора топовых российских журналов, сужает объем изучаемого числа публикаций, предоставляя доступ только к статьям, опубликованным в российских журналах. Обращение к WoS CC хотя и дает доступ к публикациям российских авторов в зарубежных журналах, включает публикации лишь из некоторых российских журналов, поэтому ограничивает анализируемый объем статей в наибольшей степени. Конечный выбор той или иной базы должен диктоваться исследовательской задачей: изучение представленности российских авторов в международном пространстве, очевидно, требует обращения к базе WoS (однако стоит рассматривать и выход за пределы ее ядра CC), а анализ сугубо российских публикаций может быть осуществлен путем анализа базы RSCI. Вместе с тем для изучения трендов развития российской науки в целом и практики ее воспроизводства стоит обратиться к базе РИНЦ или eLibrary. В идеальной ситуации анализ публикаций российских авторов должен осуществляться на основе нескольких баз данных, однако методологические вопросы выгрузки данных, поиска совпадений между базами и их объединения в единый массив пока являются открытыми и требуют дальнейшей разработки.

В качестве общей рекомендации нужно сказать, что исследователь, работающий в области библиометрического анализа, должен хорошо понимать структуру баз, с которыми он работает, чтобы получить нужную ему информацию на входе для дальней-

шего анализа, а не просто «искать где светлее» (например, в WoS, так как разработаны инструменты для анализа), а также ставить исследовательские вопросы с пониманием ограничений в покрытии баз данных. С точки зрения получения валидных результатов важной является также оценка полноты отдельных метаданных в библиографических описаниях.

Безусловно, важным вопросом является доступность баз данных. К сожалению, рассмотренная в рамках проведенного анализа база WoS с недавних пор недоступна российским исследователям, а доступ к API-сервису eLibrary отсутствует у большинства исследователей. На наш взгляд, наличие функциональных возможностей для сбора данных или открытого доступа к API-сервису eLibrary может стать важным условием для развития библиометрического анализа публикаций российских авторов и наукометрических исследований в российской практике в целом. Еще одним вариантом для получения данных может стать использование альтернативных открытых баз данных научных публикаций (например, базы OpenAlex, выбор которой уже приходит на замену традиционным базам WoS и Scopus в некоторых научно-образовательных организациях).

Дальнейшая работа над этой тематикой требует разработок в области методологии сбора, поиска совпадений и объединения массивов из различных источников. Полученные результаты в области социологии интересно сравнить с другими предметными областями. Помимо наукометрического интереса, анализ и сравнение публикационной активности исследователей на разных – международных и отечественных – площадках может помочь ответить и на многие содержательные вопросы, возникающие при изучении локальных научных сообществ, которые находятся за пределами рассмотрения данной статьи.

ЛИТЕРАТУРА

1. *Bar-Ilan J.* Informetrics at the beginning of the 21st century – A review // Journal of informetrics. 2008. Vol. 2. P. 1–52. DOI: 10.1016/j.joi.2007.11.001. EDN: MISIBR.

2. *Mingers J., Leydesdorff L.* A review of theory and practice in scientometrics // European journal of operational research. 2015. Vol. 246, № 1. P. 1–19. DOI: 10.1016/j.ejor.2015.04.002. EDN: UQPVRP.

3. *Rousseau R., Egghe L., Guns R.* Becoming metric-wise: A bibliometric guide for researchers / Ed. by W. Glänzel [et al.]. Cambridge, MA: Chandos Publishing, 2018. 850 p. ISBN: 0081024754, 9780081024751.

4. Understanding large temporal networks and spatial networks: Exploration, pattern searching, visualization and network evolution / V. Batagelj, P. Doreian, A. Ferligoj, N. Kejžar. Hoboken, NJ: WileyBlackwell, 2014. 464 p. ISBN: 978-1-118-91537-0.

5. *Моисеев С.П., Мальцева Д.В.* Отбор источников для систематического обзора литературы: сравнение экспертного и алгоритмического подходов // Социология: методология, методы, математическое моделирование (Социология: 4М). 2018. № 47. С. 7–43. EDN: MZXVXW.

6. *Булычева Е.Е., Мальцева Д.В.* Выделение актуальных тематик в социологии: взгляд сквозь призму анализа сети цитирований // Мониторинг общественного мнения: экономические и социальные перемены. 2020. № 6 (160). С. 113–140. DOI: 10.14515/monitoring.2020.6.971. EDN: UGIDGS.

7. *Harzing A.W., Alakangas S.* Google Scholar, Scopus and the Web of Science: A longitudinal and cross-disciplinary comparison // Scientometrics. 2016. Vol. 106. P. 787–804. DOI: 10.1007/s11192-015-1798-9. EDN: ZGNBBS.

8. The journal coverage of Web of Science, Scopus and Dimensions: A comparative analysis / V.K. Singh, P. Singh, M. Karmakar [et al.] // Scientometrics. 2021. Vol. 126. P. 5113–5142. DOI: 10.1007/s11192-021-03948-5. EDN: FLHAPG.

9. Google Scholar, Microsoft Academic, Scopus, Dimensions, Web of Science, and OpenCitations' COCI: a multidisciplinary comparison of coverage via citations / A. Martín-Martín, M. Thelwall, E. Orduna-Malea, E. Delgado López-Cózar // Scientometrics. 2021. Vol. 126, № 1. P. 871–906. DOI: 10.1007/s11192-020-03690-4. EDN: XNWTDQ.

10. *Harzing A.W*. Two new kids on the block: How do Crossref and Dimensions compare with Google Scholar, Microsoft Academic, Scopus and the Web of Science? // Scientometrics. 2019. Vol. 120, № 1. P. 341–349. DOI: 10.1007/s11192-019-03114-y. EDN: VKFCPM.

11. *Zhu J., Liu W.* A tale of two databases: The use of Web of Science and Scopus in academic papers // Scientometrics. 2020. Vol. 123, № 1. P. 321–335. DOI: 10.1007/s11192-020-03387-8. EDN: LZMVNM.

12. *Gusenbauer M.* Google Scholar to overshadow them all? Comparing the sizes of 12 academic search engines and bibliographic databases // Scientometrics. 2019. Vol. 118, № 1. P. 177–214. DOI: 10.1007/s11192-018-2958-5. EDN: ECWMGT.

13. *Moed H.F., Markusova V., Akoev M.* Trends in Russian research output indexed in Scopus and Web of Science // Scientometrics. 2018. Vol. 116. P. 1153–1180. DOI: 10.1007/s11192-018-2769-8. EDN: VBDLNY.

14. *Vera-Baceta M.A., Thelwall M., Kousha K.* Web of Science and Scopus language coverage // Scientometrics. 2019. Vol. 121, № 3. P. 1803–1813. DOI: 10.1007/s11192-019-03264-z. EDN: IHLJRA.

15. *Ruiz-Pérez R., López-Cózar E.D., Jiménez-Contreras E.* Spanish personal name variations in national and international biomedical databases: implications for information retrieval and bibliometric studies // Journal of the medical library association. 2002. Vol. 90, № 4. P. 411–430.

16. *Adriaanse L.S., Rensleigh C.* Web of Science, Scopus and Google Scholar: A content comprehensiveness comparison // The Electronic Library. 2013. Vol. 31, № 6. P. 727–744. DOI: 10.1108/EL-12-2011-0174.

17. *Еременко Г.О.* Сравнение уровня публикаций российских ученых в базах данных Web of Science, Scopus и RSCI: статья в открытом архиве // НЭБ. 28.02.2020. URL: https://elibrary.ru/wos_scopus_rsci.asp (дата обращения: 01.12.2023). EDN: CQMPRA.

18. Russian index of science citation: Overview and review / O. Moskaleva, V. Pislyakov, I. Sterligov [et al.] // Scientometrics. 2018. Vol. 116. P. 449–462. DOI: 10.1007/s11192-018-2758-y. EDN: XTIRDN.

19. The Russian Science Citation Index (RSCI): the first three years (2016–2018) / S. V. Gorin, A. M. Koroleva, A. N. Gerasimov, A. A. Voronov // European Science Editing. 2020. Vol. 46. DOI: 10.3897/ese.2020.e51051. EDN: XDXXDQ.

20. *Мальцева Д.В., Ващенко В.А., Капустина Л.В.* Методология обработки библиографических данных на русском языке для построения сетей коллаборации (на примере базы данных eLibrary) // Социология: методология, методы, математическое моделирование (Социология: 4М). 2022. № 54–55. С. 45–78. DOI: https://doi.org/10.19181/4m.2022.31.1-2.2. EDN: GRRLBQ.

21. *Batagelj V.* WoS2Pajek. Networks from web of science. Version 1.5 (2017). URL: http://vladowiki.fmf.uni-lj.si/doku.php?id=pajek:wos2pajek (дата обращения: 01.12.2023).

# Приложение

## АНАЛИЗ БАЗ ДАННЫХ, ИНДЕКСИРОВАННЫХ В ELIBRARY

В массиве eLibrary для каждой публикации содержится информация о том, в каких еще наукометрических базах (РИНЦ, RSCI, WoS, Scopus, ВАК) она проиндексирована. Это дает возможность посмотреть на распределение и пересечение публикаций в разных базах на основе информации, предоставляемой площадкой eLibrary.

Как видно из табл. 1, большинство публикаций из массива eLibrary представлены в журналах, индексируемых в базе РИНЦ (85%) и входящих в список ВАК (53%). Значительно меньшее количество статей опубликованы в журналах, индексируемых в RSCI (11%), Scopus (8%) и WoS CC (7%). Безусловно, множества, составляемые массивами публикаций в разных базах, являются пересекающимися (статьи могут входить в разные базы). На основе имеющихся данных о принадлежности статей в массиве eLibrary к разным базам можно посмотреть на пересечение (как множество общих единиц) и объединение (как множество всех единиц) между различными базами на уровне публикаций и журналов.

*Таблица 1*

ИНДЕКСАЦИЯ ПУБЛИКАЦИЙ ИЗ МАССИВА ELIBRARY В РАЗЛИЧНЫХ БАЗАХ

| База | Показатель | Входит в базу | Не входит в базу | Сумма |
|---|---|---|---|---|
| РИНЦ/RISC | абсолютные значения | 65 103 | 10 129 | 75 232 |
| | доля, % | 87 | 13 | 100 |
| ВАК | абсолютные значения | 40 046 | 35 186 | 75 232 |
| | доля, % | 53 | 47 | 100 |
| RSCI | абсолютные значения | 8 179 | 67 053 | 75 232 |
| | доля, % | 11 | 89 | 100 |

*Окончание табл. 1*

| База | Показатель | Входит в базу | Не входит в базу | Сумма |
|---|---|---|---|---|
| Scopus | абсолютные значения | 6 222 | 69 010 | 75 232 |
| | доля, % | 8 | 92 | 100 |
| WoS | абсолютные значения | 5 226 | 70 006 | 75 232 |
| | доля, % | 7 | 93 | 100 |

*Сходство баз на уровне публикаций.* Разные комбинации баз данных составляют разные доли от общего числа статей в массиве eLibrary (табл. 2).

● Некоторые включенности одних множеств в другие объясняются известной информацией о создании баз: так, все публикации из RSCI по определению включены в базу РИНЦ, поскольку являются подмножеством статей, опубликованных в российских топ-журналах.

● Известно, что база РИНЦ включает публикации российских авторов, представленные в журналах WoS и Scopus, публикации в массиве, входящие в эти базы, также полностью (5226 для WoS) и почти полностью (6212 из 6222 для Scopus) входят в РИНЦ.

● Ситуация для базы RSCI несколько иная: из 8179 публикаций в этой базе российских топ-журналов в журналах WoS CC также индексированы 4263 работ (52%), а из 5226 публикаций в WoS 963 статьи (18,4%) не входят в RSCI, но индексируются только в WoS CC (и попадают в базу благодаря тому, что их индексирует РИНЦ).

● Число статей из базы RSCI, также индексированных в базе Scopus, составляет 5249 (64,2% от всех публикаций в RSCI), а число уникальных статей из Scopus в нашей базе составляет 973 статьи (15,6% от всех публикаций в Scopus).

● Общее пересечение статей из собранного массива данных (75 232 публикации), входящих, по данным eLibrary, и в WoS, и в Scopus, составляет 4170 статей – что составляет 79,8% от всех публикаций WoS в массиве и 67% от всех публикаций в Scopus.

● Аналогичная доля рассчитывается и на пересечении этих двух баз и РИНЦ (опять же, по природе создания базы), но если сравнивать с базой RSCI, то число общих статей на пересечении трех баз составляет 3875 (что составляет 74,1% от всех статей в WoS, 62,3% – в Scopus и 47,4% – в RSCI).

● Общее число публикаций, представленных во всех трех базах (RSCI, Scopus, WoS), которые составляют ядро РИНЦ, рассчитанное как объединение множеств, составляет 9820 публикаций – 13% от всех публикаций в собранном массиве (75 232 публикации).

Обращает на себя внимание интересный факт: табл. 1 показывает, что не все статьи, вошедшие в массив eLibrary, входят в базу РИНЦ – только 87%. Предполагая, что оставшиеся 13% статей распределены по другим базам, указанным в массиве eLibrary, мы посмотрели на объединения баз Scopus, WoS и РИНЦ, а также объединение всех пяти баз (табл. 2). Выяснилось, что первое объединение составляет 65 113 публикаций, а второе – 65 308 публикаций – то есть снова около 87% публикаций из базы. Оставшиеся 9924 статей не входят ни в одну из пяти баз, указанных в eLibrary.

*Таблица 2*

СХОДСТВО МЕЖДУ БАЗАМИ ДАННЫХ ПО ЧИСЛУ ПУБЛИКАЦИЙ (МАССИВ ELIBRARY)

| База | Число публикаций | Доля от общего числа статей, % |
|---|---|---|
| Пересечение (множество общих статей – правило «И») | | |
| РИНЦ + RSCI | 8179 | 10,9 |
| РИНЦ + WoS | 5226 | 6,9 |
| РИНЦ + Scopus | 6212 | 8,3 |
| RSCI + WoS | 4263 | 5,7 |
| RSCI + Scopus | 5249 | 6,98 |
| WoS + Scopus | 4170 | 5,5 |
| Scopus + WoS + РИНЦ | 4170 | 5,5 |
| Scopus + WoS + RSCI | 3875 | 5,2 |
| РИНЦ + ВАК | 39 841 | 52,6 |

*Окончание табл. 2*

| База | Число публикаций | Доля от общего числа статей, % |
|---|---|---|
| Объединение (множество всех статей – правило «ИЛИ») | | |
| Scopus + WOS + RSCI (ядро РИНЦ) | 9820 | 13 |
| Scopus + WOS + РИНЦ | 65 113 | 86,5 |
| Scopus + WOS + RSCI + ВАК | 40 292 | 53,6 |
| Scopus + WOS + RSCI + ВАК + РИНЦ | 65 308 | 86,8 |
| РИНЦ + ВАК | 65 308 | 86,8 |

Более внимательный анализ этого подмассива работ показал, что они опубликованы в журналах, с которыми заключено лицензионное соглашение на размещение издания на eLibrary.ru. Кроме того, выборочный анализ некоторых журналов с помощью системы SCIENCE INDEX на eLibrary показал, что в определенные периоды времени эти журналы индексировались в РИНЦ. Топ журналов из данного подмассива приведен в табл. 3; ведущим источником выступает «Экономика и социум» с 1759 статьями (срочные платные публикации). Таким образом, в ходе анализа была уточнена реализованная стратегия сбора данных, осуществленная ООО «НЭБ»: отбор статей, индексируемых eLibrary, не аналогичен отбору по статьям, индексируемым в РИНЦ.

*Таблица 3*

ЖУРНАЛЫ С НАИБОЛЬШИМ КОЛИЧЕСТВОМ СТАТЕЙ ИЗ ПОДМАССИВА ПУБЛИКАЦИЙ, ИНДЕКСИРУЕМЫХ ТОЛЬКО В ELIBRARY

| № | Название журнала | Кол-во статей | № | Название журнала | Кол-во статей |
|---|---|---|---|---|---|
| 1 | Экономика и социум | 1759 | 12 | Студенческий | 166 |
| 2 | Молодой ученый | 441 | 13 | Гуманитарные научные исследования | 145 |

*Окончание табл. 3*

| № | Название журнала | Кол-во статей | № | Название журнала | Кол-во статей |
|---|---|---|---|---|---|
| 3 | Сборники конференций НИЦ Социосфера | 425 | 14 | Современные тенденции развития науки и технологий | 137 |
| 4 | Аллея науки | 327 | 15 | Студенческий вестник | 122 |
| 5 | NovaInfo.Ru | 250 | 16 | Научный альманах | 119 |
| 6 | Вестник современных исследований | 193 | 17 | Вестник научных конференций | 116 |
| 7 | Теория и практика современной науки | 176 | 18 | Евразийский союз ученых | 115 |
| 8 | Актуальные проблемы гуманитарных и естественных наук | 176 | 19 | Современные научные исследования и инновации | 110 |
| 9 | Стратегия устойчивого развития регионов России | 174 | 20 | Форум молодых ученых | 106 |
| 10 | Система ценностей современного общества | 171 | 21 | Colloquium-journal | 106 |
| 11 | Сборник научных трудов SWorld | 168 | 22 | Альманах современной науки и образования | 99 |

Еще одно пересечение баз данных относится к публикациям, входящим в список ВАК. По данным eLibrary, всего в собранном массиве из 75 232 работ 40 046 публикаций входят в эту базу и их подавляющее большинство (99,5%) входит в РИНЦ; разница между базами составляет 205 журналов. При объединении же

множества ВАК со Scopus, WoS и RSCI получается 40 292 публикации – всего на 246 больше, чем в базе ВАК. Получается, что список ВАК в значительной степени состоит из журналов, индексируемых в этих трех базах (ядре РИНЦ). Результат объединения пяти баз в числовом выражении аналогичен объединению множеств РИНЦ и ВАК – то есть все статьи, индексируемые в ядре РИНЦ, входят в эти два множества.

Сходство баз на уровне журналов. Количество журналов, индексируемых в разных базах по массиву eLibrary (на основании представленных площадкой данных), показано в табл. 4.

*Таблица 4*

ИНДЕКСАЦИЯ ЖУРНАЛОВ ИЗ МАССИВА ELIBRARY В РАЗЛИЧНЫХ БАЗАХ

| База | Количество журналов | Доля от общего числа журналов, % |
|---|---|---|
| РИНЦ | 3580 | 91,56 |
| RSCI | 202 | 5,17 |
| Scopus | 193 | 4,94 |
| WoS | 148 | 3,79 |
| ВАК | 1310 | 33,5 |
| Всего журналов | 3910 | 100 |

Подавляющее число источников (91,6%), в которых опубликованы статьи в базе, индексируются в РИНЦ; доли журналов, индексируемых в базах RSCI, Scopus, WoS, небольшие и составляют 4–5% от всех журналов. Треть всех журналов включены в список ВАК.

По анализируемому массиву данных из всех журналов в РИНЦ, где опубликованы работы по социологии, в число топ-журналов, отобранных для базы RSCI, входит 202 журнала. Число журналов из РИНЦ, индексируемых в WoS, составляет 148 журналов, а в Scopus – 193 журнала; на пересечении эти две зарубежные

базы дают в РИНЦ 92 журнала (что составляет 48% от всех журналов в Scopus и 62% от всех журналов в WoS). Число журналов из RSCI на пересечении с WoS дает 51 журнал, а со Scopus – 83, на пересечении трех баз находится 41 журнал.

Полученные результаты демонстрируют, что база РИНЦ, максимально близкая по размеру базе eLibrary (но не полностью покрывающая ее), в значительной степени пересекается с другими, меньшими по размеру базами библиографических данных, в том числе с базой WoS. В тексте статьи поиск такого пересечения производится через сравнениие двух рассматриваемых массивов.

***Maltseva Daria V.,***
*Candidate of Sciences in Sociology, Head of the International Laboratory for Applied Network Research, HSE University, Moscow, Russia,* dmaltseva@hse.ru

***Pavlova Irina A.,***
*Candidate of Sciences in Economics, Deputy Head of the International Laboratory for Applied Network Research, HSE University, Moscow, Russia,* iapavlova@hse.ru

***Kapustina Lika V.,***
*Research Assistant at the International Laboratory for Applied Network Research, HSE University, Moscow, Russia,* lkapustina@hse.ru

***Vashchenko Vasilisa A.,***
R*esearch Assistant at the International Laboratory for Applied Network Research, HSE University, Moscow, Russia,* vvashchenko@hse.ru

***Fiala Dalibor,***
*Associate Professor at the Faculty of Applied Sciences, Department of Computer Science and Engineering, West Bohemian University, Czech Republic, Pilsen,* dalfia@kiv.zcu.cz

**Comparative analysis of the capabilities of WoS and eLibrary for analyzing bibliographic networks**

This article presents a comparative analysis of two major scientific publication databases: Web of Science Core Collection and eLibrary – to identify their differences and unique opportunities for exploration of bibliographic networks of Russian scientific authors. Current shortage of tools and approaches for collection, processing and analysis of bibliographic data in the Russian language constitutes the relevance of this study. Empirical analysis is based on comparison of respective arrays of scientific publications in the field of sociology over the period of 2010-2021. We propose a set of comparison criteria including those related to the procedure of data access, quality of data management, quantitative and qualitative features of the data. Inspection of the databases based on the proposed criteria aids in identification of intersections between both the collections and the respective qualitative observations about them. We make conclusions regarding the comparative advantages and weaknesses of both databases in regards to their potential as the sole data source for bibliographic studies, and make recommendations for their

effective use in research on Russian science.
*Keywords:* network analysis, comparative analysis, bibliographic databases, bibliographic networks, eLibrary, Web of Science

**References**

1. Bar-Ilan J. Informetrics at the beginning of the 21st century – A review, *Journal of informetrics*. 2008, vol. 2, p. 1–52. DOI: 10.1016/j.joi.2007.11.001.
2. Mingers J., Leydesdorff L. A review of theory and practice in scientometrics, *European journal of operational research*, 2015, vol. 246, no. 1, p. 1–19. DOI: 10.1016/j.ejor.2015.04.002.
3. Rousseau R., Egghe L., Guns R. *Becoming metric-wise: A bibliometric guide for researchers*, ed. by W. Glänzel [et al.]. Cambridge, MA: Chandos Publishing, 2018. 850 p. ISBN: 0081024754, 9780081024751.
4. Batagelj V., Doreian P., Ferligoj A., Kejžar N. *Understanding large temporal networks and spatial networks: Exploration, pattern searching, visualization and network evolution*. Hoboken, NJ: WileyBlackwell, 2014, 464 p. ISBN: 978-1-118-91537-0.
5. Moiseev S.P., Maltseva D.V. Source selection for systematic literature reviews: a comparison of expert and algorithmic approaches (in Russian), *Sotsiologiya 4M (Sociology: methodology, methods, mathematical modeling)*, 2018, no. 47, p. 7–43.
6. Bylucheva E.E., Maltseva D.V. Identifying relevant topics in sociology: a bibliographic network analysis view (in Russian), *Monitoring of Public Opinion: Economic and Social Changes*, 2020, no. 6 (160), p. 113–140. DOI: 10.14515/monitoring.2020.6.971.
7. Harzing A.W., Alakangas S. Google Scholar, Scopus and the Web of Science: a longitudinal and cross-disciplinary comparison, *Scientometrics*, 2016, vol. 106, p. 787–804. DOI: 10.1007/s11192-015-1798-9.
8. Singh V.K., Singh P., Karmakar M. [et al.] The journal coverage of Web of Science, Scopus and Dimensions: A comparative analysis, *Scientometrics*, 2021, vol. 126, p. 5113–5142. DOI: 10.1007/s11192-021-03948-5.
9. Martín-Martín A., Thelwall M., Orduna-Malea E., Delgado López-Cózar E. Google Scholar, Microsoft Academic, Scopus, Dimensions, Web of Science, and OpenCitations' COCI: a multidisciplinary comparison of coverage via citations, *Scientometrics*, 2021, vol. 126, no. 1, p. 871–906. DOI: 10.1007/s11192-020-03690-4.

10. Harzing A.W. Two new kids on the block: How do Crossref and Dimensions compare with Google Scholar, Microsoft Academic, Scopus and the Web of Science? *Scientometrics*, 2019, vol. 120, no. 1, p. 341–349. DOI: 10.1007/s11192-019-03114-y.
11. Zhu J., Liu W. A tale of two databases: The use of Web of Science and Scopus in academic papers, *Scientometrics*, 2020, vol. 123, no. 1, p. 321–335. DOI: 10.1007/s11192-020-03387-8.
12. Gusenbauer M. Google Scholar to overshadow them all? Comparing the sizes of 12 academic search engines and bibliographic databases, *Scientometrics*, 2019, vol. 118, no.1, p. 177–214. DOI: 10.1007/s11192-018-2958-5.
13. Moed H.F., Markusova V., Akoev M. Trends in Russian research output indexed in Scopus and Web of Science, *Scientometrics*, 2018, vol. 116, p. 1153–1180. DOI: 10.1007/s11192-018-2769-8.
14. Vera-Baceta M.A., Thelwall M., Kousha K. Web of Science and Scopus language coverage, *Scientometrics*, 2019, vol. 121, no. 3, p. 1803–1813. DOI: 10.1007/s11192-019-03264-z.
15. Ruiz-Pérez R., López-Cózar E.D., Jiménez-Contreras E. Spanish personal name variations in national and international biomedical databases: implications for information retrieval and bibliometric studies, *Journal of the medical library association*, 2002, vol. 90, no. 4, p. 411–430.
16. Adriaanse L.S., Rensleigh C. Web of Science, Scopus and Google Scholar: A content comprehensiveness comparison, *The Electronic Library*, 2013, vol. 31, no. 6, p. 727–744. DOI: 10.1108/EL-12-2011-0174.
17. Eremenko G.O. Comparing publication levels of Russian scientists across Web of Science, Scopus and RSCI databases (in Russian), *NAB (Scientific Electoring Library)*, 28.02.2020, URL: https://elibrary.ru/wos_scopus_rsci.asp (date of access: 01.12.2023).
18. Moskaleva O., Pislyakov V., Sterligov I. [et al.] Russian index of science citation: Overview and review, *Scientometrics*, 2018, vol. 116, p. 449–462. DOI: 10.1007/s11192-018-2758-y.
19. Gorin S.V., Koroleva A.M., Gerasimov A.N., Voronov A.A. The Russian Science Citation Index (RSCI): the first three years (2016–2018), *European Science Editing*, 2020, vol. 46. DOI: 10.3897/ese.2020.e51051.
20. Maltseva D.V., Vashchenko V.A., Kapustina L.V. Methodology of processing bibliographic data in Russian language to construct collaboration networks (using the example of the eLibrary database) (in Russian),

*Sotsiologiya 4M (Sociology: methodology, methods, mathematical modeling)*, 2022, no. 54–55, p. 45–78. DOI: 10.19181/4m.2022.31.1-2.2.

21. Batagelj V. *WoS2Pajek. Networks from web of science*, Version 1.5 (2017). URL: http://vladowiki.fmf.uni-lj.si/doku.php?id=pajek:wos2pajek (date of access: 01.12.2023).

DOI: 10.19181/4m.2023.32.1.2
EDN: SJPPOZ

В.А. Ващенко
(*Москва*)

# ТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ ДЛЯ КОРОТКИХ ТЕКСТОВ: СРАВНИТЕЛЬНЫЙ АНАЛИЗ АЛГОРИТМОВ[1]

Устойчивый рост популярности социальных сетей в качестве средства коммуникации актуализирует методологические вопросы, связанные с особенностями обработки коротких текстов, обладающих меньшим семантическим контекстом, чем крупные тексты, широко используемые для обучения и тестирования моделей машинного обучения для работы с текстовыми данными. Тематическое моделирование – метод машинного обучения «без учителя», нацеленный на агрегацию текстов в тематические кластеры, – имеет множество академических и практических приложений в случаях отсутствия подробной разметки текстовых данных. Однако качество работы алгоритмов тематического моделирования может ограничиваться полнотой семантического контекста, необходимого для качественного числового представления единицы текста. В этой статье рассматриваются шесть разных подходов к тематическому моделированию, основанных на различающихся принципах концептуализации текста и тем. Сравнивается качество работы указанных алгоритмов на наборе русскоязычных комментариев в сети TikTok и проводится формальная оценка скорости и когерентности результирующих тем.

**Василиса Андреевна Ващенко** – стажер-исследователь Международной лаборатории прикладного сетевого анализа Национального исследовательского университета «Высшая школа экономики», Москва, Россия. Email: vvashchenko@hse.ru.

[1] Статья подготовлена в ходе проведения исследования в рамках Программы фундаментальных исследований Национального исследовательского университета «Высшая школа экономики» (НИУ ВШЭ).

*Ключевые слова:* тематическое моделирование, анализ текстовых данных, блокмоделинг, прикладной сетевой анализ, анализ социальных медиа, трансформерные модели

## Введение

Тематическое моделирование – метод вычленения тематических кластеров в корпусе текстов – является важным направлением в методологических исследованиях в области социологии, поскольку предоставляет возможность количественного изучения тематической композиции текстовых материалов: структуры дискуссии, художественных или научных текстов.

Хотя алгоритмы тематического моделирования склонны обобщать выделяемые темы и хуже справляются с выделением узких тематических групп, чем ручное кодирование, что иногда приводит к расхождениям с результатами ручной разметки [1], они успешно применялись в рекомендательных системах [2], наукометрических задачах [3], анализе дискурса [2; 4; 5], автоматической идентификации событий в новостях и социальных медиа [6; 7; 8] и во многих иных практических приложениях, включая анализ нетекстовых источников (например, изображений [9] и аудио [10]).

Вместе с популяризацией социальных сетей в качестве ключевых инструментов массовой коммуникации в современном мире растет и интерес исследователей к социальным медиа как к источнику знания об общественном мнении. В свою очередь крупные объемы информации, производимые пользователями социальных сетей на ежедневной основе, диктуют новые требования к масштабам эмпирической работы [5], необходимой для полноценного описания исследуемого общественного феномена, что стимулирует использование количественных техник анализа текстовых данных исследователями. Как отмечает А. Бызов, алгоритмы тематического моделирования способны оказать исследователю поддержку в разметке или выделении признаков из текстовых данных, особенно

в ситуации, когда объем массива текста затрудняет ручное кодирование [11]. Так, методы тематического моделирования использовались для исследования общественного мнения о COVID-19 на базе больших данных в Twitter [6; 12]. Однако многие алгоритмы, используемые для задач тематического моделирования, расходятся в своих базовых предпосылках с форматом коммуникации, характерным для социальных медиа: короткие тексты сообщений в социальных сетях, подобных Twitter, не позволяют достигать планок качества, сопоставимых с работой тех же инструментов на длинных текстах ввиду недостаточности контекста для слов-токенов [13; 14; 15]. Учитывая, что многие тексты, производимые пользователями социальных сетей, ограничены платформенными лимитами[1], длинные тексты зачастую недоступны исследователям онлайн-дискурса. Следовательно, необходимо формализованное сравнение качества разных алгоритмов тематического моделирования для коротких текстов с целью обнаружения наиболее эффективного с вычислительной и интерпретационной точек зрения подхода. Несмотря на наличие подобных усилий в современном академическом поле, методы, анализируемые исследователями, как правило, ограничены вариациями и адаптациями латентного размещения Дирихле (Latent Dirichlet Allocation, LDA) и словарными эмбеддингами – векторными представлениями, репрезентирующими семантический контекст слова. В свою очередь мы предлагаем проанализировать эффективность методов, альтернативных конвенционально используемому LDA. Мы включаем в анализ методы тематического моделирования, основанные как на вероятностном моделировании соприсутствия слов, так и на методах выделения сообществ в бимодальных сетях и кластеризации предобученных словарных эмбеддингов.

[1] Все платформы массовой коммуникации используют лимиты символов на разные виды текстов, производимых на платформе: публикации, комментарии, описания и т.д. Как правило, эти лимиты достаточно низкие, что заметно на примере двух из трех крупнейших социальных сетей в России на 2023 г., согласно данным исследовательской компании Mediascope [16]: для TikTok это 150 символов на комментарий, для «ВКонтакте» – 280 символов на комментарий.

Разнообразие сравниваемых методов позволяет более конкретно изучить связь между устройством подхода к тематическому моделированию и особенностями результирующих тематических кластеров, а также отметить значимые направления потенциального развития наличествующих инструментов для задач анализа коротких текстов. В настоящей статье проводится формальное сравнение качества работы новых алгоритмов тематического моделирования на базах данных новостных сводок и комментариев в социальных медиа, а также предлагаются пути выбора модели и улучшения ее качества для задач тематического моделирования в социальных исследованиях.

## Обзор исследований по теме тематического моделирования для коротких текстов

Методы тематического моделирования получили широкое признание вместе с появлением в конце 1990-х гг. ныне конвенциональных подходов вероятностного латентного семантического анализа (Probabilistic Latent Semantic Analysis, PLSA) [17] и его последующей адаптации в байесовской парадигме – LDA, моделирующей, в отличие от PLSA, глобальную структуру тем, предполагая, что слова в разных документах одного корпуса производны из одного и того же набора тем [18]. Важным нововведением LDA стало использование распределения Дирихле в качестве априорного для распределения тем в документах и слов в темах[1]. Этот шаг выполняет регуляризирующую функцию для тематической модели, выравнивая распределения тем за счет ограничения экстремальных значений и снижая чувствительность модели к отдельным документам.

Вероятностные модели представления парных связей между документами и словами-токенами хорошо подошли к анализу

[1] За более подробным описанием технической части алгоритма LDA читатель может обратиться к оригинальной статье [18] или существующим обзорам на русском языке [19; 20].

художественных и научных текстов, поскольку качество предсказания для таких моделей улучшается пропорционально количеству наблюдений и объему контекста: чем больше информации о паттернах соприсутствия слов представлено в обучающем корпусе, тем более связные латентные семантические кластеры производятся LDA. Однако для коротких текстов, доминирующих в современных медиа и представляющих непосредственный интерес для исследований общественного мнения [21], традиционные алгоритмы могут не подходить ввиду двух ограничений.

1. Вероятностные модели опираются на локальные паттерны соприсутствия в рамках документа – семантический контекст слова, который в коротких документах зачастую оказывается недостаточным.

2. Предпосылка наличия распределения тем внутри документа спорна для коротких текстов, поскольку многие из них, в отличие от более длинных документов, содержат только одну тему [15; 22].

Для разрешения этих проблем существует несколько способов: так, короткие текстовые данные предлагается обогащать метаданными [23], тегами авторов [24] и хештегами [22], использовать длинные тексты для обучения модели тематического моделирования для предсказания тем на коротких текстах [25], агрегации коротких текстов в «псевдодлинные» по заданному признаку, чтобы применять к ним традиционные методы тематического моделирования [14]. Отмечается, что использование метаданных или иных форм дополнительной информации методологически нежелательно, поскольку во многих случаях подобные данные могут быть недоступны [15]. Более того, эти адаптации не решают ряд иных значимых методологических проблем – отсутствия формального критерия выбора количества тем[1] и обоснования предпосылки о соответствии распределения тем и слов внутри

[1] Стоит отметить, что формальные критерии выбора количества тем развиваются и обсуждаются в литературе: например, примечательно применение энтропийного подхода к задаче настройки этого гиперпараметра [19].

темы распределению Дирихле [26]. Как отмечают авторы сетевого подхода к тематическому моделированию М. Герлах, Т.П. Пейшото и Э.Г. Алтманн, несмотря на высококачественные результаты работы алгоритма LDA, выбор распределения Дирихле для моделирования распределения тем в текстовом корпусе является его ограничением, поскольку репрезентация текста в LDA противоречит иным известным в лингвистике паттернам в языке (к примеру, закону Ципфа), указывающим на неравномерность в частотности появления и соприсутствия слов в текстах [26].

В качестве альтернативы семантическому обогащению модели за счет метаданных или расширения контекста Герлах, Пейшото и Алтманн предлагают подойти к задаче тематического моделирования как к задаче выделения сообществ в сетях. Доказывая формальную эквивалентность между PLSA и стохастической блокмоделью (Stochastic Blockmodel, SBM[1]), авторы рассматривают LDA как частный кейс своего непараметрического алгоритма, предлагающего возможность отказа от предпосылок об униформном распределении тематических кластеров в корпусе за счет использования иерархии априорных распределений для моделирования распределения тем [26]. Иерархическая модель допускает гетерогенность в данных и неоднородность выделяемых тем: темы могут быть более или менее «плотными», а также объединяться в структуры более высокого порядка при подъеме по «ступеням» иерархии. Более того, непараметрические модели позволяют избавиться от количества тем как предзаданного гиперпараметра и моделировать количество тем в ходе обучения [28]

Идея иерархичности в формировании тем применяется также в новых разработках среди моделей тематического моделирования,

[1] Стохастические блокмодели (SBM) являются генеративными вероятностными моделями, используемыми для группировки вершин в сетях в блоки или сообщества. Главной предпосылкой модели является положение, указывающее, что вершины, относящиеся к одному блоку, имеют более высокую вероятность связи между ними, чем вершины, относящиеся к разным блокам [27].

ориентированных на расширение семантического контекста анализируемых документов [29]. Так, активно развивается приложение трансформерных моделей к задаче тематического моделирования, поскольку эмбеддинги, используемые трансформерами, позволяют репрезентировать сложные семантические структуры в текстах на естественном языке, сохраняя больше локальной информации, чем статические предобученные эмбеддинги. «Трансформером» называется разновидность нейросетевой архитектуры, широко применяемой в обработке естественного языка; наиболее значимым отличием трансформеров является использование механизма внимания (attention) при тренировке для оценки взаимной важности слов в контексте друг для друга, а также внедрение позиционных меток в эмбеддинги слов для сохранения последовательности в текстовых данных. Создание представлений текста при помощи трансформеров предоставляет возможность не только различать омонимичные выражения, но и учитывать структурную позицию токена[1] в документе при кодировке [30]. В случае использования представлений трансформера для выделения тем, тематическое моделирование заключается в кластеризации эмбеддингов с сокращенной размерностью, где использование иерархических методов кластеризации позволяет выделять кластеры разного размера и плотности, а значит – достигается ранее упомянутая репрезентация гетерогенности в данных [29; 31].

Таким образом, область тематического моделирования активно развивается в сторону расширения типов данных, для которых выделение латентных семантических структур работает эффективно и корректно, включая короткие тексты. Наиболее популярным на-

[1] В лингвистических моделях оптимальной единицей анализа не всегда является слово. В зависимости от задачи, единицей анализа могут выступать слово, словосочетание или часть слова (слог или пары/триады букв). Такая единица анализа называется токеном (от *англ*. ‘token’ – знак, символ). При обучении моделей типа «трансформер» в качестве токенов, как правило, выступают высокочастотные сочетания букв, необязательно составляющие полный слог.

правлением анализа является обогащение контекста коротких текстов при помощи предобученных эмбеддингов или иных видов дополнительных данных для модели, однако присутствуют также и шаги в сторону смены концептуальной рамки подхода к задаче тематического моделирования, одним из которых является использование методов сетевого анализа для выделения тематических кластеров.

## Методология

*Выбор архитектур для тематического моделирования*

Поскольку ключевой целью настоящей статьи является сравнение разных подходов к тематическому моделированию, мы прибегаем к сравнению алгоритмов, относящихся к разным группам методов. Опираясь на категоризацию, предложенную в обзоре А. Абдельразика с соавторами [31], мы используем модели из трех разных выделенных в обзоре групп: алгебраических, вероятностных и нейросетевых моделей тематического моделирования (табл. 1).

К алгебраическим моделям в нашей подборке алгоритмов относится NMF. Этот алгоритм позволяет выявить темы путем разложения матрицы «документ-токен» на неотрицательные матрицы (матрицы, все элементы которых больше или равны нулю), отражающие связи «документ-тема» и «тема-токен», что обеспечивает более интерпретируемое и контекстуально значимое представление текстовых данных, поскольку в случае репрезентации связей токенов, тем и документов негативные значения неинтерпретируемы.

Вероятностные модели представлены LDA и SBMTM. LDA моделирует тематическую структуру, вероятностно распределяя слова по темам, а темы по документам на основании априорного распределения Дирихле, что позволяет обнаруживать скрытые тематические паттерны. В SBMTM темы выделяются как сооб-

*Таблица 1*

КАТЕГОРИЗАЦИЯ СРАВНИВАЕМЫХ АЛГОРИТМОВ ТЕМАТИЧЕСКОГО МОДЕЛИРОВАНИЯ

| Модель | Документация | Категория |
|---|---|---|
| Non-Negative Matrix Factorization (NMF) [32] | Реализация модели в библиотеке Gensim для Python: https://radimrehurek.com/gensim/models/nmf.html | Алгебраические |
| Latent Dirichlet Allocation (LDA) [18] | Реализация модели в библиотеке Gensim для Python: https://radimrehurek.com/gensim/models/ldamodel.html | Вероятностные |
| Hierarchical Stochastic Block Model for Topic Modeling (SBMTM) [26] | Реализация модели при помощи библиотеки graph-tool для Python: https://github.com/martingerlach/hSBM_Topicmodel | |
| Embedded Topic Model (ETM) [33] | Исходный код для Python и PyTorch: https://github.com/adjidieng/ETM | Нейросетевые |
| Product-of-Experts LDA (ProdLDA) [34] | Исходный код для PyTorch реализации: https://github.com/estebandito22/PyTorchAVITM | |
| Contextualized Topic Model (CTM) [35] | Исходный код для Python и подробные примеры применения: https://github.com/MilaNLProc/contextualized-topic-models | |
| BERTopic [29] | Описание алгоритма и инструкции по установке: https://maartengr.github.io/BERTopic/index.html<br><br>Исходный код для Python: https://github.com/MaartenGr/BERTopic | |

щества в бимодальной сети соприсутствия слов, где в качестве двух типов вершин выступают документы и слова, при помощи иерархических стохастических блокмоделей. Иерархичность подхода позволяет генерализовать предпосылки о распределениях, описывающих связь между темами и словами. Герлах, Пейшото и Альтманн утверждают, что такой подход позволяет формировать более качественные темы за счет учета как плотных, так и разреженных групп разного размера [26].

Ввиду сравнительной новизны нейросетевых моделей и перспективности их приложения к анализу коротких текстов за счет более нюансированного задействования семантического контекста, они наиболее широко представлены в нашей подборке алгоритмов для анализа. Мы рассмотрим четыре модели: ETM, ProdLDA, CTM и BERTopic. ETM объединяет в себе векторные представления слов и вероятностное моделирование тем. В отличие от традиционных методов, ETM представляет документы и темы в непрерывном семантическом пространстве, что позволяет словам вносить переменный вклад в темы. Это позволяет ETM улавливать тонкие семантические связи между словами и темами, предоставляя более гибкий и интерпретируемый подход к моделированию тем.

ProdLDA является адаптацией LDA, задействующей в своей архитектуре вариационные автокодировщики[1] (Variational

[1] Вариационные автокодировщики (VAE) – это генеративная модель, предназначенная для обучения вероятностному представлению входных данных и генерации новых образцов на основе изученного распределения. VAE состоят из кодировщика, который отображает входные данные в вероятностное скрытое пространство, и декодировщика, который восстанавливает данные из образцов, взятых из скрытого пространства. Обучение VAE побуждает выучиваемое латентное пространство следовать заданному распределению вероятностей и облегчает генерацию новых, значимых образцов данных. В целом VAE используют вероятностные принципы для обучения структурированному и непрерывному представлению входных данных, что делает их крайне полезными для таких задач, как генеративное моделирование, синтез данных и обучение латентных векторных представлений (эмбеддингов).

Autoencoders, VAE) для улучшения репрезентации тем [34]. Еще одной значимой чертой ProdLDA является использование product of experts («произведение экспертов») подхода: такие модели используют несколько «экспертных» моделей, специализирующихся на разных аспектах/частях данных, вместо одной, для вероятностного моделирования тренировочных данных. В традиционной модели LDA предполагается, что темы генерируются независимо для каждого документа. Вместо предположения о независимости ProdLDA моделирует совместное распределение тем и документов как произведение распределений моделей-«экспертов». Каждый «эксперт» связан с отдельным документом, и каждый эксперт вносит свой вклад в общее распределение. Итоговое распределение для документа получается путем взятия произведения распределений экспертов. Это позволяет модели отражать более сложные зависимости между темами и документами, что может быть особенно полезно при наличии нетривиальных зависимостей, которые немогут быть адекватно отражены при допущении независимости.

Модель CTM учитывает контекст каждого слова внутри документа, что позволяет генерировать темы, учитывающие нюансы содержания документа. Модель использует эмбеддинги предобученной модели-трансформера типа BERT для изучения контекстуализированных представлений слов, обеспечивая более точное и динамичное моделирование тем.

Наконец, в BERTopic темы выделяются при помощи иерархической кластеризации внутренних представлений текста предобученной модели трансформера. В рамках использования обеих моделей для запуска необходимо указать предобученную модель-трансформер, векторные представления слов которой ложатся в основу моделирования близости между ними. Мы используем paraphrase-multilingual-mpnet-base-v2 – наилучшую по качеству на открытых бенчмарках предобученную мультиязычную модель

из библиотеки sentence-transformers[1]. Важно отметить различие между CTM и BERTopic в том, что последняя модель выделяет темы как кластеры в пространстве словарных эмбеддингов сокращенной размерности. Мы используем Uniform Manifold Approximation and Projection (UMAP)[2] для сокращения размерности BERT-эмбеддингов и Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN)[3] для кластеризации. UMAP выбирается за счет высокого качества сохранения локальной структуры данных, в то время как HDBSCAN способен выделять кластеры разного размера и плотности, что позволяет обнаруживать разнородные темы в документах. Значимой компонентой качества BERTopic является алгоритм c-TF-IDF (class Total Frequency – Inverse Document Frequency) взвешивания – адаптация классического TF-IDF взвешивания, используемая для отражения уникальности слов в документе по отношению к другим словам той же темы. Она направлена на выделение слов, которые являются хорошо различающимися для конкретной темы, в то время как слова, которые являются общими для всего корпуса, отводятся на второй план [29].

Для обучения ETM, – обогащенного предобученными статическими словарными эмбеддингами метода LDA, – в рамках наших экспериментов сравниваются две предобученные модели статических словарных эмбеддингов (табл. 2).

1. GloVe (Global Vectors) для русского языка Navec [36]. Выбор обоснован малым объемом памяти, необходимой для их использования и дообучения при большом объеме словаря: используемые эмбеддинги, обученные на массиве русскоязычной художествен-

[1] Таблица сравнения доступных моделей представлена по ссылке: https://sbert.net/docs/pretrained_models.html (дата обращения: 05.01.2024).

[2] Подробная документация Python-имплементации UMAP доступна по ссылке: https://umap-learn.readthedocs.io (дата обращения: 05.01.2024).

[3] Подробная документация Python-имплементации HDBSCAN доступна по ссылке: https://hdbscan.readthedocs.io (дата обращения: 05.01.2024).

ной литературы[1], покрывают 98% слов в художественных текстах, занимая 50,6 Мб памяти.

2. Word2Vec (Continuous Skipgram) [37; 38] – эмбеддинги для русского языка, обученные на Национальном корпусе русского языка (НКРЯ). В отличие от Navec, эта модель обучена на словаре с тегами частей речи, что позволяет в некоторых случаях решить проблему омонимии. Предобученная модель представлена в открытом доступе как часть библиотеки Gensim для Python[2].

*Таблица 2*

ОПИСАНИЕ ИСПОЛЬЗУЕМЫХ ЭМБЕДДИНГОВ

| Модель | Размерность | Тип эмбеддинга | Размер словаря (103) | Размер модели (*Mb*) |
|---|---|---|---|---|
| Navec | 300 | GloVe | 500 | 50,6 |
| ruscorpora_300 | 300 | Continuous Skipgram | 180 | 198,8 |

Так, мы сравниваем две предобученные модели статических словарных эмбеддингов, основанные на разных методах формирования векторных представлений слов: GloVe [39] и Continuous Skip-Gram. Continuous Skip-Gram является моделью локального предсказания контекста. Ее цель – предсказать контекстные слова заданного целевого слова [37]. Модель обучается таким образом, чтобы максимизировать вероятность предсказания контекстных слов по заданному слову. GloVe, в свою очередь, основан на глобальном статистическом подходе: целью обучения является минимизация разницы между точечным произведением векторов слов

[1] В настоящей работе используется дефолтная модель Navec, документация которой доступна по ссылке: https://github.com/natasha/navec (дата обращения: 05.01.2024).

[2] Документация соответствующих методов и список доступных моделей доступны по ссылке: https://radimrehurek.com/gensim/models/word2vec.html (дата обращения: 15.12.2023).

и логарифмом вероятностей соприсутствия в корпусе на основе матрицы соприсутствия. В отличие от Word2Vec, эмбеддинги GloVe используют не только слова в непосредственной близости друг друга (контекст), но и глобальные статистики совстречаемости слов друг с другом, что позволяет также определить сравнительную значимость слов в тексте [39]. Так, выбранные эмбеддинги представляют альтернативные подходы к векторному представлению слов.

При работе с хештегами используется классический LDA, для остальных массивов данных добавляются словарные эмбеддинги и модель идентифицируется как ETM [33]. Поскольку многие хештеги содержат те или иные разновидности авторского написания (нарочитые ошибки, совмещение фразы в одно слово и т.д.), применение к ним предобученных векторных представлений слов неэффективно из-за чувствительности последних к корректности написания, роду и числу слов: многие хештеги не войдут в словарь моделей предобученных представлений слов.

Сопоставление вышеописанных алгоритмов (табл. 1), а также разных моделей предобученных эмбеддингов (табл. 2) позволяет проанализировать различия в качестве моделирования тем в зависимости от подхода к представлению текстовых данных, а также агрегации документов, включая внедрение иерархически организованных групп.

*Критерии сравнения*

Для сравнения избранных архитектур используется несколько критериев: время, затрачиваемое на вычисления, когерентность (связность) выделяемых тем, разнообразие/схожесть тем, а также качество классификации текстов как принадлежащих теме на основании смоделированных тематических кластеров. Следует отметить, что в случае с тематическим моделированием существует проблема малого количества «внешних» критериев качества – метрик качества, опирающихся на внешнюю по отношению к продукту моделирования информацию для верификации (как, например, метка

класса в задаче классификации), а не только внутренние свойства результирующих тематических групп [5]. Большинство метрик, используемых для проверки и сравнения алгоритмов тематического моделирования, являются «внутренними»: рассчитываются исходя из собственных свойств тематических групп, смоделированных алгоритмом, и не имеют внешнего референта. Однако на размеченных датасетах задача тематического моделирования может быть трансформирована в задачу классификации, что позволяет использовать «внешние» меры качества.

Метрики когерентности выделяемых алгоритмами тематического моделирования групп измеряют «расстояние» между словами внутри кластера: в случае, если слова, вошедшие в один и тот же кластер, «близки» друг к другу, когерентность кластера будет принимать высокое значение, в обратном случае – низкое. Как правило, для оценки когерентности кластера используются не все входящие в него слова, а N наиболее важных. Мы используем две меры когерентности тем: нормализованную поточечную взаимную информацию (Normalized Pointwise Mutual Information, NPMI) и UMass.

Кратко рассмотрим каждую из них.

1. NPMI количественно определяет родственность слов в теме, учитывая закономерности их совместного появления. Это нормализованная версия поточечной взаимной информации (Pointwise Mutual Information, PMI) между парами слов. PMI измеряет отличие вероятности одновременного появления двух слов в документе относительно ожидаемой при их независимости. Учитывая как силу связи, так и редкость пар слов, что обеспечивает сбалансированную оценку связности, NPMI показывает хорошую корреляцию с человеческой оценкой [28; 40], однако на качество оценки может значимо влиять выбранный размер контекстного окна[1].

[1] «Контекстным окном» называется диапазон слов относительно «центрального» слова, включаемый в анализ при расчете метрик, опирающихся на закономерности и частоты соприсутствия слов. Для примера: контекстное окно размера 3

2. UMass вычисляет PMI между главными словами в теме и их совпадением в референтном корпусе: насколько чаще эти слова встречаются вместе, чем в случайном сценарии [41]. Как и NPMI, эта метрика сенситивна к качеству базового корпуса, однако отличается тем, что измеряет вероятность асимметрично (значение зависит от того, какое слово является «центром», а какое – «контекстом») для топ-слов темы. Эта мера считается наиболее нечувствительной к шуму и является одной из самых быстрых метрик вычисления когерентности.

Разнообразие тем оценивается через метрику, обратную Rank-Biased Overlap (RBO), – меру сходства тем, рекурсивно оценивающую долю пересечений между ранжированными списками слов в темах на разных уровнях «глубины» погружения в список, и меру разнообразия топиков TopicDiversity (далее TD) [33], рассчитываемую как отношение числа уникальных слов к произведению количества тем на $k$, где $k$ – это количество топ-слов в теме, учитываемых в расчете метрики. Предыдущие исследования обнаруживают, что увеличение различности тем повышает интерпретируемость результирующих тем [42], а также принцип различия часто используется во внешней ручной оценке качества тематических моделей с привлечением экспертов, которым предлагается выделить лишнее в теме слово [43].

Для оценки схожести тем используется мера попарного сходства Жаккара (Pairwise Jaccard) – усредненное по количеству сравнений отношение количества общих слов в темах к их совокупному набору слов.

Качество предсказания оценивается при помощи усредненной по классам F1-меры: $2 * \frac{precision * recall}{precision + recall}$. Такой подход к вычислению F1-меры также называется micro F1-мерой. Micro F1 раскла-

относительно слова А предполагает, что мы считаем слово Б соприсутствующим со словом А в рассматриваемом тексте, если оно находится в пределах трех слов слева и трех слов справа от слова А. Контекстные окна используются для ограничения длины контекста, который мы считаем релевантным для каждого слова.

дывает задачу мультиклассовой классификации на несколько задач бинарной классификации, где каждый класс противопоставляется всем прочим сразу, а не по отдельности. Подобная адаптация классической F1-меры менее чувствительна к дисбалансу классов, что актуализирует ее использование с учетом сильного дисбаланса классов в используемых нами данных (см. Приложение 1).

*Эмпирическая база исследования*

Сравнение вышеописанных алгоритмов тематического моделирования проводится на четырех базах данных, преследующих разные задачи: двух размеченных (массивы новостных публикаций MLSUM и Corus) и двух неразмеченных (массивы комментариев и хештегов урбанистической тематики из русскоязычного сегмента TikTok, собранные автором самостоятельно в ходе подготовки настоящей статьи). Для восполнения ограничений, связанных с неизвестностью истинного количества тематических кластеров в базе данных комментариев в TikTok, мы начинаем с анализа размеченных баз данных. Это позволяет оценить, насколько хорошо работает каждый из избранных алгоритмов в условиях возможности проверки соответствия между созданными кластерами и темами, размеченными в массивах данных, что делает дальнейшие выводы на неразмеченной базе данных более надежными.

*Размеченные массивы данных*

Для первичного сравнения качества сравниваемых моделей мы используем два размеченных массива данных для задачи классификации текстов.

1. Размеченная темами выборка сжатых пересказов новостных публикаций из базы данных Multi Lingual Summarization (MLSUM)[1] [44]. Изначальная база данных была очищена от строк

[1] База данных MLSUM с разбиением на тренировочную, валидационную и тестовую выборки доступна на платформе Hugging Face по ссылке: https://huggingface.co/datasets/mlsum (дата обращения: 09.02.2024).

длиннее 150 символов, чтобы приблизить схожесть по длине с комментариями в социальных сетях. Были удалены низкозаполненные и несодержательные темы.

2. Размеченная темами база новостных публикаций Lenta.ru v1.1+ из корпуса Corus[1] (далее мы будем отсылаться к этому датасету как Lenta). В выборку настоящего исследования вошли топ-20 тем по количеству появлений. Для приближения объема текста к референтному – характерному для социальных сетей – в текстах сохранялись первые одно-два предложения. Итоговые тексты не превышают по длине 200 символов.

Оба массива содержат краткие изложения новостных статей и используют назначенные им опубликовавшими их изданиями категории в качестве разметки. Все данные, содержащиеся в массивах MLSUM и Corus, были получены в результате веб-скрейпинга онлайн-страниц избранных новостных изданий, подготовка данных не включала разметки экспертами со стороны авторов массива. Так, соответствие назначенной темы тексту новостной публикации зависит от качества соответствующей работы по категоризации новостей в каждом из использованных в массивах данных изданий.

*Неразмеченные массивы данных*

Для задачи оценки качества на неразмеченном датасете мы используем массив русскоязычных комментариев к урбанистическим видео в TikTok, а также хештеги, присуждаемые этим видео их создателями. Выбор TikTok в качестве платформы для анализа обусловлен устойчивой популярностью сети до блокировки в России, а также адаптацией формата в прочих социальных сетях, до сих пор доступных на территории РФ (YouTube, «ВКонтакте»). Фокус на урбанистической тематике основан на потребности

[1] Описание баз данных и кода для доступа к ним доступны на платформе GitHub по ссылке: https://github.com/natasha/corus (дата обращения: 09.02.2024).

в общей тематической когерентности анализируемых видео, которую можно использовать в качестве базового бенчмарка тематики. Сама по себе тема урбанистики удобна своей популярностью, предоставляющей достаточно объемную базу для исследования, а также способностью вовлекать аудиторию в дискуссию за счет близости темы зрителям, что обеспечивает наличие содержательных обсуждений в комментариях.

Отбор видео осуществлялся по хештегам, список которых расширялся при помощи рекомендательного алгоритма платформы. Финальный набор содержит 17 хештегов. Из массива видео, относящихся к этим хештегам, доступными оказались 2625 видео и содержащими комментарии – 1271. Выгрузка комментариев и хештегов производилась с помощью инструмента Selenium WebDriver для Chrome и языков программирования JavaScript и Python[1].

*Предобработка*

Перед анализом полученные массивы подвергались базовой предобработке.

Шаг 1. Лемматизация – слова были приведены к нормальной форме (инфинитиву для глаголов и именительному падежу единственного числа мужского рода для существительных и прилагательных).

Шаг 2. Удаление стоп-слов – из нормализованных текстов удалялись слова, входящие в наборы русскоязычных стоп-слов в библиотеках NLTK и spaCy для Python. В этот список входят слова с низкой семантической значимостью (предлоги, союзы, местоимения). Далее были удалены слова и хештеги, встречающиеся редко (менее 10 раз) или часто (более чем в 80% документов), а также

[1] Полный код для скрейпинга комментариев в TikTok, использованный для сбора данных в настоящей работе, представлен на платформе GitHub по ссылке: https://github.com/vavaschenko/TikTokScraper (дата обращения: 05.01.2024).

«отметки» – ники других пользователей, возникающие в тексте комментария, когда его автор обращается к другому пользователю. Для всех корпусов, кроме массива хештегов, в анализ включались не только отдельные слова, но и биграммы, сформированные при помощи метода Phraser библиотеки Gensim для Python.

*Таблица 3*

ОПИСАНИЕ ИСПОЛЬЗУЕМЫХ ДАННЫХ

| Датасет | Количество тем | Количество наблюдений | | Размер словаря после обработки |
|---|---|---|---|---|
| TikTok | - | Комментарии | 152 461 | 7021 |
| | | Хештеги | 2625 | 822 |
| MLSUM | 11 | 24 032 | | 6375 |
| Lenta | 20 | 239 151 | | 7617 |

## *Сравнение результатов тематического моделирования на коротких текстах*

*Оценка качества на размеченной базе данных*

Эксперименты по тренировке алгоритмов проводились на виртуальной машине Linux 5.15.109 с Intel(R) Xeon(R) Platinum 8259CL CPU @ 2,50GHz и GPU Tesla V100-SXM2-16GB. Для каждого датасета гиперпараметры (необучаемые параметры, задаются перед началом обучения) рассматриваемых моделей оптимизировались при помощи случайного поиска по сетке значений (см. Приложение 2).

Метрики качества (UMass, NPMI, IRBO, TD, Pairwise Jaccard и F1-мера) в таблицах 4 и 5 приводятся для наилучшей модели по NPMI и рассчитываются на топ-10 словах каждой темы. Для модели SBMTM оцениваются темы, результирующие на двух уровнях иерархии тем: L0 и L1 в таблицах 4 и 5 относятся к нулевому и первому уровню иерархии тем для SBMTM соответственно.

*Таблица 4*

СРАВНЕНИЕ КАЧЕСТВА ТЕМАТИЧЕСКОГО МОДЕЛИРОВАНИЯ НА РАЗМЕЧЕННЫХ ДАННЫХ

| Модель | Датасет | | Время (*мин.*) | Кол-во тем | Когерентность | | Разнообразие | | Схожесть | F1 |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | UMass | NPMI | IRBO | TD | Pairwise Jaccard | |
| BERTopic | MLSUM | | 3 | 265 | **-0,28** | -0,13 | **0,99** | **0,982** | **0,000** | 0,285 |
| | Lenta | | 7 | 719 | *-0,635* | 0,005 | *0,99* | *0,912* | *0,000* | 0,516 |
| ETM (Word2Vec) | MLSUM | | 1 | 14 | -3,311 | -0,016 | 0,974 | 0,279 | 0,029 | 0,331 |
| | Lenta | | 5 | 49 | -3,146 | 0,065 | 0,845 | 0,453 | 0,117 | 0,539 |
| ETM (Navec) | MLSUM | | 0,5 | 38 | -8,338 | -0,119 | 0,938 | 0,532 | 0,044 | 0,438 |
| | Lenta | | 6 | 35 | -3,986 | 0,059 | 0,966 | 0,737 | 0,021 | 0,506 |
| CTM | MLSUM | | 1,5 | 38 | -10,07 | -0,179 | 0,935 | 0,579 | 0,052 | 0,313 |
| | Lenta | | 9 | 49 | -3,596 | *0,164* | 0,989 | 0,745 | 0,009 | 0,582 |
| ProdLDA | MLSUM | | 1,2 | 66 | -8,636 | -0,095 | 0,956 | 0,468 | 0,029 | 0,429 |
| | Lenta | | 22,5 | 61 | -4,069 | 0,138 | 0,995 | 0,815 | 0,004 | 0,54 |
| SBMTM | MLSUM | L0 | 11 | 33 | -8,293 | -0,006 | 1,0* | 1,0* | 0,000* | **0,486** |
| | | L1 | | 6 | -4,367 | 0,023 | 1,0* | 1,0* | 0,000* | 0,376 |
| | Lenta | L0 | 21 | 674 | -10,35 | -0,019 | 1,0* | 1,0* | 0,000* | *0,653* |
| | | L1 | | 150 | -8,15 | -0,008 | 1,0* | 1,0* | 0,000* | 0,588 |
| NMF | MLSUM | | **0,5** | 11 | -4,723 | **0,01** | 0,934 | 0,67 | 0,058 | 0,344 |
| | Lenta | | 26 | 24 | -3,575 | 0,075 | 0,96 | 0,604 | 0,032 | 0,265 |

*Примечание.* Жирным шрифтом отмечены лучшие результаты для MLSUM, курсивом с подчеркиванием – для Lenta.

* В текущей версии алгоритма SBMTM не реализована возможность пересечения между выделяемыми сообществами. Как следствие, каждое слово может попадать только в один кластер, что приводит к строго разнообразным кластерам.

Для оценки различительной способности сформированных тематических кластеров при помощи сравниваемых алгоритмов мы используем случайный лес для задачи классификации: целевыми значениями становятся значения уникальных классов, а предикторами – предсказанные вероятности тематических кластеров для документов в массиве.

Нейросетевые модели превосходят альтернативы по когерентности и разнообразности тем, однако SBMTM L0 превосходит их по качеству классификации. BERTopic является абсолютным лидером по метрикам качества тем, однако значительно уступает и более простым алгоритмам в качестве классификации вследствие того, что модель выделяет очень локальные семантические группы, принадлежность к которым хуже генерализуется для предсказания класса. На высокую специфичность выделяемых тем указывает, в частности, высокое значение разнообразия тем, несмотря на техническую возможность альтернативного результата (в отличие от SBMTM). В свою очередь SBMTM хорошо приближает истинные классы, однако формирует темы с более низкой когерентностью. Между альтернативными эмбеддингами для ETM различия невелики: модель, обученная с эмбеддингами Navec, несколько превосходит эмбеддинги Word2Vec по разнообразию тем и качеству классификации на MLSUM, однако уступает для Lenta. Некоторой золотой серединой выступают CTM и ProdLDA, следующие за SBMTM и BERTopic по когерентности и F1.

Так, анализ алгоритмов тематического моделирования на размеченных данных позволил обнаружить содержательные различия в результирующих для каждого из инструментов тематических группах: несмотря на то, что BERTopic лучше агрегирует слова-токены, что позволяет достигать более когерентных тем, SBMTM превосходит BERTopic и ETM в задаче группировки документов в соответствии с выделенными темами.

*Оценка качества на неразмеченных данных*

Вторым этапом в сравнительной оценке методов тематического моделирования становится их приложение к данным с неизвестным количеством тем. Здесь мы оцениваем качество тематического моделирования на двух типах данных: хештегах, сопутствующих видеопубликациям, и комментариях, оставленных пользователями.

Анализ хештегов представляется перспективным направлением проверки качества инструментов тематического моделирования, поскольку хештеги как тип текстовых данных отличаются большей семантической когерентностью, чем свободные высказывания. За счет того, что при помощи использования хештегов пользователи стремятся привязать свое собственное высказывание к более крупному дискурсу и/или тренду, хештеги зачастую близко друг с другом связаны.

Аналогично BERTopic демонстрирует наилучшую связность и разнообразие кластеров, однако в этом случае мы наблюдаем расхождение в том, насколько хорошо модели справляются с разными типами данных.

Примечательна динамика качества тематических кластеров в зависимости от количества слов, на которых производится оценка когерентности: по мере увеличения количества анализируемых слов в кластере для расчета метрики связности значение связности падает. Этот эффект связан с качеством ранжирования слов в темах внутри самих алгоритмов. Вместе с убыванием значимости слова в теме оно должно терять свою релевантность, а значит – способствовать сокращению метрики связности темы. Устойчиво убывающие графики качества (рис. 1) указывают на хорошее качество ранжирования слов внутри тем, достигаемое за счет c-TF-IDF взвешивания внутри кластеров для BERTopic и методов выделения сообществ в бимодальной сети для SBMTM. Волатильность оценок для ETM (Navec), в свою очередь,

*Таблица 5*

СРАВНЕНИЕ КАЧЕСТВА ТЕМАТИЧЕСКОГО МОДЕЛИРОВАНИЯ НА НЕРАЗМЕЧЕННЫХ ДАННЫХ

| Модель | Датасет | | Время (*мин*) | Кол-во тем | Когерентность | | Разнообразие | | Схожесть |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | UMass | NPMI | IRBO | TD | Pairwise Jaccard |
| BERTopic | TikTok (комм.) | | 5 | 1421 | **-1,308** | **0,0** | **0,999** | **0,78** | **0,000** |
| | TikTok (хешт.) | | 7 | 719 | *-0,635* | 0,005 | *0,999* | *0,912* | *0,000* |
| ETM (Word2Vec) | TikTok (комм.) | | 3 | 73 | -3,11 | 0,061 | 0,036 | 0,014 | 1,0 |
| ETM (Navec) | | | 3 | 38 | -6,487 | -0,036 | 0,903 | 0,421 | 0,067 |
| CTM | TikTok (комм.) | | 10 | 146 | -7,803 | -0,016 | 0,871 | 0,125 | 0,114 |
| | TikTok (хешт.) | | 0,1 | 12 | -9,124 | *0,092* | 0,959 | 0,733 | 0,037 |
| ProdLDA | TikTok (комм.) | | 10 | 55 | -6,237 | -0,021 | 0,856 | 0,193 | 0,106 |
| | TikTok (хешт.) | | 0,08 | 18 | -7,117 | -0,004 | 0,77 | 0,333 | 0,179 |
| LDA | TikTok (хешт.) | | 0,05 | 70 | -2,992 | 0,054 | 0,031 | 0,016 | 0,914 |
| SBMTM | TikTok (комм.) | L0 | 17 | 11 | -7,66 | -0,05 | 1,0* | 1,0* | 0,000* |
| | | L1 | | 2 | -8,086 | 0,006 | 1,0* | 1,0* | 0,000* |
| | TikTok (хешт.) | L0 | 1 | 44 | -11,24 | 0,042 | 1,0* | 1,0* | 0,000* |
| | | L1 | | 10 | -10,81 | -0,029 | 1,0* | 1,0* | 0,000* |
| NMF | TikTok (комм.) | | 14 | 47 | -4,799 | 0,043 | 0,93 | 0,364 | 0,048 |
| | TikTok (хешт.) | | 0,25 | 46 | -7,949 | 0,052 | 0,881 | 0,315 | 0,091 |

*Примечание.* Жирным шрифтом отмечены лучшие результаты для MLSUM, курсивом с подчеркиванием – для Lenta.

* В текущей версии алгоритма SBMTM не реализована возможность пересечения между выделяемыми сообществами. Как следствие, каждое слово может попадать только в один кластер, что приводит к строго разнообразным кластерам.

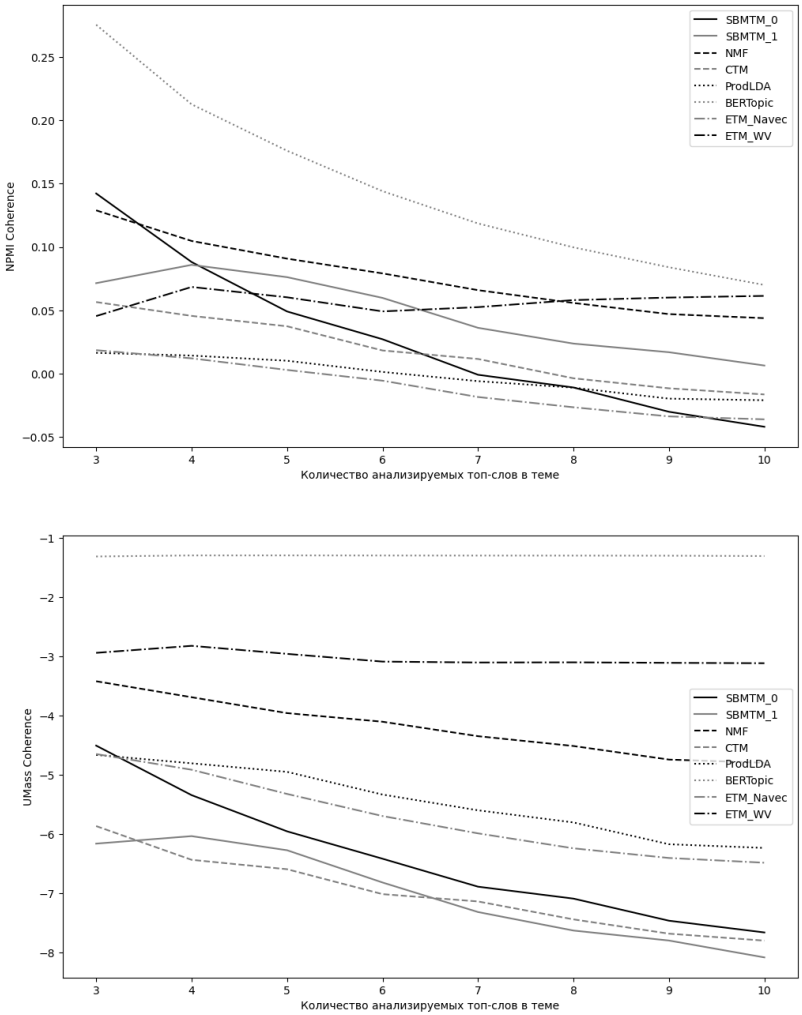



***Рис. 1. Динамика метрик качества тем для комментариев в зависимости от количества включаемых в анализ слов в теме***

демонстрирует худшее качество ранжирования важности слов в тематических кластерах. Тем не менее наблюдаемая динамика может частично объяснять превосходство прочих моделей перед BERTopic в качестве классификации на размеченных данных: модель отлично выбирает наиболее значимые слова в теме, однако далее по списку значимости качество тем резко падает, в то время как CTM и ProdLDA сохраняют близкие значения когерентности при увеличении числа анализируемых топ-слов.

Следует обратить внимание на то, что между моделями различается и разброс значений когерентности (рис. 2): если BERTopic склонен создавать темы примерно одного и того же среднего уровня качества, то SBMTM производит больше как очень низко-, так и очень высококогерентных тем.

ETM (Word2Vec) демонстрирует очень высокую волатильность качества в зависимости от числа анализируемых слов. В меньшей степени это характерно для ETM (Navec) и NMF. Наибольшую стабильность демонстрируют CTM и ProdLDA – анализ большего числа слов незначительно меняет распределение когерентности тем, указывая на хорошее и стабильное качество ранжирования слов между темами.

## Ограничения

Среди ограничений настоящего исследования стоит обратить внимание на несколько деталей, связанных с ограничениями вычислительных мощностей, доступных автору при проведении вышеописанных экспериментов. Во-первых, ввиду ограничений доступных объемов оперативной памяти CPU и GPU (50 и 16 Гб соответственно) для экспериментальной оценки качества алгоритмов тематического моделирования были выбраны сравнительно небольшие массивы данных. Во-вторых, используемый массив неразмеченных коротких текстов из TikTok по той же причине ограничен тематически, однако дополнительным ограничением

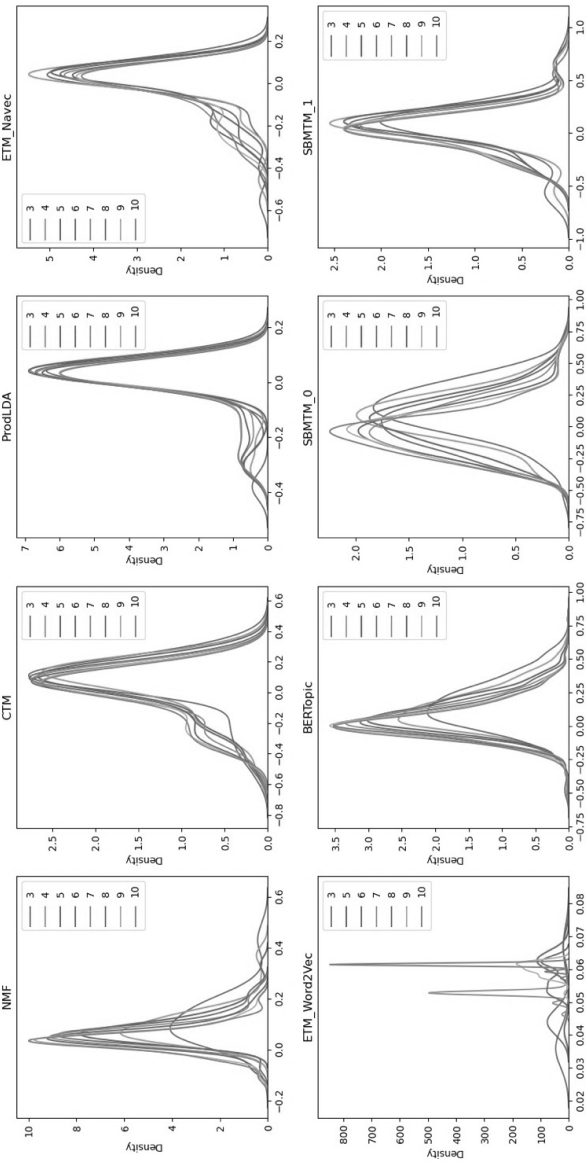



***Рис. 2. Распределение метрик качества тем для комментариев в зависимости от количества включаемых в анализ слов в теме***

доступа к данным комментариев в TikTok являются внутренние лимиты на автоматизированный сбор данных в социальной сети: в целях противодействия веб-скрейпингу TikTok ограничивает объем выдачи как видео, так и комментариев к ним по заданному запросу. Другими словами, в собранном массиве данных присутствуют не все видео урбанистической тематики, а также хоть и абсолютное большинство, но не все комментарии к полученным видео. Наконец, ограничения доступной памяти и недоступность параллельных вычислений в нашем случае не позволяют расширить сетку значений гиперпараметров, рассматриваемую при оптимизации выдачи каждого из анализируемых алгоритмов, что позволяет предположить, что избранное решение с наилучшим результатом для каждой из моделей не является абсолютным максимумом ее возможного качества.

В качестве методологического ограничения проделанной работы можно выделить разницу в источнике сравниваемых размеченных и неразмеченных массивов данных. Оба размеченных массива содержат тексты новостных публикаций, в то время как неразмеченные данные представляют комментарии пользователей в социальной сети TikTok. Поскольку новостные публикации зачастую имеют тематическую категоризацию на сайтах изданий, они представляют собой крупный и доступный объем данных для тематического моделирования, и многие базы данных для тематического моделирования опираются именно на них. Выбор урбанистической тематики в качестве фильтра позволяет ожидать пересечения в темах обсуждения в комментариях в TikTok с новостными публикациями, освещающими проблемы общества, вопросы транспорта, и территориальными категориями «Москва» и «Московская область» в массиве MLSUM. Тем не менее тексты комментариев могут отличаться лексически, содержать больше сленговых выражений, ненормативной лексики, именованных сущностей, что может затруднять работу моделей, основанных на предобученных словарных эмбеддингах. Так, сравниваемые

массивы различаются не только наличием разметки, но и структурой данных.

Наконец, в рамках настоящего исследования оптимальное число тем подбиралось по сетке значений вместе с прочими гиперпараметрами для моделей, требующих указания числа тем в качестве гиперпараметра, на основании увеличения метрики когерентности. Использование автоматических методов оптимизации тематических моделей, таких как энтропийные тематические модели [19], и регуляризаторов для итеративного сокращения числа тем [20] может помочь быстрее находить оптимальное количество тем и улучшить результаты тематического моделирования.

## Заключение

В рамках настоящей работы представлен сравнительный анализ качества неконвенциональных алгоритмов тематического моделирования на четырех разных корпусах коротких текстов. В сравнение вошли методы, дополняющие традиционно используемый LDA семантической информацией при помощи предобученных статических (ETM) и контекстуальных (CTM) эмбеддингов, расширяющие формальное определение LDA за счет внедрения иерархического подхода к моделированию тем в сетевом представлении документов и слов (SBMTM), или product-of-experts структуры (ProdLDA), а также реформирующие задачу тематического моделирования как задачу кластеризации на векторных представлениях слов и документов, произведенных на выходе кодировщика BERT-модели (BERTopic) или матричного разложения (NMF). Избранные алгоритмы сравнивались в задачах классификации и тематического моделирования, для этого были использованы наборы размеченных по темам новостных публикаций и неразмеченных хештегов и комментариев к урбанистическим видео в TikTok.

Анализ качества тематического моделирования по метрикам когерентности и разнообразия тем для всех четырех корпусов

текстов указывает на превосходство BERTopic в задаче создания когерентных тем. Это достигается за счет c-TF-IDF – алгоритма взвешивания слов по их значимости внутри темы, прямые аналоги которого отсутствуют у альтернатив. Однако мы обнаруживаем, что, несмотря на то что темы, выделяемые BERTopic, имеют более высокие показатели когерентности в среднем по модели, SBMTM на нулевом уровне иерархии формирует больше тем с высокими значениями когерентности, которые при усреднении уравновешиваются темами с низкими метриками качества. Вероятно, эта характеристика позволяет SBMTM значительно превосходить BERTopic в задаче классификации документов. Совместное моделирование сообществ для документов и слов в бимодальной сети позволяет SBMTM лучше кластеризировать документы, однако качество предсказания также указывает на то, что темы, производимые SBMTM, лучше подлежат генерализации. Иерархическая структура алгоритма позволяет эффективно привнести гетерогенность в набор производимых тем, в то время как BERTopic склонен фокусироваться исключительно на локальных семантических паттернах.

Проведенные эксперименты подтверждают перспективность использования альтернативных LDA-методов тематического моделирования на коротких текстах и, в частности, сетевого подхода к представлению текстовых данных в задаче тематического моделирования. Мы обнаруживаем, что метод иерархического блокмоделирования превосходит методы, основанные на словарных эмбеддингах в задаче классификации, ранее продемонстрировавших наиболее высокое среди сродных альтернатив качество для классификации документов на англоязычных датасетах [15]PLSA and LDA. Тем не менее следует отметить, что SBMTM уступает альтернативам по производительности, а также качеству ранжирования слов в темах по значимости, на что указывает высокий разброс значений и заметные различия распределения когерентности по темам – в зависимости от числа анализируемых слов.

На трех задействованных в анализе базах текстов: MLSUM, Lenta v1.1+ и TikTok наиболее стабильное качество демонстрируют CTM и ProdLDA. Несмотря на то, что во всех экспериментах по анализу качества тем лидирует BERTopic, высокое значение когерентности в этой модели нестабильно и резко падает с увеличением числа анализируемых топ-слов в теме; BERTopic также хуже справляется с классификацией документов на основании произведенных моделью тем. Лидер классификации, SBMTM, в свою очередь отличается нестабильностью качества тем. Мы отмечаем потенциал применения методов сетевого анализа для тематического моделирования на коротких текстах, однако заключаем, что методы, основанные на создании более сложных внутренних представлений текста в модели, демонстрируют более высокое качество за счет лучшей репрезентации близости между текстами и результирующими темами.

Наконец, в то время как разнообразия сравниваемых размеченных баз данных недостаточно для формулировки формализованных рекомендаций к выбору методов тематического моделирования на основе характеристик доступных данных, представляется возможным обобщение наблюдений о предпочтительности той или иной модели в зависимости от исследовательской задачи. Так, если тематическое моделирование в рамках исследования предусмотрено с целью различения документов, то следует рассмотреть модели, группирующие документы по тематической близости (в рамках настоящего исследования это SBMTM). Для данных с предположительно сложным тематическим составом (особенно если предполагается, что в данных могут присутствовать малые по частоте встречаемости, но значимые для задачи исследования темы) могут быть очень полезны инструменты тематического моделирования с использованием контекстуальных эмбеддингов (здесь CTM, BERTopic). Однако мы подчеркиваем преимущество CTM с точки зрения интерпретации за счет выделения меньшего количества тем, что делает их доступными для ручной проверки

на связность. Тем не менее многие альтернативы LDA могут быть затратны с точки зрения вычислительных ресурсов, поскольку имеют меньше альтернативных эффективных реализаций и не всегда доступны для параллельных вычислений: в случае, если предполагается сравнительная тематическая однородность, не ожидается значимых различий в значении слов в зависимости от контекста и порядок слов в тексте не является принципиальным, LDA может оказаться предпочтительнее моделей, обращающихся к контексту.

## ЛИТЕРАТУРА

1. *Brookes G., McEnery T.* The utility of topic modelling for discourse studies: A critical evaluation // Discourse Studies. 2019. Vol. 21, № 1. С. 3–21. DOI: 10.1177/1461445618814032.

2. Using topic models for Twitter hashtag recommendation / F. Godin, V. Slavkovikj, W. De Neve [et al.] // Proceedings of the 22nd International Conference on World Wide Web. Rio de Janeiro, Brazil: ACM, 2013. P. 593–596. DOI: 10.1145/2487788.2488002.

3. *Asmussen C.B., Møller C.* Smart literature review: a practical topic modelling approach to exploratory literature review // Journal of Big Data. 2019. Vol. 6, № 1. P. 93. DOI: 10.1186/s40537-019-0255-7. EDN: XBRIWK.

4. On the Globalization of the QAnon Conspiracy Theory Through Telegram / M. Hoseini, P. Melo, F. Benevenuto [et al.] // Proceedings of the 15th ACM Web Science Conference 2023. Austin TX, USA: ACM, 2023. P. 75–85. DOI: 10.1145/3578503.3583603.

5. *Кольцова О.Ю., Маслинский К.А.* Выявление тематической структуры российской блогосферы: автоматические методы анализа текстов // Социология: методология, методы, математическое моделирование (Социология: 4М). 2013. № 36. С. 113–139. EDN: RCFOWJ.

6. *Lyu J.C., Han E.L., Luli G.K.* COVID-19 Vaccine–Related Discussion on Twitter: Topic Modeling and Sentiment Analysis // Journal of Medical Internet Research. 2021. Vol. 23, № 6. P. e24435. DOI: 10.2196/24435.

7. ET-LDA: Joint topic modeling for aligning, analyzing and sensemaking of public events and their Twitter feeds / Y. Hu, A. John, F. Wang [et al.] // Cornwall University [site]. 08.10.2012. URL: https://arxiv.org/abs/1210.2164 (дата обращения: 01.09.2023).

8. Multi-modal event topic model for social event analysis / S. Qian, T. Zhang, C. Xu, J. Shao // IEEE Transactions on Multimedia. 2016. Vol. 18, № 2. P. 233–246. DOI: 10.1109/TMM.2015.2510329.

9. *Zheng Y., Zhang Y.-J., Larochelle H.* Topic Modeling of Multimodal Data: An Autoregressive Approach // 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, OH, USA: IEEE, 2014. P. 1370–1377. DOI: 10.1109/CVPR.2014.178.

10. *Gong Y., Poellabauer C.* Topic Modeling Based Multi-modal Depression Detection // Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge. Mountain View, California, USA: ACM, 2017. P. 69–76. DOI: 10.1145/3133944.3133945.

11. *Бызов А.А.* Интеллектуальный анализ текстов в социальных науках // Социология: методология, методы, математическое моделирование (Социология: 4М).2019. № 49. С. 131–160. EDN: GCIIVL.

12. *Boon-Itt S., Skunkan Y.* Public Perception of the COVID-19 Pandemic on Twitter: Sentiment Analysis and Topic Modeling Study // JMIR Public Health and Surveillance. 2020. Vol. 6, № 4. P. e21978. DOI: 10.2196/21978.

13. *Albalawi R., Yeap T.H., Benyoucef M.* Using topic modeling methods for short-text data: A comparative analysis // Frontiers in artificial intelligence. 2020. Vol. 3. P. 42. DOI: 10.3389/frai.2020.00042.

14. *Hong L., Davison B.D.* Empirical study of topic modeling in Twitter // Proceedings of the First Workshop on Social Media Analytics. Washington, D.C.: ACM, 2010. P. 80–88. DOI: 10.3390/ijerph18126487.

15. Short Text Topic Modeling Techniques, Applications, and Performance: A Survey / Q. Jipeng, Q. Zhenyu, L. Yun [et al.] // IEEE Trans. Knowl. Data Eng. 2022. Vol. 34, № 3. P. 1427–1445. DOI: 10.1109/TKDE.2020.2992485. EDN: ACFCRC.

16. Медиапотребление 2023 // Mediascope [сайт]. [2023]. URL: https://mediascope.net/upload/iblock/226/e7lwh96qizxpwhf1rj2ttfzkwlie8vr8/медиапотребление%202023.pdf (дата обращения: 09.02.2024).

17. *Hofmann T.* Probabilistic latent semantic analysis // Cornwall University [site]. 22.01.2013. URL: https://arxiv.org/abs/1301.6705 (дата обращения: 01.09.2023).

18. *Blei D.M., Ng A.Y., Jordan M.I.* Latent dirichlet allocation // Journal of machine learning research. 2003. Vol. 3. P. 993–1022.

19. *Кольцов С.Н.* Применение энтропийного подхода к проблеме выбора числа тем в тематических моделях // Социофизика и социоинженерия'2018: труды второй Всероссийской междисциплинарной конференции. Москва, 23–25 мая 2018 г. М.: Ин-т проблем управления им. В.А. Трапезникова РАН, 2018. С. 235–236. DOI: 10.21883/PJTF.2017.12.44713.16725. EDN: XYERBR.

20. *Потапенко А.А.* Семантические векторные представления текста на основе вероятностного тематического моделирования: дис. … канд. физ.-мат. наук / НИУ ВШЭ. М., 2017. 147 с. EDN: DNXEFS.

21. Digital Trace Data in the Study of Public Opinion: An Indicator of Attention Toward Politics Rather Than Political Support / A. Jungherr, H. Schoen, O. Posegga, P. Jürgens // Social Science Computer Review. 2017. Vol. 35, № 3. P. 336–356. DOI: 10.1177/0894439316631043.

22. *Ahuja A., Wei W., Carley K.M.* Topic modeling in large scale social network data // SSRN electronic journal. January 2015. DOI: 10.2139/ssrn.2720333.

23. Using hashtag graph-based topic model to connect semantically-related words without co-occurrence in microblogs / Y. Wang, J. Liu, Y. Huang, X. Feng // IEEE Transactions on Knowledge and Data Engineering. 2016. Vol. 28, № 7. P. 1919–1933. DOI: 10.1109/TKDE.2016.2531661.

24. The author-topic model for authors and documents / M. Rosen-Zvi, T. Griffiths, M. Steyvers, P. Smyth // Cornwall University [site]. 11.01.2012. URL: https://arxiv.org/abs/1207.4169 (дата обращения: 01.09.2023).

25. *Phan X.-H., Nguyen L.-M., Horiguchi S.* Learning to classify short and sparse text & web with hidden topics from large-scale data collections // Proceedings of the 17th international conference on World Wide Web. Beijing, China: ACM, 2008. P. 91–100. DOI: 10.1145/1367497.1367510.

26. *Gerlach M., Peixoto T.P., Altmann E.G.* A network approach to topic models // Sci. Adv. 2018. Vol. 4, № 7. P. eaaq1360. DOI: 10.1126/sciadv.aaq1360.

27. Mixed Membership Stochastic Blockmodels / E.M. Airoldi, D. Blei, S. Fienberg, E. Xing // Advances in Neural Information Processing Systems. Vancouver, British Columbia, Canada: Curran Associates Inc., 2008. P. 33–40.

28. *Коршунов А., Гомзин А.* Тематическое моделирование текстов на естественном языке // Труды Института системного программирования РАН. 2012. Т. 23. С. 215–244. DOI: 10.15514/ISPRAS-2012-23-13. EDN: PLUXDR.

29. *Grootendorst M.* BERTopic: Neural topic modeling with a class-based TF-IDF procedure // Cornwall University [site]. 11.03.2022. URL: https://arxiv.org/abs/2203.05794 (дата обращения: 01.09.2023).

30. Attention is All you Need / A. Vaswani, N. Shazeer, N. Parmar [et al.] // Advances in Neural Information Processing Systems. Long Beach, CA, USA: Curran Associates Inc., 2017. P. 5998–6008.

31. Topic modeling algorithms and applications: A survey / A. Abdelrazek, Y. Eid, E. Gawish [et al.] // Information Systems. 2022. Vol. 112. P. 102131. DOI: 10.1016/j.is.2022.102131. EDN: WLYLKR.

32. *Lee D., Seung H.S.* Algorithms for Non-negative Matrix Factorization // Advances in Neural Information Processing Systems. Denver, CO, USA: MIT Press, 2000. P. 556–562.

33. *Dieng A.B., Ruiz F.J.R., Blei D.M.* Topic Modeling in Embedding Spaces // Transactions of the Association for Computational Linguistics. 2020. Vol. 8. P. 439–453. DOI: 10.1162/tacl_a_00325.

34. *Srivastava A., Sutton C.* Autoencoding Variational Inference for Topic Models // Cornwall University [site]. 04.03.2017. URL: https://arxiv.org/abs/1703.01488 (дата обращения: 01.09.2023).

35. Cross-lingual Contextualized Topic Models with Zero-shot Learning / F. Bianchi, S. Terragni, D. Hovy [et al.] // Proceedings of the 16th Conference of the

European Chapter of the Association for Computational Linguistics. April 19–23, 2021 / Ed. by P. Merlo, J. Tiedemann, R. Tsarfaty. Potsdam, Germany: Association for Computational Linguistics, 2021. P. 1676–1683. DOI: 10.18653/v1/2021.eacl-main.143.

36. *Кукушкин А.* Navec – компактные эмбеддинги для русского языка // Проект Natasha – набор Python-библиотек для обработки текстов на естественном русском языке [сайт]. 2022. URL: https://natasha.github.io/navec/ (дата обращения: 05.01.2024).

37. Efficient estimation of word representations in vector space / T. Mikolov, K. Chen, G. Corrado, J. Dean // Cornwall University [site]. 16.01.2013. URL: https://arxiv.org/abs/1301.3781 (дата обращения: 01.09.2023).

38. Distributed representations of words and phrases and their compositionality / T. Mikolov, I. Sutskever, K. Chen [et al.] // Advances in Neural Information Processing Systems. 2013. Vol. 26. P. 3111–3119.

39. *Pennington J., Socher R., Manning C.D.* Glove: Global vectors for word representation // Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha, Qatar: Association for Computational Linguistics, 2014. P. 1532–1543. DOI: 10.3115/v1/D14-1162.

40. *Aletras N., Stevenson M.* Evaluating topic coherence using distributional semantics // Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) – Long Papers. Potsdam, Germany: Association for Computational Linguistics, 2013. P. 13–22.

41. Optimizing Semantic Coherence in Topic Models / D. Mimno, H.M. Wallach, E. Talley [et al.] // Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP). Edinburgh, Scotland, UK: Association for Computational Linguistics, 2011. P. 262–272.

42. *Tan Y., Ou Z.* Topic-weak-correlated Latent Dirichlet allocation // 2010 7th International Symposium on Chinese Spoken Language Processing. Tainan, Taiwan: IEEE, 2010. P. 224–228. DOI: 10.1109/ISCSLP.2010.5684906.

43 *Newman D., Karimi S., Cavedon L.* External Evaluation of Topic Models // ADCS 2009 – Proceedings of the Fourteenth Australasian Document Computing Symposium. Sydney, Australia: University of Sydney, 2011. P. 1–8.

44. MLSUM: The Multilingual Summarization Corpus / T. Scialom, P.-A. Dray, S. Lamprier [et al.] // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). [s. l.]: Association for Computational Linguistics, 2020. P. 8051–8067. DOI: https://doi.org/10.18653/v1/2020.emnlp-main.647.

# Приложение 1

## РАСПРЕДЕЛЕНИЕ КЛАССОВ В РАЗМЕЧЕННЫХ МАССИВАХ ТЕКСТОВ

### РАСПРЕДЕЛЕНИЕ ТЕМ В ВЫБОРКЕ ИЗ MLSUM

| Тема | Количество | Тема | Количество |
|---|---|---|---|
| Общество | 7168 | Москва | 1528 |
| Политика | 5310 | Экономика | 1514 |
| Спорт | 3073 | Московская область | 604 |
| Культура | 2623 | Наука | 506 |
| Происшествия | 1634 | Авто | 72 |

### РАСПРЕДЕЛЕНИЕ ТЕМ В ВЫБОРКЕ ИЗ Lenta.ru v1.1+

| Тема | Количество | Тема | Количество |
|---|---|---|---|
| Политика | 36 093 | Музыка | 7054 |
| Общество | 31 963 | Наука | 6693 |
| Украина | 19 920 | Люди | 6250 |
| Происшествия | 19 212 | Квартира | 4656 |
| Футбол | 14 301 | Преступность | 4652 |
| Госэкономика | 14 277 | ТВ и радио | 3945 |
| Кино | 10 117 | Космос | 3686 |
| Интернет | 8516 | События | 3414 |
| Бизнес | 8018 | Конфликты | 3380 |
| Следствие и суд | 7784 | Соцсети | 3314 |

## Приложение 2

### ОПИСАНИЕ ГИПЕРПАРАМЕТРОВ, ПОДБИРАЕМЫХ В РАМКАХ ОПТИМИЗАЦИИ АЛГОРИТМОВ ТЕМАТИЧЕСКОГО МОДЕЛИРОВАНИЯ

| Модель | Гиперпараметр | Влияние на результат | Рассматриваемые значения |
|---|---|---|---|
| NMF | Количество тем | Задает количество тем, которое будет выделять алгоритм | {10, 20, … 200} |
| | Темп обучения (learning rate) – определяет размер шага оптимизатора при обучении модели | Высокий темп обучения способствует ускорению обучения, однако может мешать определению глобального минимума функции потерь | {0.005, 0.01, ... 0.03} |
| LDA | Количество тем | См. выше | {10, 20, … 200} |
| ETM | Количество тем | См. выше | {10, 20, … 200} |
| | ρ (ро) | Контролирует разреженность эмбеддингов сокращенной размерности в модели. Чем выше это значение, тем более разрежены вектора, отображающие отношения документов, и тем меньше тем выделяется в каждом документе | {0.1, 0.2, 0.5, 1.0} |

*Продолжение прилож. 2*

| Модель | Гиперпараметр | Влияние на результат | Рассматриваемые значения |
|---|---|---|---|
| ETM | Dropout rate – доля нейронов, случайным образом удаляемых из заданного слоя модели в ходе обучения | Значения dropout rate отображают силу регуляризации в модели: чем выше dropout rate, тем сильнее противодействие переобучению при тренировке модели | {0, 0.1, … 0.6} |
| | Темп обучения (learning rate) | См. выше | {0.005, 0.01.. 0.03} |
| ProdLDA и CTM | Количество тем | См. выше | {10, 20, … 200} |
| | Функция активации – функция, преобразующая выходы полносвязных слоев в нейросетевой архитектуре | Функции активации добавляют нелинейность в модель. От выбора функции активации зависят качество и эффективность обучения нейросетевой модели | {ReLU, LeakyReLU, Softplus} |
| | Dropout rate | См. выше | {0, 0.1, … 0.6} |
| | Количество слоев – отображает глубину обучаемой архитектуры VAE | Более глубокие архитектуры способны улавливать более сложные связи в массиве обучающих данных, однако более склонны к переобучению | {1, 2, 3, 4} |

*Окончание прилож. 2*

| Модель | Гиперпараметр | Влияние на результат | Рассматриваемые значения |
|---|---|---|---|
| BERTopic | Минимальный размер темы – минимально допустимый размер кластера при кластеризации эмбеддингов сокращенной размерности | Более низкие значения позволяют лучше улавливать малые темы, однако увеличивают время вычисления | {10, 20, 30, 40, 50} |
| | Размерность сокращенного пространства эмбеддингов перед кластеризацией | Чем больше измерений используется при сокращении размерности эмбеддингов, тем лучше сохраняется структура оригинальных данных, однако это увеличивает время вычисления | {2, 3, 4, 5} |
| | Количество «соседей» – количество ближайших наблюдений, задействуемых при вычислении эмбеддингов сокращенной размерности | Более высокие значения количества соседей стимулируют модель сокращения размерности отображать более глобальную структуру данных, более низкие – локальную | {10, 15, 20, 25, 30} |

***Vashchenko Vasilisa A.,***
*Research Assistant at the International Laboratory for Applied Network Research, HSE University, Moscow, Russia,* vvashchenko@hse.ru

**Topic modeling for short texts: comparative analysis of algorithms**

The steady increase in the popularity of social media as a means of communication actualizes methodological issues related to processing of short texts with less semantic context than large corpora, which are widely used for training and testing machine learning models for textual data. Topic modeling, an unsupervised machine learning technique aimed at aggregating texts into topic clusters, has many academic and practical applications where information on true groupings of texts is not available. However, the performance of topic modeling algorithms may be limited by requirement of a sufficient semantic context for a high-quality numerical representation of a unit of text, which may not be derived effectively from a short document. This paper is dedicated to discussing 6 different approaches to topic modeling, comparing their performance on a set of Russian-language comments on TikTok and formally evaluating their performance based on speed and coherence of the resulting topics.
*Keywords:* topic modeling, analysis of textual data, blockmodeling, applied network analysis, social media analysis, transformer models

**References**

1. Brookes G., McEnery T. The utility of topic modelling for discourse studies: A critical evaluation, *Discourse Studies*, 2019, vol. 21, no. 1, p. 3–21. DOI: 10.1177/1461445618814032.
2. Godin F., Slavkoviki V., De Neve W. et al. Using topic models for Twitter hashtag recommendation, *Proceedings of the 22nd International Conference on World Wide Web*. ACM, Rio de Janeiro, 2013, p. 593-596. DOI: 10.1145/2487788.2488002.
3. Asmussen C.B., Møller C. Smart literature review: a practical topic modelling approach to exploratory literature review, *Journal of Big Data*, 2019, vol. 6, no 1, p. 93. DOI: 10.1186/s40537-019-0255-7.
4. Hoseini M., Melo P., Benevenuto F. et al. On the Globalization of the QAnon Conspiracy Theory Through Telegram, *Proceedings of the 15th ACM Web Science Conference*. ACM: Austin, 2023, p. 75-85. DOI: 10.1145/3578503.3583603.

5. Koltsova O., Maslinsky K. Revealing the thematic structure of the Russian blogosphere: automatic methods of text analysis (in Russian), *Sotsiologiya 4M (Sociology: methodology, methods, mathematical modeling)*, 2013, no. 36, p. 113-139.
6. Lyu J.C., Han E.L., Luli G.K. COVID-19 Vaccine–Related Discussion on Twitter: Topic Modeling and Sentiment Analysis, *Journal of Medical Internet Research*, 2021, vol. 23, no. 6, p. e24435. DOI: 10.2196/24435.
7. Hu Y., John A., Wang F., et al. ET-LDA: Joint topic modeling for aligning, analyzing and sensemaking of public events and their Twitter feeds, *Cornwall University* [site], 08.10.2012. URL: https://arxiv.org/abs/1210.2164 (date of access: 01.09.2023).
8. Qian S., Zhang T., Xu C., Shao J. Multi-modal event topic model for social event analysis, *IEEE Transactions on Multimedia*, 2016, vol. 18, no. 2, p. 233–246. DOI: 10.1109/TMM.2015.2510329.
9. Zheng Y., Zhang Y.-J., Larochelle H. "Topic Modeling of Multimodal Data: An Autoregressive Approach", in: *2014 IEEE Conference on Computer Vision and Pattern Recognition*. Columbus, OH, USA: IEEE, 2014, p. 1370–1377. DOI: 10.1109/CVPR.2014.178.
10. Gong Y., Poellabauer C. "Topic Modeling Based Multi-modal Depression Detection", in: *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*. Mountain View, California, USA: ACM, 2017, p. 69–76. DOI: 10.1145/3133944.3133945.
11. Byzov A. Text mining in social sciences (in Russian), *Sotsiologiya 4M (Sociology: methodology, methods, mathematical modeling)*, 2019, no. 49, p. 131-160.
12. Boon-Itt S., Skunkan Y. Public Perception of the COVID-19 Pandemic on Twitter: Sentiment Analysis and Topic Modeling Study, *JMIR Public Health and Surveillance*, 2020, vol. 6, no. 4, p. e21978. DOI: 10.2196/21978.
13. Albalawi R., Yeap T.H., Benyoucef M. Using topic modeling methods for short-text data: A comparative analysis, *Frontiers in artificial intelligence*, 2020, vol. 3, p. 42. DOI: 10.3389/frai.2020.00042.
14. Hong L., Davison B. D. "Empirical study of topic modeling in Twitter", in: *Proceedings of the First Workshop on Social Media Analytics*. Washington, D.C.: ACM, 2010, p. 80–88. DOI: 10.3390/ijerph18126487.
15. Jipeng Q., Zhenyu Q., Yun L., et al. Short Text Topic Modeling Techniques, Applications, and Performance: A Survey, *IEEE Trans. Knowl. Data Eng*. 2022, vol. 34, no. 3, p. 1427–1445. DOI: 10.1109/TKDE.2020.2992485.

16. Mediaconsumption 2023 (in Russian), *Mediascope* [site], 2023. URL: https://mediascope.net/upload/iblock/226/e7lwh96qizxpwhf1rj2ttfzkwlie8vr8/медиапотребление%202023.pdf (date of access: 09.02.2024).
17. Hofmann T. Probabilistic latent semantic analysis, *Cornwall University* [site], 22.01.2013. URL: https://arxiv.org/abs/1301.6705 (date of access: 01.09.2023).
18. Blei D.M., Ng A.Y., Jordan M.I. Latent dirichlet allocation, *Journal of machine learning research*, 2003, vol. 3, p. 993–1022.
19. Koltsov S. "Applying the entropy approach to the problem of choosing the number of topics in topic models", in: *Sociophysics and Socioengineering 2018: Proceedings of the Second All-Russian Cross-disciplinary Conference*. Moscow: V.A. Trapeznikov Institute of Control Sciences of the Russian Academy of Sciences, 2018, p. 235–236. DOI: 10.21883/PJTF.2017.12.44713.16725.
20. Potapenko A. *Sematic vector embeddings of text based on probabilistic topic modeling* (in Russian) [Doct. Diss.]. Moscow: Higher School of Economics, 2017, 147 p.
21. Jungherr A., Schoen H., Posegga O., Jürgens P. Digital Trace Data in the Study of Public Opinion: An Indicator of Attention Toward Politics Rather Than Political Support, *Social Science Computer Review*, 2017, vol. 35, no. 3. p. 336–356. DOI: 10.1177/0894439316631043.
22. Ahuja A., Wei W., Carley K.M. Topic modeling in large scale social network data, *SSRN electronic journal*, January 2015. DOI: 10.2139/ssrn.2720333.
23. Wang Y., Liu J., Huang Y., Feng X. Using hashtag graph-based topic model to connect semantically-related words without co-occurrence in microblogs, *IEEE Transactions on Knowledge and Data Engineering*, 2016, vol. 28, no. 7, p. 1919–1933. DOI: 10.1109/TKDE.2016.2531661.
24. Rosen-Zvi M., Griffiths T., Steyvers M., Smyth P. The author-topic model for authors and documents, *Cornwall University* [site], 11.01.2012. URL: https://arxiv.org/abs/1207.4169 (date of access: 01.09.2023).
25. Phan X.-H., Nguyen L.-M., Horiguchi S. "Learning to classify short and sparse text & web with hidden topics from large-scale data collections", in: *Proceedings of the 17th international conference on World Wide Web*. Beijing, China: ACM, 2008, p. 91–100. DOI: 10.1145/1367497.1367510.
26. Gerlach M., Peixoto T.P., Altmann E.G. A network approach to topic models, *Sci. Adv*, 2018, vol. 4, no. 7, p. eaaq1360. DOI: 10.1126/sciadv.aaq1360.

27. Airoldi E.M., Blei D., Fienberg S., Xing E. “Mixed Membership Stochastic Blockmodels”, in: *Advances in Neural Information Processing Systems*. Vancouver, British Columbia, Canada: Curran Associates Inc., 2008, p. 33–40.
28. Korshunov A., Gomzin A. Topic modelling for natural language texts (in Russian), *Proceedings of the Institute for System Programming of the Russian Academy of Sciences*, 2012, vol. 34, p. 215-244. DOI: 10.15514/ISPRAS-2012-23-13.
29. Grootendorst M. BERTopic: Neural topic modeling with a class-based TF-IDF procedure, *Cornwall University* [site], 11.03.2022. URL: https://arxiv.org/abs/2203.05794 (date of access: 01.09.2023).
30. Vaswani A., Shazeer N., Parmar N., et al. “Attention is All you Need”, in: *Advances in Neural Information Processing Systems*. Long Beach, CA, USA: Curran Associates Inc., 2017, p. 5998–6008.
31. Abdelrazek A., Eid Y., Gawish E., et al. Topic modeling algorithms and applications: A survey, *Information Systems*, 2022, vol. 112. p. 102131. DOI: 10.1016/j.is.2022.102131.
32. Lee D., Seung H.S. “Algorithms for Non-negative Matrix Factorization”, in: *Advances in Neural Information Processing Systems*. Denver, CO, USA: MIT Press, 2000, p. 556–562.
33. Dieng A.B., Ruiz F.J.R., Blei D.M. Topic Modeling in Embedding Spaces, *Transactions of the Association for Computational Linguistics*, 2020, vol. 8, p. 439–453. DOI: 10.1162/tacl_a_00325.
34. Srivastava A., Sutton C. Autoencoding Variational Inference for Topic Models, *Cornwall University* [site], 04.03.2017. URL: https://arxiv.org/abs/1703.01488 (date of access: 01.09.2023).
35. Bianchi F., Terragni S., Hovy D., et al. “Cross-lingual Contextualized Topic Models with Zero-shot Learning”, in: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, ed. by P. Merlo, J. Tiedemann, R. Tsarfaty. Potsdam, Germany: Association for Computational Linguistics, 2021, p. 1676–1683. DOI: 10.18653/v1/2021.eacl-main.143.
36. Kukushkin A. Navec – compact embeddings for the Russian language (in Russian), *Project Natasha – an array of Python libraries for text processing in natural Russian language* (in Russian) [site], 2022. URL: https://natasha.github.io/navec/ (date of access: 05.01.2024).

37. Mikolov T., Chen K., Corrado G., Dean J. Efficient estimation of word representations in vector space, *Cornwall University* [site], 16.01.2013. URL: https://arxiv.org/abs/1301.3781 (date of access: 01.09.2023).
38. Mikolov T., Sutskever I., Chen K, et al. Distributed representations of words and phrases and their compositionality, *Advances in Neural Information Processing Systems*, 2013, vol. 26, p. 3111–3119.
39. Pennington J., Socher R., Manning C.D. "Glove: Global vectors for word representation", in: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, 2014, p. 1532–1543. DOI: 10.3115/v1/D14-1162.
40. Aletras N., Stevenson M. "Evaluating topic coherence using distributional semantics", in: *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) – Long Papers*. Potsdam, Germany: Association for Computational Linguistics, 2013, p. 13–22.
41. Mimno D., Wallach H.M., Talley E., et al. "Optimizing Semantic Coherence in Topic Models", in: *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Edinburgh, Scotland, UK: Association for Computational Linguistics, 2011, p. 262–272.
42. Tan Y., Ou Z. "Topic-weak-correlated Latent Dirichlet allocation", in: *2010 7th International Symposium on Chinese Spoken Language Processing*. Tainan: IEEE, 2010, p. 224–228. DOI: 10.1109/ISCSLP.2010.5684906.
43. Newman D., Karimi S., Cavedon L. "External Evaluation of Topic Models", in: *ADCS 2009 – Proceedings of the Fourteenth Australasian Document Computing Symposium*. Sydney: University of Sydney, 2011, p. 1–8.
44. Scialom T., Dray P.-A., Lamprier S., et al. "MLSUM: The Multilingual Summarization Corpus", in: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2020, p. 8051–8067. DOI: https://doi.org/10.18653/v1/2020.emnlp-main.647.

DOI: 10.19181/4m.2023.32.1.3
EDN: GXCDWB

А.В. Ким
(*Москва*)

# КАЧЕСТВЕННЫЙ СЕТЕВОЙ АНАЛИЗ НА ПРАКТИКЕ: СРАВНЕНИЕ СПОСОБОВ ПОСТРОЕНИЯ СЕТЕВЫХ КАРТ[1]

Данная статья посвящена применению качественного сетевого анализа на примере изучения социальных взаимодействий молодых родителей в миграции. Представлено концептуальное описание качественного сетевого анализа, особенности визуализации сети, дизайн качественного сетевого исследования, сравнение двух подходов к построению сетевых карт и пример их анализа. Качественный сетевой анализ направлен на изучение отношений в сети и состоит из интерпретативного и структурного компонентов, которые реализуются посредством проведения интервью и построения сетевой карты. Два подхода к построению сетевых карт отличаются тем, что в одном случае сетевые карты строятся исследователем на основе интервью, а в другом случае построение карты делегируется самому информанту. Первый способ лимитирован данными интервью

**Арюна Витальевна Ким** – младший научный сотрудник Международной лаборатории прикладного сетевого анализа, Национальный исследовательский университет «Высшая школа экономики», Москва, Россия. Email: aryunakimm@gmail.com

[1] Статья подготовлена в ходе проведения исследования в рамках Программы фундаментальных исследований Национального исследовательского университета «Высшая школа экономики» (НИУ ВШЭ).

Благодарю Е.Ю. Рождественскую за ценные советы при планировании исследования, а также редакцию журнала за ценные замечания и комментарии по данной статье.

и является их структурной визуализацией. Второй способ построения сетевой карты самим информантом создает дополнительные ресурсы для интерпретации – контрастирующие нарративы и визуальное структурирование взаимодействий информанта.

*Ключевые слова:* качественный сетевой анализ, сетевой анализ, визуализация сети, сетевые карты, качественные методы, родительство, миграция, российские мигранты

## Введение

Сетевые исследования в социальных науках основываются на «сетевом подходе», понимаемом как «комплекс теоретико-методологических направлений, объединенных использованием понятия сети для объяснения социальных явлений» [1, с. 4]. Сетевые исследования реализуют разнообразные комбинации способов сбора и анализа данных, однако использование качественных методов в сетевых исследованиях становится все более популярным в социальных науках [2]. Несмотря на то, что качественные методы всегда использовались в сетевых исследованиях, современная практика показывает, что есть необходимость в переосмыслении использования качественных методов в сетевых исследованиях [3, 4]. Качественный сетевой анализ (далее КСА) основан на применении качественных методов сбора и анализа данных и сетевом анализе. Однако позиция КСА является неоднозначной, так как одни исследователи убеждены в невозможности обозначения метода независимым, что КСА возможен только при смешивании методов [5], другие считают, что данный метод может быть определен как обособленный методологический подход [2, 3, 4, 6], поскольку в сравнении с количественным сетевым анализом, который направлен на выявление социальных структур, КСА способен раскрыть социальные структуры «изнутри», определив восприятие социальных отношений внутри сети.

Как будет показано ниже, КСА является независимым методом в социальных науках, направленным на изучение восприятия и понимания сетевых отношений [7]. Как и в количественном сетевом анализе, у него есть визуальный компонент в виде сетевой карты (далее СК), строящейся на основе эгосети, или сети одного индивида [8, 9]. На СК представлены взаимодействия одного индиида в виде окружающих его акторов разной близости, взаимодействий с ним и между собой. Существуют разные подходы к построению СК, однако еще не сравнивались СК, построенные исследователем и информантом [9, 10, 11], это будет сделано в данной статье. Статья посвящена двум исследовательским вопросам.

1. Чем отличаются способы построения СК, где в одном случае СК составляется исследователем, а в другом случае информант самостоятельно визуализирует свои контакты в СК?

2. Как анализировать СК, которые создаются в ходе интервью указанными двумя способами (исследователем и информантом)?

На примере изучения изменения сетей взаимодействия молодых родителей в условиях миграции в статье показано применение КСА. Нами рассматриваются российские мигранты, у которых в новой стране родился первый ребенок. Сами мигранты самоидентифицируют себя как «релоканты», однако поскольку релокация понимается как перевод сотрудников на новое место жительства, связанный с деловыми целями компаний [12, 13], россияне, переехавшие вслед за работодателем, также относятся к более широкой категории мигрантов. Данный эмпирический объект интересен по нескольким аспектам. С точки зрения актуальной повестки у мигрантов меняются связи и отношения в период адаптации на новом месте, а также в связи с первым опытом родительства их взаимодействия приобретают новый фокус. С точки зрения методологии КСА позволяет структурно выделить меняющиеся социальные круги отношений, новые смыслы этих отношений и стратегию адаптации к новой жизни в родительстве и миграции. КСА способен отобразить то, как создаются и поддерживаются

отношения в сетях мигрантов и через привязанный нарратив охарактеризовать эти отношения ввиду изменений в социальном кругу до и после рождения ребенка и переезда.

В начале статьи представлено концептуальное описание КСА как методологического подхода к изучению восприятия отношений в сети. Далее рассмотрены особенности визуализации качественных сетевых данных. Затем предложен дизайн качественного сетевого исследования по теме изменения сети взаимодействий в контексте родительства и миграции, приводятся примеры таких изменений. После этого сравниваются два подхода к построению СК, где в одном случае СК строятся исследователем после интервью, а в другом построение СК предлагается самому информанту, и демонстрируется пример анализа полученных сетей взаимодействий. Статья завершается выводами относительно апробации КСА и сравнения подходов к построению СК.

## *Качественный сетевой анализ как методологический подход к изучению восприятия отношений в сети*

КСА может иметь разные варианты названий: качественный эгосетевой анализ, качественный структурный анализ или качественный подход в сетевом анализе и др. Предметом данного методологического подхода является изучение восприятия сетевых отношений: что означают эти отношения, как они проявляются и как воспринимаются в контексте сети в целом и в сравнении с другими отношениями. Существует разнородная практика использования КСА в стратегии смешанных методов, а также как отдельного методологического подхода [6, 7]. Поскольку в русскоязычном сегменте нет примера применения КСА на практике, то это и будет реализовано в данной статье.

КСА определяется как методологический подход, выявляющий восприятие сетевых отношений в персональных сетях

взаимодействий. Сетевые отношения рассматриваются исходя из двух компонентов: структурной позиции в сети и смыслов отношений. Структурная позиция актора в сети обозначает место актора в структуре сетевого взаимодействия [14]. Смыслы отношений в контексте коммуникативно-сетевого взаимодействия понимаются нами как продукт коммуникативных событий, в ходе которых формируются и воспроизводятся отношения акторов. Ключевой особенностью КСА является возможность включения двух компонентов одновременно – и смыслов в коммуникативно-сетевом взаимодействии, и структурной позиции, что позволяет синтезировать структуру и интерпретацию сетевых отношений. Смыслы отношений формируются из данных интервью, а структурная визуализация сети строится при помощи СК.

Данный методологический подход состоит из интерпретативной и структурной составляющих в виде интерпретации смыслов относительно коммуникации в сетевых отношениях, а также положений сетевых отношений на СК. На этапе сбора данных применяются качественные методы, такие как наблюдение, интервью и анализ документов, а также собираются сетевые данные, которые могут быть визуализированы на СК [15]. Качественное сетевое исследование состоит из следующих этапов: разработка дизайна исследования, сбор данных качественными методами, сетевая визуализация, анализ данных, интерпретация данных и формулировка выводов.

*Особенности визуализации качественных сетевых данных*

Качественные сетевые исследования могут быть направлены на изучение эгосетей, в которых сеть строится с точки зрения отдельного актора (эго), где рассматриваются взаимодействия между эго и другими акторами, а также все связи между другими акторами. А. Герц определяет эгосети как «связи между одним субъектом (эго) и другими субъектами (альтерами) в его или ее непосредственном соседстве внутри сети, а также связи между

этими субъектами (альтер-альтер)» [8, s. 133]. Для визуализации отношений в эгосети строится СК взаимодействий, которая отражает взаимодействия эго с альтерами и альтеров между собой.

Практика КСА породила два вида визуализаций на основе качественных сетевых данных: сетевые изображения и СК [9]. Сетевые изображения создаются информантом в виде свободных рисунков без каких-либо спецификаций, тогда как СК структурированы по секторам. М. Гампер разделяет СК на три вида: СК, построенная на концентрических кругах (1), СК, дополненная секторами или дополнительными СК (2), цифровая СК (3) [10]. В первом типе СК информант либо рисует отношения на бумаге, либо ему предоставляется карта сети с концентрическими кругами, в которой нужно расположить контакты в зависимости от эмоциональной близости [16], затем СК описываются нарративной интерпретацией отношений. Второй тип СК был выделен в работе Ф. Страуса [9] и обозначен как более сложный вариант первого типа визуализации, поскольку дополняется секторами различных сфер жизни (например, работа, семья, друзья), или создается несколько СК, содержащих различные аспекты отношений (важность, близость, поддержка). Особенность третьего типа СК в том, что информант может построить цифровую СК в компьютерной программе, например в VennMaker[1], где контакты размещаются ближе или дальше от центра, в зависимости от их доступности, и размер актора определяется его важностью.

В другом исследовании Л. Йесперсен сравнивает три способа сбора сетевых данных в качественном сетевом исследовании, такие как совместное сетевое картографирование (1), интервью с СК (2), визуальный сетевой опрос (3) [11]. Первый способ подразумевает совместное создание и интерпретацию СК, где участники обсуждают, кто влияет на определенную проблему взаимодействий

[1] Программа для визуализации и анализа эгосетевых данных VennMaker: URL: https://www.vennmaker.com/?lang=en (дата обращения: 27.05.2024).

и размышляют об отношениях, которые они считают важными, сложными или нуждающимися в улучшении. Второй способ – это использование СК в рамках интервью, он позволяет углубленно погрузиться в сетевые отношения: информанты могут поразмышлять о сходствах, различиях и взаимозависимостях между отношениями [16, 17]. Третий способ включает в себя использование визуальных сетевых опросов, которые предполагают создание стандартизированных СК с помощью «социометрического опроса» [18]. Данные собираются с помощью анкетного опроса, который сочетает элементы обычного сетевого опроса с визуальными элементами. Программные пакеты, такие как EgoNet.QF, E-NET, GENSI, Network Canvas и VennMaker, преобразуют ответы в опросе в СК, которые затем предоставляются респондентам для проверки [11].

В случае сбора данных при помощи интервью исследователи либо предлагали своим информантам самостоятельно визуализировать СК своего окружения, либо строили СК за них в контексте интервью. Пример визуализации СК информантами приведен в исследовании социального капитала мигрантов, где исследователи Е. Зоммер и М. Гампер при помощи СК сравнили бизнес-сети на начальном этапе и на момент интервью, наглядно показывая изменения в структуре и в отношениях с коллегами и заказчиками (см. рис. 1 в Приложении) [15]. Другой пример построения СК самими исследователями приведен в исследовании сетей инноваций [19]. На основе полуструктурированных интервью была собрана информация о внутриорганизационных и внеорганизационных отношениях, способствующих инновационному процессу [19]. После сбора данных исследователи визуализировали результаты благодаря самостоятельно разработанной СК (см. рис. 2 в Приложении).

В случае других качественных способов сбора данных, таких как анализ документов или наблюдение, СК строятся исследователем на основе имеющихся у него данных [20; 21] либо визуализация сети отсутствует [22]. Например, в исследовании

эффективного взаимодействия с целью обучения [22], где использовалось наблюдение как метод сбора данных, были приведены описания взаимодействий между учителями и учениками, однако визуализации сетевых отношений не было. В другом отечественном исследовании Д. Мальцевой и С. Моисеева рассматривалась биография Татьяны Заславской, где на основе уже собранных Б. Докторовым интервью строились эгосети в разные временные периоды [21].

Таким образом, на основе описанных подходов к визуализации сетевых данных мы можем теоретически предположить различные сценарии проведения КСА с разными последовательностями.

1. Проведение интервью, затем построение СК исследователем.

2. Проведение интервью, затем построение СК информантом.

3. Построение СК информантом, затем проведение с ним интервью.

4. Проведение основного интервью, затем построение СК исследователем, после чего предлагается совместное обсуждение СК с информантом с получением его комментария.

5. Проведение основного интервью, затем построение СК информантом, после чего предлагается совместное обсуждение СК с информантом с получением его комментария.

Поскольку первый и второй сценарии являются более нормативными, в рамках данной статьи будет изучена разница между подходами, где после интервью СК строится исследователем или информантом. Также будет предложен пример анализа СК в контексте интервью.

## Дизайн качественного сетевого исследования

Примером применения КСА является исследование изменения сетей взаимодействия в контекстах родительства и миграции. В связи с транзитом к родительству отношения с окружающими

людьми могут поменяться после рождения ребенка. По причине миграции в новую страну некоторые связи могут перестать быть нужными, а также могут появиться новые знакомства и отношения в кругу переехавших. Транзит к родительству (transition to parenthood) обозначает период адаптации, который проходят молодые родители первенца в новой роли, и описывается как переходный период нестабильности и внутреннего конфликта по поводу потерь и приобретений, приводящих к реорганизации внутренней жизни [23; 24; 25; 26]. Социальное окружение может полностью поменяться в период транзита к родительству. Несмотря на изученность темы транзита к родительству, условия миграции привносят дополнительную сложность для адаптации в новой роли. Одним из важных аспектов адаптации является поддержание привычных социальных взаимодействий и появление новых социальных связей. Возможно, сети мигрантов, переезжающих на новое место, становятся похожими на тип «дезинтегрированных» сетей без друзей и знакомых в новом для них городе [27]. С другой стороны, включенность в сети соотечественников на новом месте помогает интегрироваться в новой среде [28]. КСА позволит раскрыть восприятие сетевых отношений в целом, а также возможные изменения этих отношений в связи с родительством и миграцией.

Типичный сценарий применения КСА состоит из сбора интервью, после чего следует построение СК информантом. В данном исследовании сбор данных осуществлялся при помощи полуформализованного интервью и СК, в которых информант структурировал свои отношения до и после переезда и родительства. Сбор данных проводился в сентябре – ноябре 2023 года. В исследовании принимали участие женщины, у которых в новой стране родился первый ребенок. Поиск информанток проходил через мигрантские сообщества в Telegram, которые были отобраны по ключевым словам «россияне» / «релоканты» / «мамы» + «Казахстан» / «Армения» / «Грузия».

*Сбор качественных данных*

Гайд интервью включал в себя шесть блоков: обстоятельства переезда, беременность и опыт родов и родительства, отношения между супругами, работа и социальные взаимодействия информанта. Всего собрано 15 интервью – по 5 интервью из стран ближнего зарубежья (Казахстан, Армения, Грузия). В табл. (см. в Приложении) представлены данные о возрасте информантов, способе сбора данных и длительности интервью. Интервью проводились двумя способами: в формате непрерывной беседы по Zoom и посредством голосовых сообщений в Telegram, так как некоторые женщины не могли уделить нужное количество времени для непрерывной беседы. Длительность беседы в Zoom варьируется от 1 часа 2 минут до 2 часов 2 минут, тогда как беседа при помощи голосовых сообщений в Telegram могла длиться в сумме от 1 часа 1 минуты до 1 часа 43 минут.

*Сбор сетевых данных*

После интервью информантки строили СК и отправляли ее фотографию исследователю. Они строили свои СК до и после переезда и рождения ребенка. В инструкции по построению СК также были показаны примеры СК, состоящих из трех концентрированных кругов, на основе которых информантки рисовали свои карты: самый малый круг был обозначен как круг самых близких людей, средний круг означал среднюю степень близости и дальний круг состоял из неблизких людей.

## Изменения сетей взаимодействий в контексте родительства и миграции: кейсы Ольги и Арины

Для демонстрации применения КСА я предлагаю рассмотреть два кейса семей мигрантов, которые переехали в Казахстан. Дан-

ные примеры являются периферийными и пограничными с точки зрения транзита к родительству и адаптации к новому месту жительства, где в одном случае демонстрируется сложный адаптационный процесс, а в другом случае сравнительно легкий опыт.

*Кейс Ольги*

СК построены после интервью самим исследователем. На рис. 1 изображены две СК до и после рождения ребенка и миграции. На СК зафиксированы все упомянутые в интервью люди, также о каждом из них в интервью приводится история развития отношений до и после переезда и рождения ребенка. Исследователь выбирает позицию актора на СК, исходя из описания отношений из интервью.

Слева на рисунке изображена СК до рождения ребенка и миграции, где черным цветом в центре обозначена Ольга, у которой брали интервью, затем зеленым цветом выделены члены семьи, оранжевым – друзья и голубым – коллеги. Справа представлена СК Ольги после рождения ребенка и миграции, где фиолетовым цветом обозначен психолог. Ольга отмечала, что именно после рождения ребенка у нее стали меняться отношения с людьми. После рождения дочери ухудшились отношения со свекровью, хотя до рождения ребенка проблем не было. Также она близко общалась с подругами Ольгой и Мариной, но после рождения ребенка она стала чуть меньше общаться с Ольгой, но появилась новая подруга Арина. Общение с другими своими друзьями и друзьями мужа практически прекратилось, теперь Ольга периодически общается с соседками с детьми из своего дома. При этом Ольга замечает, что если раньше круг общения состоял из людей такого же уровня образования, возраста и дохода, как у нее, то теперь взаимодействия носят несколько случайный характер по этим характеристикам, но есть потребность в общении с такими же молодыми мамами, как она. Ольга нечасто общалась неформально с коллегами, в связи с чем после выхода в декрет она перестала

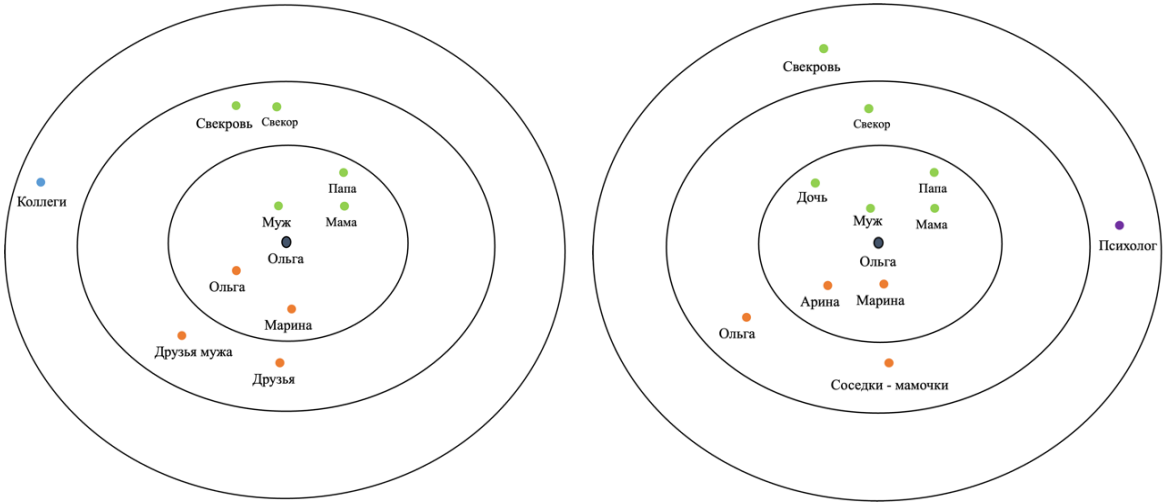



***Рис. 1. Сетевые карты, составленные исследователем: «Сеть Ольги до (слева) и после (справа) рождения ребенка и релокации»***

*Примечание.* Черным цветом в центре обозначена информантка, зеленым цветом выделены члены семьи, оранжевым цветом выделены друзья, голубым цветом обозначены коллеги, фиолетовым цветом обозначен психолог.

общаться с коллегами вовсе. После родов возникла потребность в психологической поддержке, поэтому она регулярно взаимодействует с психологом.

Особенностью сложного транзита к родительству в случае Ольги является послеродовая депрессия. После родов Ольга погрузилась в депрессию. «*Вроде бы, счастье такое случилось, но одновременно появилось ощущение, что вся твоя жизнь рухнула. И ты стоишь на руинах своей жизни, а новую ты еще не построила*» (Ольга, 32 г.). Ольге не хватало поддержки со стороны родителей, близких подруг и поскольку рядом никого не было, Ольга работала с психологом. Спустя месяц после родов к Ольге приехала на выходные подруга Марина из Москвы. Именно после приезда подруги Марины Ольга ощутила тепло и поддержку, начала выздоравливать и ее депрессия прошла. В интервью Ольга подчеркивала, насколько этот кризисный период сблизил ее с подругой Мариной, что теперь их отношения стали крепче, чем раньше.

*Кейс Арины*

СК информантка строила самостоятельно в процессе интервью. На рис. 2 изображены две СК до и после рождения ребенка и миграции. Слева изображена СК до рождения ребенка и миграции, а справа представлена СК Арины после этих событий. СК преобразованы в цифровой вид для улучшения качества (оригинал см. на рис. 3 в Приложении).

Судя по СК, узлов и взаимодействий у Арины стало больше в миграции. Однако в интервью Арина рассказывала, что общения стало меньше после переезда и появления ребенка. Хотя в окружении появились новые знакомства и появились регулярные контакты со старыми знакомыми, которые тоже переехали в Казахстан. До рождения ребенка в кругу самых близких у Арины были муж и две подруги, но после рождения ребенка и релокации с одной подругой она стала общаться реже, но зато познакомилась и сблизилась с новой подругой. Довольно интересно, что со вре-

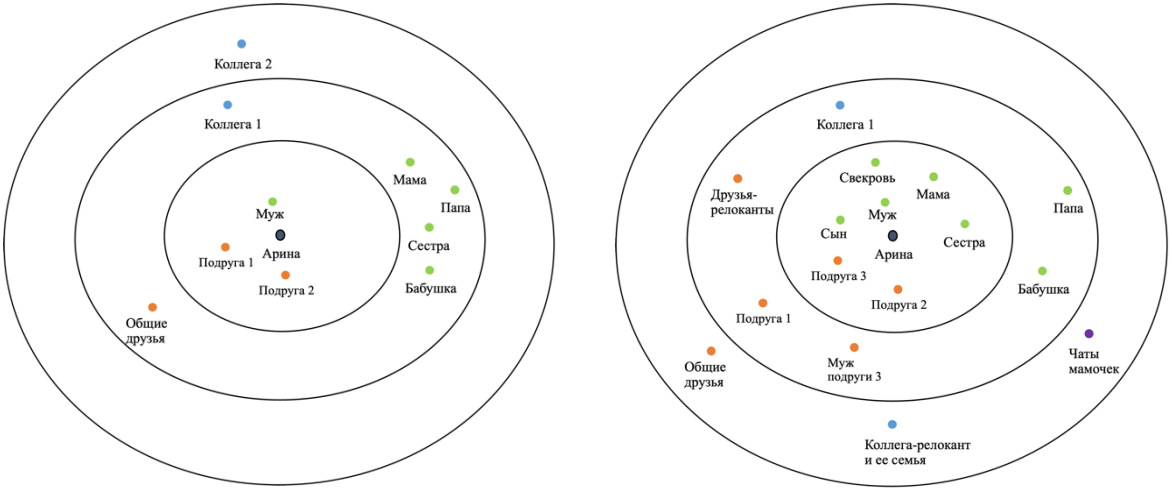



***Рис. 2. Сетевые карты, составленные информантом: «Сеть Арины до (слева) и после (справа) рождения ребенка и релокации»***

*Примечание.* Черным цветом в центре обозначена информантка, зеленым цветом выделены члены семьи, оранжевым цветом выделены друзья, голубым цветом обозначены коллеги, фиолетовым цветом выделено сообщество мам.

менем отношения с мамой, сестрой после рождения ребенка стали более близкими. «*В беременность и после родов я стала больше общаться с мамой и теперь я понимаю, насколько ей, возможно, было трудно и тяжело со мной и моей сестрой. Благодаря материнству я смогла взглянуть на маму не с позиции ребенка, а с позиции такого же взрослого*» (Арина, 27 л.). С мамой мужа отношения стали более близкими, тогда как раньше они редко общались, поэтому ее нет на СК слева.

До переезда Арина с мужем часто общались в компании общих друзей, но после переезда они с друзьями стали реже контактировать. Но некоторые друзья тоже переехали в Казахстан и с ними Арина встречается и общается регулярно. Также Арина вышла из декрета и работает удаленно несколько часов в неделю. Если раньше она общалась с двумя коллегами, то теперь регулярно взаимодействует только с одной. Также в Казахстан переехала коллега Арины со своей семьей: если ранее они часто не общались, то теперь периодически поддерживают связь. Отдельно Арина отметила материнские онлайн-чаты, в которых общается группа девушек, вместе посещавших курсы для беременных, и другие чаты с мамами в Казахстане. Арина часто обращается в эти чаты за советом и поддержкой, так как дети пользователей данных чатов ровесники и все сталкиваются с одними и теми же проблемами.

## Сравнение способов сбора сетевых карт

В данном разделе описано сравнение способов сбора сетевых данных для построения СК. Оба способа протестированы на конкретном примере, с единым гайдом интервью, оба интервью проводились по Zoom, длительность интервью приблизительно одинаковая (1 час 42 минуты с Ольгой и 1 час 45 минут с Ариной).

В случае построения СК информанткой (кейс Арины) ей было предложено расположить свой круг общения в трех концентрированных кругах [15]. Логика построения СК информантом

начиналась с отбора людей, затем в интервью описывались отношения с ними, в том числе их близость к информанту, выделялась история этих отношений и подчеркивался контекст отношений. Когда исследовательница строила СК на основе интервью (кейс Ольги), то сначала она разбирала истории отношений из интервью, затем выделяла людей, которых можно перенести на СК, и далее обозначала отношения с ними. Ключевая особенность построения СК исследователем – это необходимость полностью полагаться на данные, полученные из интервью, тогда как в другом случае может быть обозначена связь с интервью. Результат в виде построенной СК зависит от инструкции информанту – если он самостоятельно строит СК, и от замысла исследователя – в случае построения СК исследователем.

Ключевой особенностью построения СК информантом является возможное расхождение с данными интервью, так как информант может не вспомнить каких-то людей или не обозначить изменения в отношениях. Поскольку используются два источника информации – данные СК и данные интервью, то возможны сущностные различия. Сравнивая данные, полученные из интервью, и данные с СК, могут быть предъявлены разные социальные круги взаимодействий – например, из данных интервью по описанию отношений их можно было бы отнести к близким отношениям, но информант помещает их на СК в более отдаленное от себя место. Такое расхождение объясняется тем, что нарратив состоит из оценочно-ситуативной логики описания отношений, а положение отношений в структуре позволяет оценить важность актора не только исходя из текущего контекста, а опираясь на весь предшествующий опыт и нормативный статус. Например, информантка в интервью описывала сложности в отношениях с мужем и позже, располагая позицию мужа на СК, она все равно поместила ее в ближайшем круге, объясняя это нормативным статусом мужа: «*Это же муж*» (Арина, 27 л.). Описание динамики в отношениях из интервью и структурирование отношений на СК обладают той же особен-

ностью. В нарративах отношения могут меняться и описываться контексты их изменений, однако их отображение на СК может не отличаться между собой – в случае если СК визуализирует сам информант.

К основным ограничениям построения СК информантом можно отнести необходимость в организации процесса сбора сетевых данных. Необходимо дать подробную инструкцию к построению СК, поскольку от нее зависит результат визуализации. В данном исследовании информантка отправляла СК исследователю по почте, вследствие чего качество картинки было невысоким, на карте присутствовали помарки, почерк мог быть непонятным и др. Но данную СК можно изобразить в цифровом виде, преобразовав оригинальное изображение. Основным ограничением построения СК исследователем является его полагание на данные из интервью, в ходе которого информант может вспомнить не всех акторов или отметить изменения не во всех отношениях. Преимуществом данного способа сбора сетевых данных является его экономичность, поскольку не тратится время информанта.

## Анализ сетевых карт на примере Ольги и Арины

Анализ СК, по мнению П. Атхенс, возможен двумя способами: 1) рассматриваются СК отдельно от интервью, 2) СК анализируются в контексте интервью [29]. В первом случае можно сравнить собранные СК между собой, а также совершить проверку интервьюеров, как было описано у Атхенс. При использовании второго способа позиция каждого актора на СК объясняется в интервью. А. Херз с соавторами предлагают подход к анализу СК, который используется и в количественном сетевом анализе и предполагает фокус на структуре, акторах и характере их взаимодействий [30]. Анализ структуры сети показывает общие паттерны сети в виде плотности и эквивалентности сети. Концентрируясь на акторах

в сети, исследователем определяются возможные связи между акторами и характеристики, которыми обладают акторы. Фокус на взаимодействия направлен на описания и интерпретации взаимодействий в сети.

*Общая структура сети*

Сравнивая сети Ольги и Арины на уровне общей структуры, можно заметить разницу в увеличении числа акторов в сети у Арины после переезда и родительства и уменьшении числа акторов у Ольги. Поскольку Арина переехала на втором триместре беременности, у нее было время включиться в сообщество будущих мам, а также взаимодействовать с другими знакомыми релокантами, сформировав слабые связи с ними. Ольга переехала накануне родов, она не включалась в общение с местным сообществом мам и российских мигрантов.

*Акторы в сети*

Фокусируясь на акторах, можно сказать, что в обоих случаях семьи и друзья поддерживали решение об миграции и родительстве. Также подчеркнем, что и Ольга, и Арина не живут в одном городе с родителями, поэтому непривычное физическое отдаление ощущается только от привычного круга друзей и они переживают нехватку в общении с ними, тогда как с родителями уже налажена опосредованная связь онлайн. «*Привычный круг друзей потерян, связь с ними обрывается и становится поверхностной*» (Арина, 27 л.).

*Взаимодействия в сети*

Анализируя взаимодействия в сетях, можно отметить, что у обеих девушек схожие паттерны в отношениях с друзьями, однако есть разница во взаимодействии с родителями: у Арины наблюдаются поддержка и помощь со стороны родителей, тогда как со стороны семьи Ольги имеется только эмоциональная поддержка. Но тем не менее несмотря на то, что родители живут в другом месте,

Ольга ожидала поддержку и физическую помощь с ребенком со стороны родителей, включающую в себя непосредственное участие в уходе за ребенком, однако эти ожидания не оправдались. В интервью Ольга проговаривает нормативные ожидания от родителей: «*Родители обычно помогают, принимают и поддерживают*» (Ольга, 32 г.). Тогда как у Арины ожидания помощи и поддержки оправдались, а также улучшились отношения с мамой и сестрой.

## Заключение

На основе эмпирического исследования изменения сетей взаимодействий молодых родителей в условиях миграции нами была апробирована методология КСА, проведено сравнение двух подходов к построению СК и представлен пример их анализа. Однако ограничением данной работы является сравнение только двух эмпирических кейсов. КСА представлен как методологический подход к изучению отношений в сети, где синтезируются структурализм и интерпретативизм. С помощью КСА можно изучать эгосети, для визуализации используются СК. При всем разнообразии видов и подходов к построению СК [8; 9; 10] сравнения между построением карт исследователем либо самим информантом ранее не проводились, это было реализовано в данной статье.

Данная работа иллюстрирует описанные в литературе [10; 11; 15] эвристические возможности КСА в зависимости от последовательности располагаемых интервью и СК. В случае самостоятельного заполнения СК информантом анализируются два вида собранных данных – из интервью и СК. Возможна разница между самопрезентацией отношений в нарративах и положения в структуре. Синтезируя смыслы отношений из нарративов и структурализацию этих отношений, тем самым наслаивая их друг на друга, можно достичь большей глубины в восприятии и понимании отношений. В другом случае, когда исследователь полностью полагается на данные из интервью и самостоятельно

структурирует отношения в СК, анализируются только данные из интервью, которые дополняются визуализацией СК. КСА способен охарактеризовать восприятие сетевых отношений в полной мере – в случае сбора данных при помощи интервью и построения СК информантом.

## ЛИТЕРАТУРА

1. *Мальцева Д.В.* Сетевой подход в социологии: генезис идей, современное состояние и возможности применения: дис. ... канд. социол. наук: 22.00.01. М., 2014. 177 с. EDN: ZSKVZF.

2. *Kim A., Maltseva D.* Qualitative social network analysis: studying the field through the bibliographic approach // Quality and Quantity. 2023. № 58. P. 385–411. DOI: 10.1007/s11135-023-01651-6. EDN: KIJGES.

3. *Häussling R.* Allocation to Social Positions in Class: Interactions and Relationships in First Grade School Classes and Their Consequences // Current Sociology. 2010. Vol. 58, № 1. P. 119–138. DOI: 10.3102/0091732X20903311.

4. *Mische A*. Culture, Networks, and Interaction in Social Movement Publics // Simposio de Berlín. Marzo. 2008. № 20. P. 1–6.

5. *Diaz-Bone R.* Gibt es eine qualitative Netzwerkanalyse? // Historical Social Research. 2008. Vol. 33, № 4. P. 311–343. DOI: 10.12759/hsr.33.2008.4.311-343.

6. *Ким А.В.* Качественный сетевой анализ в стратегии смешивания методов в социальных науках: систематический обзор литературы // Социология: методология, методы, математическое моделирование. 2021. Т. 2, № 53. С. 83–116. DOI: 10.19181/4m.2021.53.3. EDN: XJRIYF.

7. *Ким А.В.* Методологический подход к изучению отношений в сети: качественный сетевой анализ // Интеракция. Интервью. Интерпретация. 2023. Т. 15, № 3. С. 11–30. DOI: 10.19181/inter.2023.15.3.1. EDN: BADYQC.

8. *Herz A.* Ego-zentrierte Netzwerkanalysen zur Erforschung von Sozialräumen // Soziale Netzwerkanalyse. Theorie – Praxis – Methoden. 2012. № 2. S. 133–152.

9. *Straus F.* Netzwerkanalysen: Gemeindepsychologische Perspektiven für Forschung und Praxis. Wiesbaden, Germany: Dt. Univ.-Verl., 2002. 354 s.

10. *Gamper M., Schönhuth M., Kronenwett M.* Bringing qualitative and quantitative data together: Collecting network data with the help of the software tool VennMaker // Social networking and community behavior modeling: Qualitative and quantitative measures. Hershey, PA: IGI Global, 2012. P. 193–213. DOI: 10.4018/978-1-61350-444-4.ch011.

11. *Jaspersen L.J., Stein C.* Beyond the Matrix: Visual Methods for Qualitative Network Research // British Journal of Management. 2019. Vol. 30, № 3. P. 748–763. DOI: 10.1111/1467-8551.12339.

12. *Белослудцев А.Н., Дзюба Е.В.* Релокация как социально-экономический и правовой институт // Россия в глобальном мире. 2023. Т. 26, №2. С. 97–123. DOI: 10.48612/rg/RGW.26.2.7. EDN: VPEHQH.

13. *Рахмонов А.Х.* Вынужденная эмиграция из России 2022: потенциал российских IT-специалистов для центральноазиатских стран-участниц СНГ // Вестник университета. 2023. № 7. С. 162–170. DOI: 10.26425/1816-4277-2023-7-162-170. EDN: NDXEZA.

14. *Wasserman S., Faust K.* Social Network Analysis: Methods and Applications. Cambridge: Cambridge University Press, 1994. 825 p. DOI: 10.1017/CBO9780511815478.

15. *Sommer E., Gamper M.* Beyond structural determinism: advantages and challenges of qualitative social network analysis for studying social capital of migrants // Global Networks. 2020. Vol. 21, № 2. P. 608–625. DOI: 10.1111/glob.12302.

16. *Antonucci T.C.* Measuring social support networks: Hierarchical mapping technique // Generations: Journal of the American Society on Aging. 1986. Vol. 10, № 4. P. 10–12.

17. *McCarty C.* A comparison of social network mapping and personal network visualization // Field methods. 2007. Vol. 19, № 2. P. 145–162. DOI: 10.1177/1525822X06298592. EDN: JLZJXH.

18. *Zwijze-Koning K.H., De Jong M.D.T.* Auditing information structures in organizations: A review of data collection techniques for network analysis // Organizational Research Methods. 2005. Vol. 8, № 4. P. 429–453. DOI: 10.1177/1094428105280120. EDN: JOPJCP.

19. *Conway S., Steward F.* Mapping innovation networks //International Journal of Innovation Management. 1998. Vol. 2, № 2. P. 223–254. DOI: 10.1142/S1363919698000110. EDN: ESDQBP.

20. *Malandrino A.* Comparing qualitative and quantitative text analysis methods in combination with document-based social network analysis to understand policy networks // Quality & Quantity. 2023. № 58. P. 2543–2570. DOI: 10.1007/s11135-023-01753-1. EDN: XXIQRG.

21. *Мальцева Д.В., Моисеев С.П.* Сетевой анализ биографических интервью: кейс Т.И. Заславской // Телескоп: журнал социологических и маркетинговых исследований. 2018. № 2 (128). С. 15–24. EDN: YVKHJF.

22. *Bertolotti F., Tagliaventi M.R.* Discovering complex interdependencies in organizational settings: the role of social network analysis in qualitative research // Qualitative Rsearch in Organizations and Management: An International Journal. 2007. Vol. 2, № 1. P. 43–61. DOI: 10.1108/17465640710749126.

23. *Belsky J.* Transition to parenthood // Medical Aspects of Human Sexuality. 1986. Vol.20, № 9. P. 56–59.

24. *Cowan P.* Individual and Family Life Transitions: A Proposal for a New Definition. Hillsdale, NJ: Lawrence Erlbaum Associates, 1991. 393 p.

25. *Falicov C.J.* Family transitions: Continuity and change over the life cycle. New York: Guilford Press, 1991. 476 p.

26. *Lois D.* Types of social networks and the transition to parenthood // Demographic Research. 2016. Vol. 34, № 23. P. 657–688. DOI: 10.4054/DemRes.2016.34.23.

27. *Ширманова И.* 800 тысяч россиян могли покинуть страну в 2022 году // Если быть точным: [сайт]. 27.02.2023. URL: https://tochno.st/materials/rossiyan-mogli-pokinut-stranu-v-2022-godu (дата обращения: 27.05.2024).

28. Российская ризома: социальный портрет новой эмиграции / Н. Костенко, М. Завадская, Э. Камалов, И. Сергеева // Re: Russia. 27.08.2023. URL: https://re-russia.net/expertise/045/ (дата обращения: 27.05.2024).

29. *Ahrens P.* Qualitative network analysis: A useful tool for investigating policy networks in transnational settings? // Methodological Innovations. 2018. Vol. 11, №. 1. P. 2059799118769816. DOI: 10.1177/2059799118769816.

30. *Herz A., Peters L., Truschkat I.* How to do qualitative structural analysis? The qualitative interpretation of network maps and narrative interviews // Forum: Qualitative Social Research. 2015. Vol.16, № 1. P. 1–24. DOI: 10.17169/fqs-16.1.2092.

# Приложение

**Figure 1: Business network of an IT company – start-up phase, migrant market**

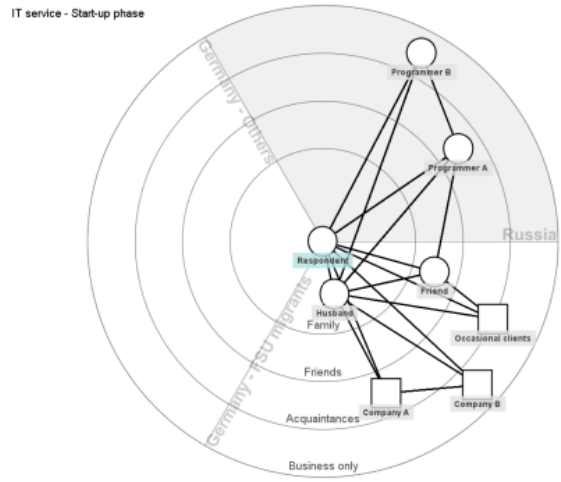
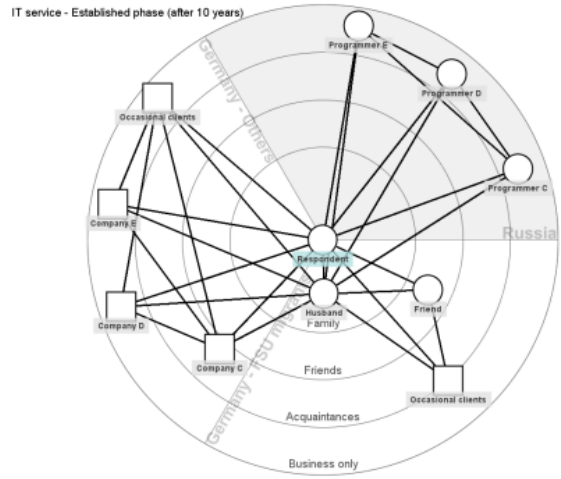



**Figure 2: Business network of an IT company – after ten years, mainstream market**




Note: circles indicate individuals; squares indicate groups of people or organizations.

***Рис. 1. Пример сетевой карты, построенной информантом***

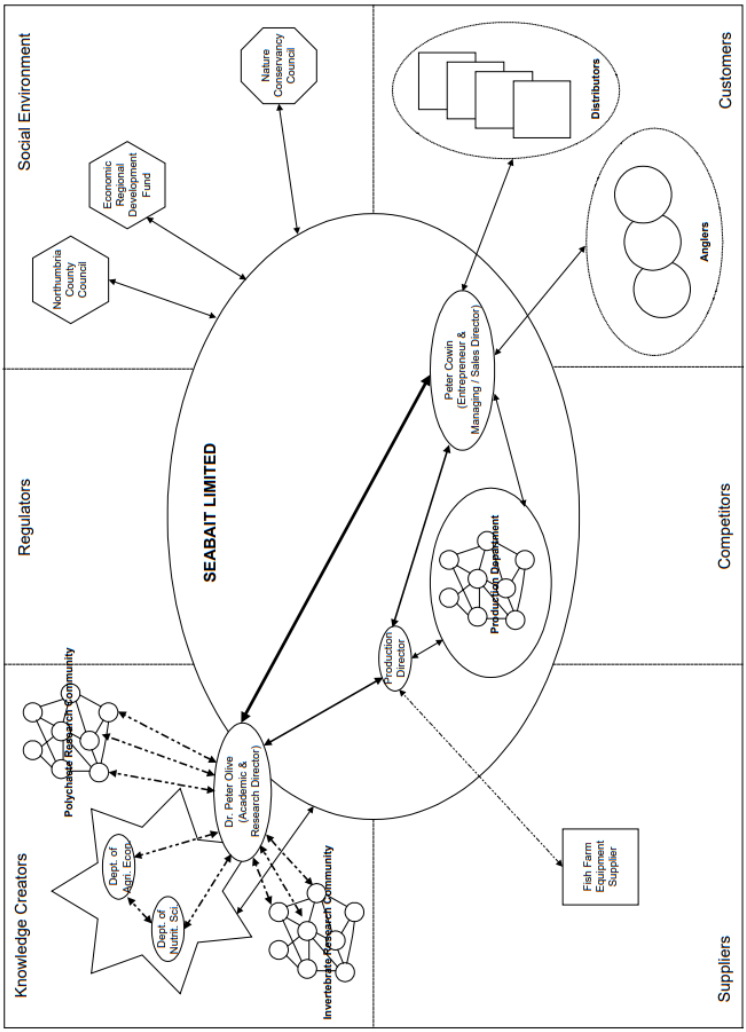



***Рис. 2. Пример сетевой карты, построенной исследователем***

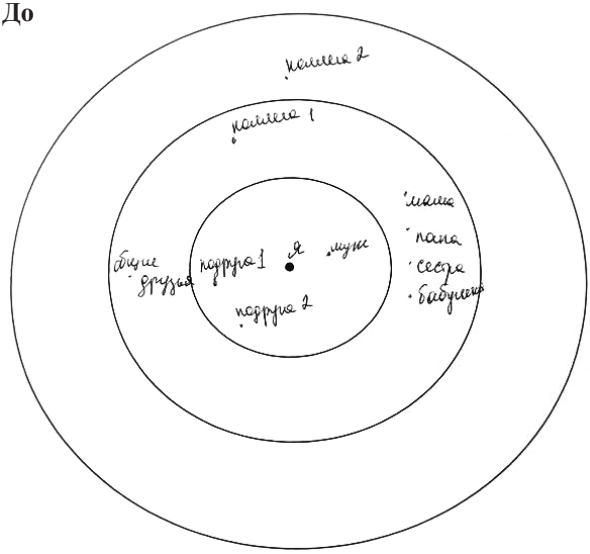
*Таблица*

ДАННЫЕ ОБ ИНФОРМАНТКАХ И СПОСОБЫ СБОРА ДАННЫХ

| | | | | | |
|---|---|---|---|---|---|
| Казахстан | Ольга, 32 года, Zoom, 1 час 42 минуты | Арина, 27 лет, Zoom, 1 час 45 минут | Ирина, 37 лет, Zoom, 1 час 33 минуты | Ярослава, 30 лет, Телеграм, 1 час 37 минут | Алла, 28 лет, Zoom, 1 час 14 минут |
| Армения | Олеся, 24 года, Zoom, 1 час 24 минуты | Екатерина, 32 года, Zoom, 1час 2 минуты | Алена,35 лет, Zoom, 1 час 15 минут | Катя, 30 лет, Телеграм, 1 час 33 минуты | Маша,27 лет, Телеграм, 1 час 43 минуты |
| Грузия | Анна, 36 лет, Zoom, 2 часа 2 vинуты | Екатерина, 23 года, Телеграм, 1 час 5 минут | Ксения, 31 год, Телеграм, 1 час 3 минуты | Анастасия, 23 года, Zoom, 1 час 10 минут | Вероника, 31 год, Телеграм, 1 час 1 минута |

*Примечание.* Жирным шрифтом выделены информанты, чьи кейсы рассматривались в данной статье.

**До**




**После**

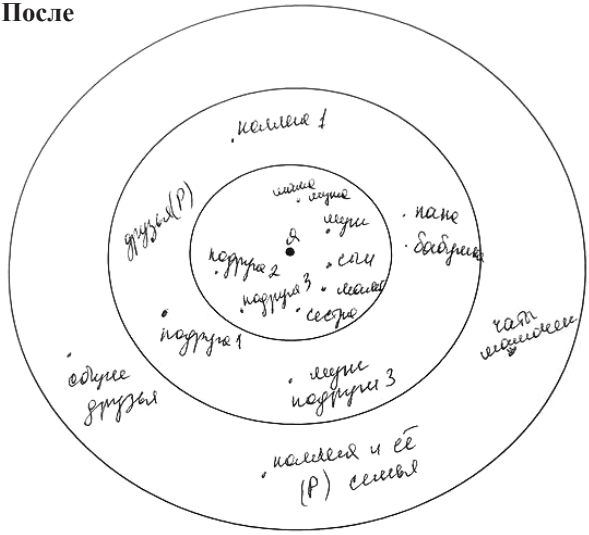



***Рис. 3. Сетевые карты, нарисованные информантом: «Сеть Арины до и после» (оригинал)***

***Kim Aryuna V.,***
*Junior Research Fellow at the International Laboratory for Applied Network Research, HSE University, Moscow, Russia,* avkim@hse.ru

**Qualitative social network analysis in practice: comparison of methods for network maps construction**

This article is devoted to the application of qualitative network analysis on the example of studying the social interactions of young parents in migration. A conceptual description of qualitative network analysis, network visualization features, design of qualitative network research, comparison of two approaches to building network maps and an example of their analysis are presented. Qualitative network analysis is aimed at studying relationships in the network and consists of interpretative and structural components, which are implemented through interviews and building a network map. The two approaches to building network maps differ in that in one case, network maps are built by the researcher based on an interview, and in the other case, the construction of the map is delegated to the informant himself. The first method is limited by the interview data and is their structural visualization. The second method of building a network map by the informant himself creates additional resources for interpretation – contrasting narratives and visual structuring of the informant's interactions.

*Keywords:* qualitative social network analysis, network analysis, network visualization, network maps, qualitative methods, parenting, migration, Russian migrants

**References**

1. Maltseva D.V. *Network approach in sociology: the genesis of ideas, the current state and possibilities of application* (in Russian), dis. ... cand. in Sociology: 22.00.01. Moscow, 2014. 177 p.
2. Kim A., Maltseva D. Qualitative social network analysis: studying the field through the bibliographic approach, *Quality and Quantity*, 2023, no. 58, p. 385–411. DOI: 10.1007/s11135-023-01651-6.
3. Häussling R. Allocation to Social Positions in Class: Interactions and Relationships in First Grade School Classes and Their Consequences, *Current Sociology*, 2010, vol. 1, no. 5), p. 119–138. DOI: 10.3102/0091732X20903311.
4. Mische A. Culture, Networks, and Interaction in Social Movement Publics, *Simposio de Berlín. Marzo*, 2008, no. 20, p. 1–6.

5. Diaz-Bone R. Gibt es eine qualitative Netzwerkanalyse? *Historical Social Research*, 2008, vol. 4, no. 33, p. 311–343. DOI: 10.12759/hsr.33.2008.4.311-343.
6. Kim A.V. Qualitative social network analysis in the strategy of mixing methods in the social sciences: a systematic review of the literature (in Russian), *Sotsiologiya 4M (Sociology: methodology, methods, mathematical modeling)*, 2021, no. 53, p. 83–116. DOI: 10.19181/4m.2021.53.3. EDN: XJRIYF.
7. Kim A.V. Methodological approach to the study of relationships in the network: qualitative social network analysis (in Russian), *Interaction. Interview. Interpretation*, 2023, vol. 15, no. 3, P. 11–30. DOI: 10.19181/inter.2023.15.3.1.
8. Herz A. Ego-zentrierte Netzwerkanalysen zur Erforschung von Sozialräumen, Soziale Netzwerkanalyse. *Theorie – Praxis – Methoden*. 2012, N. 2, S. 133–152.
9. Straus F. *Netzwerkanalysen: Gemeindepsychologische Perspektiven für Forschung und Praxis*. Wiesbaden, Germany: Dt. Univ.-Verl., 2002. 354 s.
10. Gamper M., Schönhuth M., Kronenwett M. Bringing qualitative and quantitative data together: Collecting network data with the help of the software tool VennMaker, *Social networking and community behavior modeling: Qualitative and quantitative measures*. Hershey, PA: IGI Global, 2012, P. 193–213. DOI: 10.4018/978-1-61350-444-4.ch011.
11. Jaspersen L.J., Stein C. Beyond the Matrix: Visual Methods for Qualitative Network Research, *British Journal of Management*, 2019, vol. 30, no. 3, p. 748–763. DOI: 10.1111/1467-8551.12339.
12. Belosludtsev A.N., Dzyuba E.V. Relocation as a socio-economic and legal institution (in Russian), *Russia in the global world*, 2023, vol. 26, no. 2, p. 97–123. DOI: 10.48612/rg/RGW.26.2.7.
13. Rakhmonov A.H. Forced emigration from Russia 2022: the potential of Russian IT specialists for Central Asian CIS member countries (in Russian), *Bulletin of the University*, 2023, no. 7, p. 162–170. DOI: 10.26425/1816-4277-2023-7-162-170.
14. Wasserman S., Faust K. *Social Network Analysis: Methods and Applications*. Cambridge: Cambridge University Press, 1994, 825 p. DOI: 10.1017/CBO9780511815478.
15. Sommer E., Gamper M. Beyond structural determinism: advantages and challenges of qualitative social network analysis for studying social

capital of migrants, *Global Networks*, 2020, vol. 21, no. 2, p. 608–625. DOI: 10.1111/glob.12302.
16. Antonucci T.C. Measuring social support networks: Hierarchical mapping technique, *Generations: Journal of the American Society on Aging*, 1986, vol. 10, no. 4, p. 10–12.
17. McCarty C. A comparison of social network mapping and personal network visualization, *Field Methods*, 2007, vol. 19, no. 2, p. 145–162. DOI: 10.1177/1525822X06298592.
18. Zwijze-Koning K.H., De Jong M.D.T. Auditing information structures in organizations: A review of data collection techniques for network analysis, *Organizational Research Methods*, 2005, vol. 8, no. 4, p. 429–453. DOI: 10.1177/1094428105280120.
19. Conway S., Steward F. Mapping innovation networks, International *Journal of Innovation Management*, 1998, vol. 2, no. 2, p. 223–254. DOI: 10.1142/S1363919698000110.
20. Malandrino A. Comparing qualitative and quantitative text analysis methods in combination with document-based social network analysis to understand policy networks, *Quality & Quantity*, 2023, no. 58, p. 2543–2570. DOI: 10.1007/s11135-023-01753-1.
21. Maltseva D.V., Moiseev S.P. Network analysis of biographical interviews: the case of T.I. Zaslavskaya (in Russian), *Telescope: Journal of Sociological and Marketing Research*, 2018, no. 2 (128), p. 15–24.
22. Bertolotti F., Tagliaventi M.R. Discovering complex interdependencies in organizational settings: the role of social network analysis in qualitative research, *Qualitative Research in Organizations and Management: An International Journal*, 2007, vol. 2, no. 1, p. 43–61. DOI: https://doi.org/10.1108/17465640710749126.
23. Belsky J. Transition to parenthood, *Medical Aspects of Human Sexuality*, 1986, no. 20 (9), p. 56–59.
24. Cowan P. *Individual and Family Life Transitions: A Proposal for a New Definition*. Hillsdale, NJ: Lawrence Erlbaum Associates, 1991. 393 p.
25. Falicov C.J. *Family transitions: Continuity and change over the life cycle*. New York: Guilford Press, 1991. 476 p.
26. Lois D. Types of social networks and the transition to parenthood, *Demographic Research*, 2016, vol. 34, no. 23, p. 657–688. DOI: 10.4054/DemRes.2016.34.23.

27. Shirmanova I. 800 thousand Russians could leave the country in 2022 (in Russian), *To be precise*: [website]. 27.02.2023. URL: https://tochno.st/materials/rossiyan-mogli-pokinut-stranu-v-2022-godu (date of application: 27.05.2024).
28. The Russian rhizome: a social portrait of a new emigration (in Russian), *Re: Russia*. 27.08.2023. URL: https://re-russia.net/expertise/045/ (date of access: 27.05.2024).
29. Ahrens P. Qualitative network analysis: A useful tool for investigating policy networks in transnational settings? *Methodological Innovations*, 2018, vol. 11, no. 1, p. 2059799118769816. DOI: 10.1177/2059799118769816.
30. Herz A., Peters L., Truschkat I. How to do qualitative structural analysis? The qualitative interpretation of network maps and narrative interviews, *Forum: Qualitative Social Research*, 2015, no. 16 (1), p. 1–24. DOI: 10.17169/fqs-16.1.2092.

## *ОПЫТ ПРАКТИЧЕСКОГО ПРИМЕНЕНИЯ СЕТЕВОГО АНАЛИЗА*

DOI: 10.19181/4m.2023.32.1.4
EDN: FDYTSV

С. Ткач, П.Д. Воробьева, М.М. Русакова
(*Санкт-Петербург*)

# ОПЫТ РЕАЛИЗАЦИИ ДИСКУРС-АНАЛИЗА И КОНЦЕПТУАЛЬНОГО КАРТИРОВАНИЯ СООБЩЕСТВ ЗДОРОВОГО ПИТАНИЯ[1]

В статье приводится опыт имплементации методов дискурс-анализа в трактовке Э. Лакло и Ш. Муфф и метода концептуального картирования в интерпретации У. Трочима посредством методики сетевого анализа на примере тематики здорового питания. Результатом анализа выступают графы, которые позволяют выделить борьбу агентов дискурса за означение ключевых дискурсивных знаков, а также рассмотренная в статье в качестве примера концептуальная карта участников онлайн-дискуссии по спорным в рамках тематики вопросам. В качестве эмпирической базы дискурс-анализа выступили 3 тыс. собранных комментариев в четырех сообществах о здоровом питании социальных сетей «ВКонтакте» и «Одноклассники». Разработанная реализация методов концептуального картирования и дискурс-анализа адаптирована для анализа онлайн-дискуссий. Последующая валидация разработанных методов видится

**Сергей Ткач** – социолог, Центр прикладной социологии, Санкт-Петербургский государственный университет, Санкт-Петербург, Россия. Email: s.tkach@spbu.ru.
**Полина Дмитриевна Воробьева** – социолог, Центр прикладной социологии, Санкт-Петербургский государственный университет, Санкт-Петербург, Россия. Email: st098355@student.spbu.ru.
**Майя Михайловна Русакова** – кандидат социологических наук, директор, Центр прикладной социологии, Санкт-Петербургский государственный университет, Санкт-Петербург, Россия. Email: m.rusakova@spbu.ru.

направлением дальнейших исследований. У предлагаемых дизайнов есть ряд ограничений, которые обсуждаются в статье.

*Ключевые слова:* дискурс-анализ, концептуальное картирование, анализ графов, сентимент-анализ, анализ естественного языка

## Введение

Социальные сети обладают рядом особенностей, которые затрудняют их исследование. Структура интерфейса социальных сетей может значительно влиять на то, как происходит дискуссия [1; 2], и делать ее отличной от офлайн-дискуссий или дискуссий в печатных и телевизионных СМИ. Это требует от исследователей разработки методов, более подходящих под формат онлайн-дискуссии. Возможное решение видится в совершенствовании существующих исследовательских методов посредством имплементации цифровых элементов анализа. Методы дискурс-анализа и концептуального картирования имеют длинную традицию эффективного применения учеными для решения исследовательских задач анализа текста. Однако в своем изначальном дизайне они не ориентированы на анализ дискуссий в социальных сетях. Имплементация цифровых элементов анализа позволит ухватить данную специфику. Также комбинация методов позволит получить более детальное представление об исследуемом объекте. Реализация этих методов будет продемонстрирована на примере дискуссий в онлайн-сообществах здорового питания.

## Концептуальное картирование

Концептуальное картирование было разработано как самостоятельный метод У. Трочимом в начале 1980-х гг. Метод совмещал в себе элементы качественного анализа (мозговой штурм, неструктурированная сортировка, интерпретация) и количественного

анализа (многомерное шкалирование, иерархический кластерный анализ). Метод позволяет группе людей изобразить наглядную последовательную концептуальную схему интересующего их вопроса [3]. Сам Трочим определяет концептуальное картирование как структурированный процесс, сосредоточенный на интересующей теме, включающий вклад нескольких участников, итогом которого становится интерпретируемое графическое представление идей и концепций участников построения, а также то, как эти идеи взаимосвязаны. Под концептуальной картой Трочим понимает структурную концептуализацию или многомерную графическую репрезентацию набора идей, сгенерированных группой [4]. Изначально метод включал в себя шесть этапов: подготовка к проведению исследования (включая отбор участников дискуссии); производство высказываний участниками дискуссии; структурирование полученных высказываний; представление высказываний в виде концептуальной карты (с использованием многомерного шкалирования и кластерного анализа – сам Трочим рекомендовал для удобства ограничиться двумерным представлением); интерпретация полученных карт; практическое использование карт в решении прикладных задач [5]. Шестой этап является прикладным, поэтому исследователи, как правило, выделяют пять аналитических этапов. За прошедшие десятилетия метод приобрел множество вариаций и интерпретаций в рамках исследований в общественных науках. В частности, можно упомянуть работу Н. Абрамовой и Ю. Николаевой об использовании концептуального картирования как метода повышения валидности результатов оценочного исследования [6], а также шестикомпонентную модель реализации метода, разработанную Трочимом совместно с М. Кейном [7]. Метод концептуального картирования предполагает работу с офлайн-дискуссиями специально отобранных экспертов. Однако он показал свою эффективность и для анализа групповых дискуссий экспертов опыта [8] (информантов, обладающих экспертностью, приобретенной посредством проживания уникального опыта,

а не наличием квалификации [9]), что делает перспективным использование этого метода для анализа онлайн-дискуссий. Однако сама специфика онлайн-дискуссий требует доработки метода посредством имплементации элементов анализа цифровых следов. Авторы в статье опираются на модель Трочима «концептуальной карты» и общую логику его подхода, внеся изменения в метод концептуального картирования таким образом, чтобы он стал более подходящим для анализа онлайн-дискуссий. Мы понимаем под концептуальной картой интерпретируемое графическое представление сообщений участников онлайн-дискуссии, а также то, как идеи, изложенные в сообщениях, взаимосвязаны между собой.

## Дискурс-анализ: концептуализация

В академической литературе не существует единого мнения о сущности понятия «дискурс» и, как следствие, о понятии «дискурс-анализ». В понимании Э. Лакло и Ш. Муфф дискурсом можно назвать упорядоченную тотальность знаков. Лакло и Муфф полагают, что некоторые знаки более ценны, чем другие. К таким знакам обращено больше внимания, их чаще изображают, о них чаще говорят, от их значения зависят значения других знаков, менее ценных. Их можно назвать *узловыми точками* – они сплетают воедино дискурс, фиксируют его. Содержание узловых точек пусто и изменчиво: разные дискурсы могут наделить его значением, тем самым обозначив свое превосходство над другими дискурсами. Здесь появляется другое ключевое понятие теории Лакло и Муфф – *борьба*. Различные дискурсы постоянно *ведут борьбу*, стремясь зафиксировать свое значение в языке. Отсюда становится понятна одна из основных целей дискурс-анализа в предложенной трактовке: проследить процессы борьбы за определенный способ определения значений в языке [10].

Новым методологическим вызовом для дискурс-анализа становится развитие цифровых технологий. Несмотря на то, что существу-

ющее в теории Лакло и Муфф всеобъемлющее постмодернистское понятие дискурса позволяет обращаться к различным формам данных в качестве объектов исследования, аналитическое переложение теоретических концепций исходной теории к интернет-дискуссиям – достаточно нетривиальная задача [11]. Дискурс-анализ в осмыслении Лакло и Муф – это метод, используемый для анализа политического дискурса, и возможности применения аналитического аппарата этого метода для исследования других сфер общественной жизни неочевидны. Так, потенциал использования метода дискурс-анализа для исследований медиа показали Н. Карпентер и Б. Де Клин [12]. В их совместной статье в том числе рассматривается возможность изучения медиа в качестве поля дискурсивной борьбы. О. Игнатьевой отмечается, что сама специфика интернет-дискуссий, такая как особый интернет-язык, для которого характерны сокращения или киберорфография, усложняет процесс аналитической обработки дискурса. В дополнение интернет-дискуссии могут быть рассмотрены в качестве устной речи, что требует особых методов анализа, например – конверсационного анализа [13]. Большое число исследований, направленных на изучение большого количества текстовых данных и предполагающих цифровые этапы обработки дискурса, как отмечает Дж. Хадитаги и другие ученые, имеют существенный недостаток в виде назначения узловых точек «внешней силой», что может не соответствовать эмпирическому объекту изучения [14]. В статье П. Бакумова приводятся примеры нескольких исследований, в рамках которых в качестве инструментов анализа использовалась цифровая имплементация метода дискурс-анализа в трактовке Лакло и Муфф. Бакумов отмечает, что большинство попыток использования выбранного метода дискурс-анализа начинаются с переопределения теоретических концепций исходного метода, а также последовательности шагов анализа [15]. Однако не сложно заметить, что каждая имплементация этого метода крайне специфична и напрямую зависит от предмета и целей исследования. Это к приводит выводу, что проведение дальнейших

исследований в сфере использования цифровых методов анализа дискурса необходимо.

Авторы обращаются к традиции критического дискурс-анализа в интерпретации Лакло и Муфф. Ключевым изменением в сравнении с традиционным дискурс-анализом является использование методов визуализации, посредством которых наглядно изображается совстречаемость тех или иных знаков дискурса. В понимании Лакло и Муфф поле дискурса – это поле борьбы. Значение, скрывающееся за знаками, не постоянно – оно устанавливается различными агентами, преследующими свои интересы посредством высказываний – дискурсивных практик. Интересанты, чью позицию выражают агенты дискурса, в большинстве случаев не могут быть напрямую соотнесены с конкретными людьми. В повседневной речи люди упоминают слова с разным контекстуальным значением, транслируя тем самым разные интересы, скрытые за этим. Агент (действующее лицо или группа) производит высказывание, в котором упоминает какую-то *узловую точку*. Делает он это в определенном контексте: в высказывании упоминается не только сама узловая точка, но и другие знаки. Этот контекст (множество знаков, упоминаемое в высказывании) способен, согласно Лакло и Муфф, переопределить значение *узловой точки*. Так как этот контекст разный у разных агентов, значение *узловой точки* постоянно переопределяется. Те агенты, кто делают это успешнее, побеждают на поле боя дискурса. Победа может быть с некоторыми допущениями названа конечной, когда определенное контекстуальное толкование слова полностью исчезает из речи. Такую окончательную победу, как и промежуточные, можно зафиксировать в рамках дискурс-анализа.

## Дискурс-анализ: методология

### Методология отбора данных

Исследовательской группой были отобраны четыре сообщества социальных сетей «ВКонтакте» и «Одноклассники», посвя-

щенные здоровому питанию[1]. Отбор комментариев проходил с 16 апреля по 6 мая 2023 года.

Сообщества были отобраны по следующим критериям.

1. Количество подписчиков: отбирались наиболее популярные сообщества. Ранжирование происходило при помощи поиска в сетях «ВКонтакте» и «Одноклассники».

2. Активность сообщества. Если сообщество было неактивным, то есть в нем не было ни одного поста за последнюю неделю, оно удалялось. По итогам выборки ни одно сообщество не было удалено.

Из отобранных сообществ были агрегированы все комментарии, когда-либо оставленные пользователями, – 15 618 комментариев. К комментариям сохранялись также метаданные: пост, под которым они были размещены, пользователи, оставившие их, время комментария, количество отметок «нравится». Критерии для отбора 3000 комментариев для анализа указаны ниже.

*Методология анализа*

Главной метафорой, вокруг которой выстраивалась разработка метода – поле дискурса, представляющее собой систему упорядоченных знаков посредством связей совстречаемости. Связь и борьба,

[1] Работа выполнена при поддержке СПбГУ, шифр проекта 121062300141-5. Исследование, в котором были реализованы концептуальное картирование и дискурс-анализ, было посвящено различным потребительским практикам россиян в зависимости от их уровня финансового благополучия.

Анализируемые сообщества:
1) «КЕТО диета», ссылка на сообщество «ВКонтакте»: https://vk.com/ketodieta;
2) «Диета Углеводов.нет», ссылка на сообщество «ВКонтакте»: https://vk.com/uglevodovnet;
3) Greenway Global, ссылка на сообщество «ВКонтакте»: https://vk.com/greenwayglobaloffical;
4) «Павел Корпачев – здоровье, фитнес, тренировки», ссылка на сообщество в «Одноклассниках»: https://ok.ru/zenofit.

которые подразумевались Э. Лакло и Ш. Муфф иносказательно, могут быть изображены наглядно. Встречаемость (контекстность) одного знака по отношению к другому может быть представлена в виде ребра графа. В исследовании под знаком операционально понимается слово. Тогда знаки с наибольшим числом ребер становятся узловыми точками дискурса. Под агентами понимаются пользователи сообщества. То, какие агенты используют знаки, можно выделить цветом. Так, словосочетания, которые используются агентами с одной позицией, будут окрашены в один цвет, а агентами с другой позицией – в другой. Под позицией понимается ценностная установка, которая кажется агентам правильной и которую они стараются донести в своих высказываниях. Борьба агентов сводится к образованию как можно большего числа ребер с узловой точкой, другими словами – к расширению пространства контекста точки дискурсивной борьбы. Теперь перейдем непосредственно к шагам реализации такой модели.

Первым шагом была кодировка комментариев. Исследовательская группа знакомилась с подвыборкой из 1000 комментариев. Данная подвыборка состояла из комментариев, которые получили наибольшее количество отметок «нравится». Наибольшее число отметок «нравится» составило 180, а наименьшее – 4. Среди всех комментариев с 4 отметками «нравится» были отобраны самые недавние – как представляющие наиболее актуальный дискурс в сообществе. Размер выборочной совокупности ограничивался трудовыми ресурсами исследовательской группы.

**Процедура кодировки.** Прочитав комментарии, группа выделяла позиции агентов дискурса (классы) (например, «худеющие» – те, кто придерживается позиции, что здоровое питание – это инструмент в процессе похудения и т.д.) и присваивала каждому комментарию соответствующую метку – от 1 до n (при n выделенных позиций). Описание позиций агентов записывалось. Ручная кодировка комментариев, получивших наибольшее количество оценок «нравится», обусловлена тем, что позволила выделить те группы

агентов, которые наиболее влиятельны в исходном дискурсе. На первом этапе были сформированы и уточнены инструкции кодирования, на втором этапе эти инструкции использовались дополнительными кодировщиками. Исследовательской группой были вручную классифицированы в соответствии с выделенными при кодировке позициями другие случайно отобранные 2000 комментариев. Случайный отбор осуществлен функцией sample библиотеки Pandas.

Следующим пунктом была подготовка текста комментариев для дальнейшего анализа. В предобработке использовались библиотеки языка Python: pymorphy2 и NLTK:

1) была удалена вся пунктуация в тексте;

2) слова были приведены к нормальной форме и нижнему регистру;

3) среди слов были отобраны только прилагательные и существительные;

4) все слова в тексте были разбиты на словосочетания (биграммы методом nltk.bigrams) с указанием частоты их встречаемости;

5) из полученных словосочетаний были дополнительно удалены те, которые содержали наиболее общеупотребимые слова – слова бытовой повседневной речи (при помощи словаря библиотеки NLTK языка Python);

6) из словосочетаний был построен граф при помощи программы Gephi – ребрами графа выступала совстречаемость слов на расстоянии 1 в тексте.

## Результаты

Граф поля дискурса строился из словосочетаний, приведем наиболее популярные из них (рис. 1).

Основная часть посвящена ведению диет: день кето, неделя кето, неделя два, вернуться кето – эти словосочетания посвящены кетодиете. Другие популярные словосочетания посвящены прак-

тикам приема пищи, и в особенности их частоте и продолжительности: прием пища, каждый день, раз [в] день, первый день, течение день, пара день [дней], каждый прием, прием неделя. Третья группа наиболее популярных словосочетаний включает в себя пищевые продукты и соединения: сливочный масло, яблочный уксус, фульвовая кислота, цитрат магний. В последнюю группу

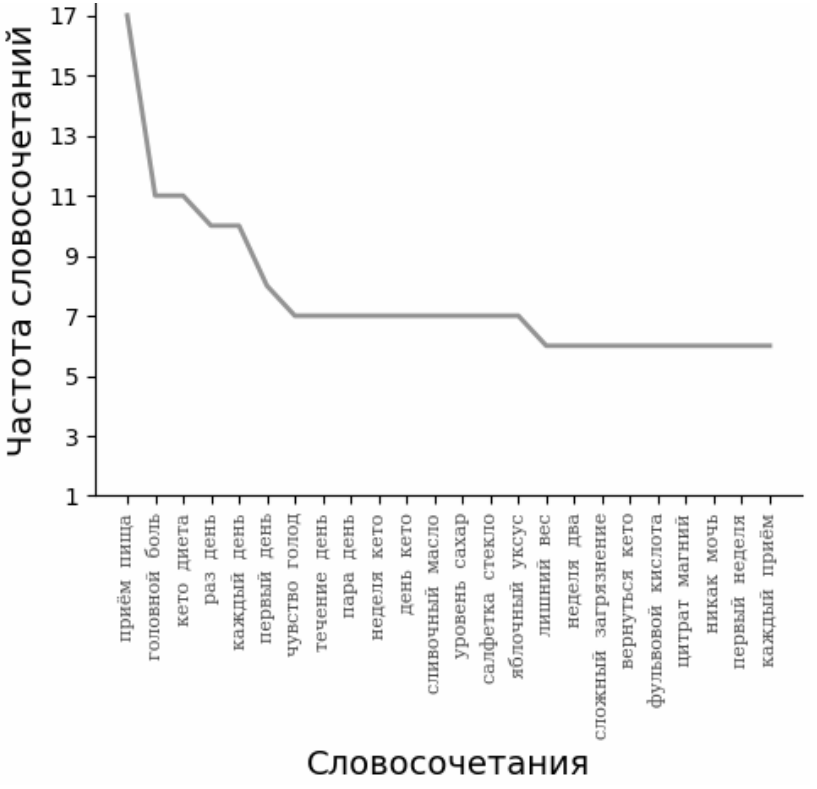



***Рис. 1. Частотное распределение наиболее популярных словосочетаний сообществ здорового питания***

можно отнести ментальные и соматические состояния: головной боль, чувство голод, уровень сахар, лишний вес, никак [не] мочь.

Число узлов составило 648, а связей – 440 (рис. 2). Плотность полученного графа составила 0,002, средняя степень центральности – 1,358, средняя длина пути в графе – 3,916, средняя степень близости узла – 0,76636, модулярность – 0,94, число обнаруженных компонент методом Гирвана-Ньюмана составило 222.

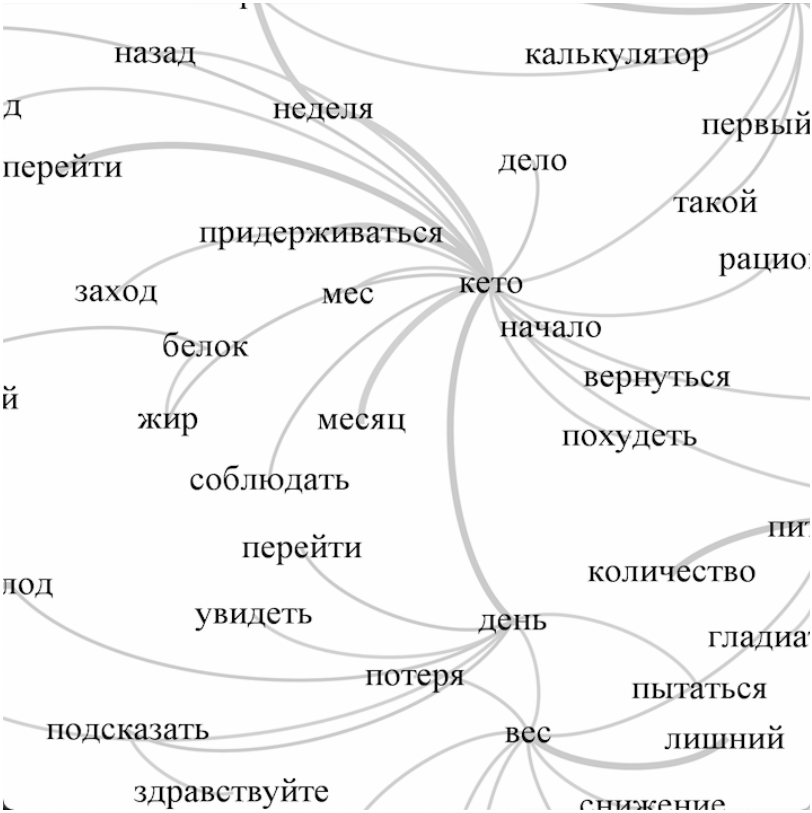



***Рис. 2. Фрагмент поля дискурса здорового питания***

## Дискуссия

В рамках описанной в методологии кодировки были выделены шесть групп агентов дискурса.

1. Худеющие – агенты, которые делятся результатами своего похудения. Они воспринимают здоровое питание как способ эффективно похудеть, привести себя в форму (10,62% комментариев, здесь и далее – от общего объема выборочной совокупности в 3000 комментариев).

2. Кулинары – агенты, которые с интересом делятся рецептами и рационами. Для них здоровое питание – это возможность вкусно приготовить и поесть здоровую пищу (19,01% комментариев).

3. Комментаторы – агенты, которые реагируют на публикации сообщества, критикуют, восхищаются и др. В их представлении здоровое питание – это бренд, организация, собеседник (19,01% комментариев).

4. Лайфстайл – агенты, обсуждающие обычные повседневные привычки и радости. Практика здорового питания не имеет ключевого значения в их речи (17,53% комментариев).

5. Опытные – агенты, которым важно поделиться опытом в освоении здорового питания, возникшими ранее проблемами и путями их решения. Для них здоровое питание уже было важной частью жизни (15,8% комментариев).

6. Члены комьюнити – агенты, активно участвующие в общении с другими членами сообщества, дающие советы, задающие вопросы, поддерживающие других. В их представлении здоровое питание – это общее дело, возможность найти единомышленников или обогатиться новыми знаниями (18,03% комментариев).

Агенты образуют поле данного дискурса (рис. 2). В качестве точек дискурсивной борьбы выступают знаки «кето», «день», «питание», «вес». Эти знаки содержат наибольшее число связей с другими – подсчет производился членами исследовательской группы. Результаты анализа приводятся в табл. – для компактно-

*Таблица*

ПРОАНАЛИЗИРОВАННЫЕ ЗНАКИ НА ПОЛЕ ДИСКУРСА

| Знак | Встречаемые совместно знаки | Агенты, использующие встречаемые совместно знаки |
|---|---|---|
| Кето | Неделя, месяц, год, вернуться, заход, придерживаться, похуделый, похудеть | Опытные |
| | Пропорция, жир, белок, питание | Члены комьюнити |
| День | Первый, третий, пара, следующий, раз, вес | Опытные |
| | Добрый, подсказать | Члены комьюнити |
| | Пытаться, нагрузка | Лайфтсайл |
| Питание | Скидка, спортивный | Лайфстайл |
| | Начинать, низкоуглеводный, кето | Члены комьюнити |
| Вес | Потеря, держать, терять | Худеющие |
| | Встать, потеря | Опытные |

сти приведены только наиболее многочисленно представленные позиции агентов.

Точкой наиболее ожесточенной дискурсивной борьбы оказывается знак «кето» – его используют 5 классов агентов из 6. Рассмотрим только наиболее активные группы агентов в этой борьбе. «Опытные» используют знак «кето» совместно со знаками временных рамок (неделя; месяц; год), а также в контексте возвращения и повторяющейся практики (вернуться; заход; придерживаться и др.). Они также указывают на эффективность диеты в похудении (похуделый; похудеть). «Члены комьюнити» используют знаки для уточнения деталей и нюансов практики «кето» (пропорция; белок; жир; питание), пытаясь точно сформулировать стратегию здорового питания. Итогом борьбы за знак «кето» становится выигрыш «опытных» агентов – знак «кето» обладает наибольшим числом контекстуальных связей в речи «опытных» агентов.

Следующей точкой дискурсивной борьбы является «день». Если агенты класса «опытные» используют данное слово в контексте своего прошлого опыта, а также для обозначения распорядка практики здорового питания (первый; третий; пара; следующий; раз; вес), то «члены комьюнити» больше фокусируются на доброжелательном тоне общения и взаимной поддержке (добрый [день]; подсказать и др.). Для представителей класса «лайфстайл» соблюдение практики здорового питания является тяжелой задачей (пытаться; нагрузка и др.). Победу в данной точке дискурсивной борьбы вновь одерживают «опытные» агенты.

Следующей точкой дискурсивной борьбы является знак «питание». Активнее всего борьбу за данный знак ведут классы агентов «лайфстайл» и «члены комьюнити». Для представителей класса «лайфстайл», важен разумный экономический подход (используя знак «скидка»), они также пытаются связать с практикой здорового питания другие сферы своего интереса (спортивный). Говоря о питании, «члены комьюнити» ищут единомышленников на начальных этапах практики здорового питания (начинать; низкоуглеводный; кето). Именно этот класс побеждает в борьбе за знак «питание».

Последней точкой дискурсивной борьбы является знак «вес». За него борются классы «худеющих» и «опытных». Другие классы участвуют в борьбе несущественно. Для класса «худеющих» «вес» фигурирует в контексте достижения основной цели – похудения (потеря; держать; терять). Агенты из класса «опытных» используют знаки «стоять», «встать» наряду со словом «снижение», что говорит о достижении цели и желании сохранить результат. В результате борьбы данный знак остается за классом «худеющих» агентов.

## Концептуальное картирование

*Методология отбора данных*

Для построения концептуальной карты был отобран пост в сообществе о здоровом питании, под которым была дискуссия, содержащая наибольшее число комментариев для данного сообщества, – 19. Данная дискуссия была выбрана в качестве примера для реализации концептуального картирования. На ее основе создавался свой ориентированный граф – концептуальная карта. Данные, с помощью которых была получена концептуальная карта, содержали следующие характеристики: текст комментария, автор комментария, время его публикации и количество выставленных сообщению лайков.

*Методология анализа*

Процесс построения концептуальной карты можно разделить на несколько этапов. Первым этапом построения является первичная обработка сообщений – естественного языка, – направленная на упрощение текста для последующей машинной обработки. Также анализировалась тональность высказывания. Текст был подвергнут минимальной первичной обработке – приведен к нижнему регистру, очищен от эмодзи и специальных символов. После описанной предобработки из полученного массива данных удалялись пустые комментарии.

Вторым этапом реализации является определение семантической близости между высказываниями участников дискуссии. Определение семантической близости текстов – широко распространенная задача в области обработки естественного языка [16] и заключает в себе две подзадачи: векторное представление слов и непосредственно определение семантической близости. В результате векторного представления слов анализируемый текст трансформируется в вектор. Это позволяет применять к анализу

текстов методы машинного обучения. Векторное представление слов осуществляется с помощью библиотеки sentence_transformers. Выбранный из библиотеки метод SentenceTransformer реализуется на базе предобученной мультиязычной нейронной сети Sentence-BERT (SBERT), позволяющей корректно сравнивать русскоязычные комментарии [17]. После векторной трансформации текстов комментариев необходимо определить семантическую близость каждой пары комментариев. Для этого создается матрица, заполненная нулями, размерности NxN, где N – количество комментариев в дискуссии. К каждой паре комментариев применяется метод cosine_similarity из библиотеки sklearn.metrics.pairwise для расчета значений косинусного расстояния. Семантическая близость между комментариями, таким образом, равна косинусному расстоянию между двумя векторами, образованными при помощи векторного представления этих комментариев нейронной сетью Sentence-BERT.

На третьем этапе определяется эмоциональная окраска комментариев. Данная информация была необходима для определения логической связи между сообщениями дискуссии. Если негативно окрашенный комментарий был семантически близок к комментарию с положительной тональностью, предполагалось, что логическая связь между комментариями строго дизъюнктивна. Нейтральная связь предполагала только наличие причинной связи и т.д. Эмоциональная окраска комментариев дискуссии определялась с помощью библиотеки языка Python dostoevsky. Ее особенностью является обучение модели на корпусе русскоязычных текстов [18]. Модель определяет вероятность принадлежности каждого комментария к одному из пяти классов тональности: neutral (или нейтральный тон), negative (негативный), positive (положительный), а также speech (текст является элементом разговорного языка и не содержит ярко выраженной эмоциональной окраски), skip (данный класс присваивается тексту, если модель не может определить тональность). В рамках построения концеп-

туальных карт с помощью модели каждому комментарию присваивался наиболее вероятный класс из перечисленных выше. Итогом данного этапа является перечень меток эмоциональной окраски сообщений, информация о которых добавляется в изначальный массив комментариев.

Наконец, четвертый этап: на основе исходного массива комментариев, с добавленной информацией об их тональности, а также сконструированной на предыдущем этапе матрицы косинусных расстояний, создается ориентированный граф. В силу ограниченности трудовых ресурсов экспертной группы проекта значение косинусного расстояния, при котором в итоговый ориентированный граф добавляются связи между вершинами, подбирался вручную таким образом, чтобы граф имел не более 20 ребер между вершинами-сообщениями – общего числа ребер графа. При увеличении числа ребер количество анализируемых цепочек комментариев многократно росло, отчего было принято решение ограничиться наиболее связанными комментариями – ядром дискуссии. Такой граф представляет собой концептуальную карту дискуссии, в которой отражен ход проанализированной дискуссии. Его визуальное представление строится посредством программы Gephi. Вершинами графа выступали комментарии дискуссии, а ребра между ними обозначались, если семантически комментарии были достаточно близки друг к другу. Ребра были направлены от более ранних комментариев к более новым, что позволяло сохранить причинно-следственные связи между сообщениями.

## Результаты

В дискуссии явно выделяется слово «кг», сокращенное от «килограмм» (рис. 3). Оно употребляется заметно чаще остальных – 19 раз. Ближайшие по частоте слова – «вес» и «месяц» – употреблены более чем в два раза реже (каждое – по 7 раз). Еще

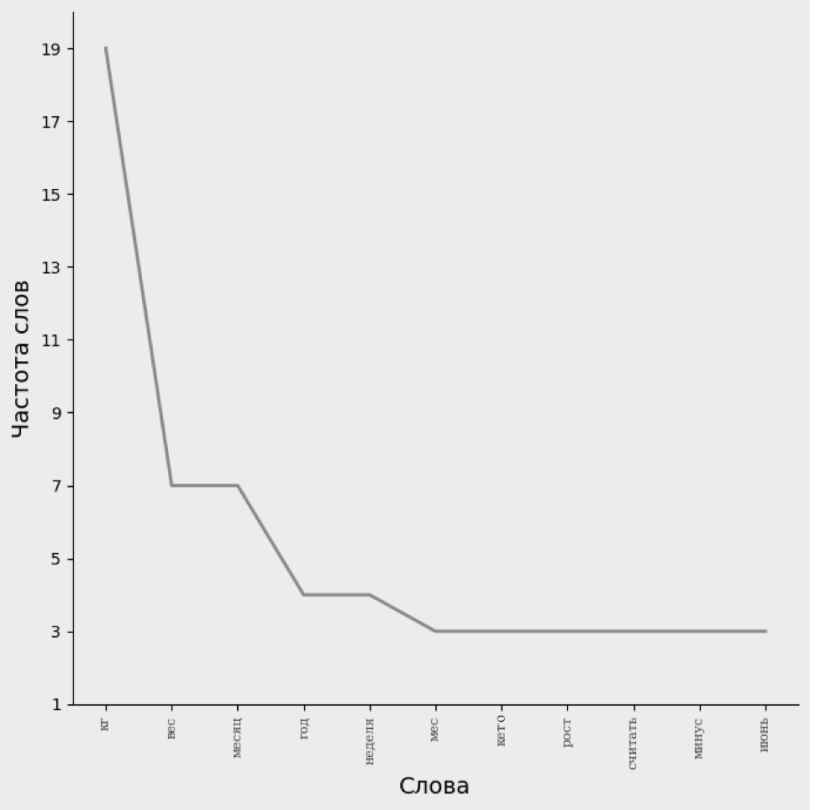



***Рис. 3. Частотное распределение наиболее популярных слов в проанализированной дискуссии***

реже употребляются слова «неделя» и «год» (4 раза). Другое слово, означающее длительность – «мес», сокращение от «месяц», – употребляется только 3 раза. Оставшиеся слова, представленные на рисунке, употреблены в комментариях дважды. Это слова: «кето», «рост», «считать», «минус», «июнь».

На основании данного набора слов можно заключить, что практика здорового питания тесно связана с практикой похуде-

ния: участники сообщества формируют дискуссию вокруг опыта похудения (слова «кг», «вес»), приобретаемого в течение разных отрезков времени (об этом свидетельствуют словосочетания, описывающие различные промежутки времени).

## Анализ концептуальных карт

На рис. 4 представлена концептуальная карта обсуждения в сообществе здорового питания.

Для простоты визуального восприятия представленной коллективной концептуальной карты вершины, не имеющие связей с другими, были предварительно исключены. Так, из 19 исходных комментариев на рисунке представлены только 12. Все связи ориентированы от наиболее ранних к наиболее поздним комментариям, чтобы сохранить причинно-следственную связь и порядок дискуссии. Так, концептуальную карту можно рассматривать как ориентированный несвязный граф, состоящий из трех компонент связности. Вершины связаны 18 ребрами. Общая плотность графа составляет 0,18. Основной темой, раскрываемой в приведенном обсуждении практики здорового питания, является похудение. Данная концептуальная карта фрагментарна. Причем заметно, что обособленные две пары вершин, во-первых, сообщают исключительно факты о количестве сброшенных килограммов за указанный промежуток времени, во-вторых, сильно схожи по структуре.

Перейдем к анализу основного фрагмента обсуждения, в который включается большинство комментариев. Из 12 комментариев, представленных на рис. 4, 8 включены в данный фрагмент и соединены 16 связями. Плотность представленного ориентированного графа составляет 0,57. Данный фрагмент имеет единственную стартовую вершину, которая порождает сразу 5 альтернативных продолжений. Для примера рассмотрим только связи с наибольшим весом со стартовой вершиной цепочки сообщений.

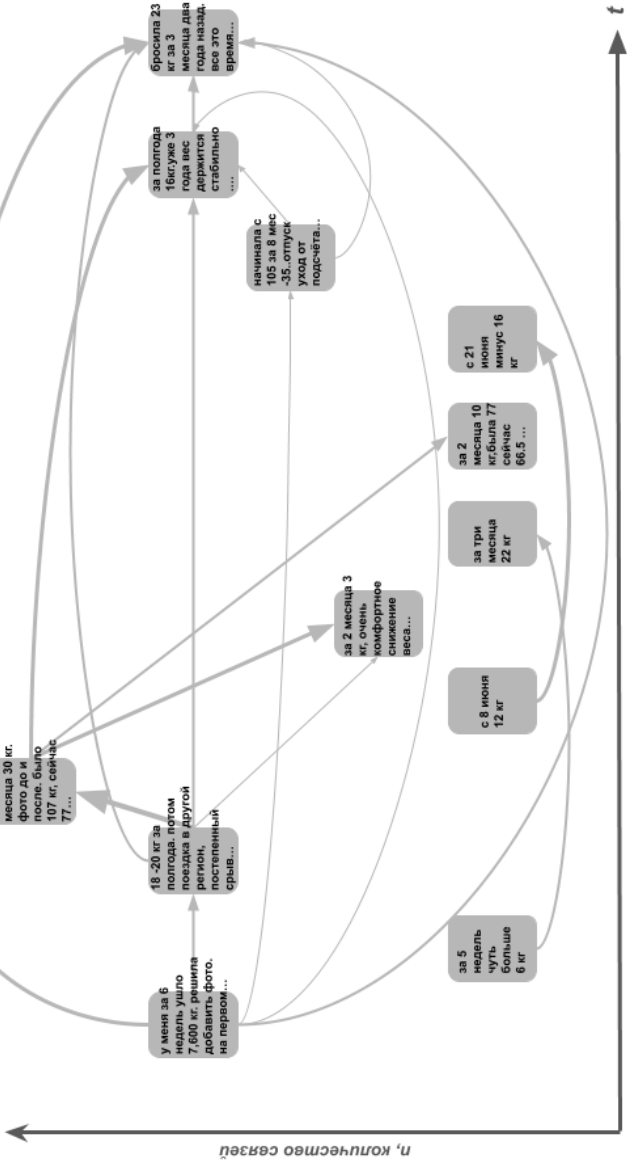



*Рис. 4. Коллективная концептуальная карта дискуссии в сообществе о здоровом питании*

1. ***у меня за 6 недель ушло 7,600 кг. решила добавить фото. на первом 2 недели на кето, на втором – сегодня***
    - ***a.*** *сбросила 23 кг за 3 месяца два года назад. все это время держу вес. три кг туда-сюда гуляет, но это погрешности режима*
    - ***b. за 1 год и 2 месяца 30 кг. фото до и после. было 107 кг, сейчас 77, рост 180 см. я худею очень медленно, сейчас на кето, но уже калории не считаю, считаю только белок и углеводы, жира ем выше нормы. нормы на калькуляторе 149, я под 200 иначе мне голодно. пока прибавки в весе нет, надеюсь так и дальше будет***
        - ***i.*** *за 2 месяца 10 кг,была 77 сейчас 66.5рост 179см, меньше не хочу??*
        - ***ii.*** *за 2 месяца 3 кг, очень комфортное снижение веса и объёмов. но вес был 61 кг, сейчас 57,5–58 при росте 160.*
        - ***iii.*** *(сбросила 23 кг за 3 месяца два года назад. все это время держу вес. три кг туда-сюда гуляет, но это погрешности режима)*
        - ***iv.*** *за полгода 16кг.уже 3 года вес держится стабильно. ккал не считаю, лень. пью воду. убрала мусорные продукты.*
            - **o** *(сбросила 23 кг за 3 месяца два года назад. все это время держу вес. три кг туда-сюда гуляет, но это погрешности режим)*

[…]

Автор *комментария 1* начинает со стандартной для данного обсуждения структуры: делится количеством килограммов и сроком диеты, добавляя фотографии до и после. Автор к*омментария a* тоже делится успехами похудения, добавляя информацию о прошлом опыте и о нынешнем состоянии. Написанный таким образом комментарий сохраняет структуру противопоставления «до»–«после». *Автор комментария b* делает то же самое, подробно описывая свои антропометрические данные (рост, а также вес в начале процесса похудения и на момент написания комментария). Также автор подробно описывает строгость соблюдения диеты.

*Комментарий b* продолжается четырьмя различными сообщениями. При этом *комментарии i, ii* связаны с предыдущей

вершиной сходным образом: авторы добавляют антропометрическое описание. В отличие от *комментария i*, в *комментарии ii* присутствует субъективная характеристика процесса похудения. В *комментарий iii* связь усиливается, так как в *комментариях b* и *iii* присутствуют описания некоторого пренебрежения диетой. Наконец, *комментарий iv* связан с предыдущим комментарием аналогичным описанием пренебрежения диетой. Порожденная этим комментарием цепочка продолжается рассмотренным ранее комментарием и дополнительно анализироваться не будет.

## Дискуссия

Коллективное обсуждение практик здорового питания характеризуется отсутствием выраженного полилога между агентами – они не обсуждают поставленную тему, а, скорее, делятся списком фактов о собственном опыте. Часто этот опыт – некоторые нарушения режима питания, но он не вызывает у агентов никаких особенных эмоций. Так, большинство комментариев среди всех проанализированных в концептуальном картировании имеют нейтральную эмоциональную окраску, а тон остальных модель определить не смогла. Семантически близкие друг к другу сообщения сходны внутренней структурой (срок практикования здорового питания и количество сброшенных килограммов). В дополнение к ней агенты часто упоминают другие антропометрические факты: рост, вес «до» и «после», иногда добавляя фотографии.

## Заключение

Полученные результаты исследования и сделанные на их основе выводы кажутся контринтуитивными. При анализе сообществ здорового питания исследовательская группа ожидала увидеть дискуссии о здоровье и влиянии на здоровье продуктов и пищевых привычек. В действительности же точкой кристалли-

зации дискуссии оказывается похудение. Этот вывод отчетливо прослеживается как в дискурс-анализе, так и в концептуальном картировании. Несмотря на то, что комментарии примерно в равной степени распределены между представителями различных групп агентов, основная дискуссия как борьба дискурсивных позиций развивается между пользователями класса «опытные» и теми, кто только начинает свой путь похудения. Пользователи класса «опытные» чаще остальных участвуют в борьбе за ключевые знаки дискурса. Они же стараются сместить акцент с похудения как процесса изменения веса к его удержанию. Они чаще рассказывают, что вес не теряется, а стоит, килограммы не уходят, а стоят. Этот же вывод оказывается справедлив для представленной в статье концептуальной карты. Здесь важно указать на существующее терминологическое противоречие выбранных методов: агенты могут быть внешне доброжелательны (что может быть обнаружено в результатах концептуального картирования), но имплицитно они будут бороться именно за свою единственно верную интерпретацию того или иного знака. Оттого без явного противоречия с выводами дискурс-анализа мы можем наблюдать, что явное противостояние пользователей не встречается в анализируемой дискуссии. Как правило, пользователи фактологически описывают, каких результатов они достигли в похудении. Они строго связывают здоровье с утратой веса. Вес понимается как нечто ненужное, что следовало бы потерять. Такое понимание веса – как элемента фактологического изложения объемов его утраты – приводится в работе Е. Костяшкиной, где указывается, применительно к медиа-дискурсу, что потеря веса начинает восприниматься как новостной факт. Как и в нашем исследовании, потеря веса в работе Е. Костяшкиной рассматривается в контексте медиадискурса здоровья [19].

В представленной в качестве примера в статье дискуссии существует явный консенсус в понимании здорового питания как процесса, ориентированного на похудение с помощью кетодиеты.

Разногласие происходит в обозначении значимости элементов похудения – что важнее: потерять вес или удержать его? Также разногласия существуют в оценках сложности кетодиеты. Однако концептуальное картирование позволяет получить информацию о принимаемой форме разногласия в конкретной анализируемой дискуссии. Пользователи не вступают в открытый конфликт, а склонны через высказывания о себе популяризировать свою позицию и поддерживать позиции других пользователей, сходные с их собственной. Важно упомянуть, что речь идет о наиболее сильно семантически связанных комментариях. Анализ с включением нерассмотренных комментариев позволил бы получить дополнительные выводы.

В настоящей статье представлено первое приближение пути реализации методов концептуального картирования и дискурс-анализа применительно к анализу онлайн-дискуссий. Опишем подробнее ограничения методов и, соответственно, исследования. Сопоставление результатов дискурс-анализа и концептуального картирования видится небесспорным. С одной стороны, оба метода направлены на единый объект и анализируют один и тот же эмпирический материал – все слова из комментариев концептуальной карты есть на поле дискурса (кроме удаленных стоп-слов). Отсюда методы описывают единую дискуссию о здоровом питании с разных сторон. Однако концептуальное картирование описано только для дискуссии, приводимой в статье в качестве примера. В других дискуссиях в сообществе знаки, выделенные в дискурс-анализе, могут использоваться иначе. Важным ограничением оказывается роль ботов – алгоритмов искусственного интеллекта, автоматически размещающих рекламные и иные сообщения в социальных сетях, нередко выдающих себя за живых пользователей. Эта проблема активно обсуждается в работах С. Брэдшоу и П. Говарда [20], Дж. Прэра [21]. С. Джанвеккио [22] и других авторов. В нашем исследовании это позволяет задать важный вопрос – могут ли эти боты быть субъектами коллективной концептуальной карты,

можем ли мы говорить о разуме таких ботов? Технооптимисты и технопессимисты дают разный ответ на этот вопрос, пишет Д. МакДермотт: у первых искусственный интеллект либо обретет сознание, либо обладает им уже сейчас, у технопессимистов, соответственно, наоборот [23]. Также ограничением выступает выбор авторами уровня анализа в концептуальном картировании: авторы не обобщают высказывания до уровня концептов, что реализовано во многих интерпретациях метода, а остаются на уровне конкретных высказываний. С одной стороны, это позволяет использовать методы поиска семантической близости для обнаружения связи в концептуальной карте и более детально анализировать содержание комментариев, но, с другой стороны, из вида упускается более высокоуровневое обобщение результатов.

Авторы представили в данной статье адаптацию и применение метода дискурс-анализа в интерпретации Лакло и Муфф, а также метода концептуального картирования в интерпретации Трочима для анализа онлайн-дискуссий. Оценка валидности разработанных методов видится задачей для дальнейших исследований.

## ЛИТЕРАТУРА

1. Network analysis reveals open forums and echo chambers in social media discussions of climate change / H.T. Williams, J.R. McMurray, T. Kurz, F.H. Lambert // Global environmental change. 2015. № 32. P. 126−138. DOI: 10.1016/j.gloenvcha.2015.03.006.

2. Using social media to monitor mental health discussions − evidence from Twitter / C. McClellan, M. M Ali, R. Mutter [et al.] // Journal of the American Medical Informatics Association. 2017. Vol. 24, № 3. P. 496–502. DOI: 10.1093/jamia/ocw133.

3. *Trochim W.M., McLinden D.* Introduction to a special issue on concept mapping // Evaluation and program planning. 2017. № 60. P. 166–175. DOI: 10.1016/j.evalprogplan.2016.10.006.

4. *Тумский С.В.* Концептуальное картирование как междисциплинарный метод анализа // Когнитивные исследования языка. 2014. № 17. С. 182–188. EDN: SALHET.

5. *Trochim W.M.* An introduction to concept mapping for planning and evaluation // Evaluation and program planning. 1989. Vol. 12, №1. P. 1–16. DOI: 10.1016/0149-7189(89)90016-5.

6. *Абрамова Н.В., Николаева Ю.В.* Построение концептуальных карт как метод повышения валидности результатов оценочного исследования // Социология: Методология, методы, математические модели (Социология: 4М). 2006. № 23. С. 83–99. EDN: KVKIXH.

7. *Kane M., Trochim W.M.* Concept mapping for planning and evaluation. CA: Sage Publications, 2007. 216 p. ISBN: 1412940273, 9781412940276.

8. *Lee S., Chun J.* Conceptualizing the impacts of cyberbullying victimization among Korean male adolescents // Children and Youth Services Review. 2020. № 117. art. 105275. DOI: 10.1016/j.childyouth.2020.105275.

9. *Scourfield P.* A Critical Reflection on the Involvement of "Experts by Experience" in Inspections // The British Journal of Social Work. 2010. Vol. 40, № 6. P. 1890–1907. DOI: 10.1093/bjsw/bcp119.

10. *Йоргенсен М.В., Филлипс Л.Дж.* Дискурс-анализ. Теория и метод / Пер. с англ.; 2-е изд., испр. Харьков: Гуманитарный центр, 2008. 352 с. ISBN: 0-7619-7112-2.

11. J*ones R.H., Chik A., Hafner C.A.* Discourse and digital practices: Doing discourse analysis in the digital age. London: Taylor&Francis, 2015. 262 p. ISBN: 1317537009, 9781317537007.

12. *Carpentier N., De Cleen B.* Bringing discourse theory into media studies: The applicability of discourse theoretical analysis (DTA) for the study of media practises and discourses // Journal of language and politics. 2007. Vol. 6, №. 2. P. 265–293. DOI:10.1075/jlp.6.2.08car.

13. *Игнатьева О.А.* Дискурс-анализ политических суждений в контексте цифровизации // Политическая экспертиза: ПОЛИТЭКС. 2021. Т. 17, № 3. С. 259–272. DOI: 10.21638/spbu23.2021.303. EDN: GBFXQD.

14. *Haditaghi J., Hassasskhah J., Sorahi M.A.* A network-based approach for discourse analysis from Laclau and Mouffe's perspectives // Journal of Computer-Assisted Linguistic Research. 2020. Vol. 4. P. 1–22. DOI: 10.4995/jclr.2020.12105.

15. *Bakumov P.* An Alternative Model for the Operationalization of Discourse Theory of Laclau and Mouffe // Laboratorium: Russian Review of Social Research. 2022. Vol. 14, №. 3. P. 119–134. DOI: 10.25285/2078-1938-2022-14-3-119-134. EDN: SUGLDK.

16. *Reimers N., Gurevych I.* Sentence-bert: Sentence embeddings using siamese bert-networks // Cornwall University [site]. 27.08.2019. URL: https://arxiv.org/abs/1908.10084 (date of access: 05.07.2023).

17. *Vatolin A.S., Smirnova E.Y., Shkarin S.S.* Russian News Similarity Detection with SBERT: Pre-training and fine-tuning // Komp'juternaja Lingvistika i Intellektual'nye Tehnologii. 2021. № 20. P. 692–697. DOI: 10.28995/2075-7182-2021-20-692-697. EDN: NKSZTA.

18. GitHub – bureaucratic-labs/dostoevsky: Sentiment analysis library for russian language // GitHub [site]. URL: https://github.com/bureaucratic-labs/dostoevsky (date of access: 05.07.2023).

19. *Костяшина Е.А.* Дискурсивная организация картины мира научно-популярного медицинского журнала // Вестник Томского государственного университета. Филология. 2010. Т. 3, №11. С. 41–47. EDN: NEFMVJ.

20. *Bradshaw S., Howard P.* Troops, trolls and troublemakers: A global inventory of organized social media manipulation // Computational Propaganda Research Project. 2017. № 12. P. 1–37.

21. *Prier J.* Commanding the trend: Social media as information warfare. In Information warfare in the age of cyber conflict. London: Routledge, 2020. P. 88–113. ISBN: 9780429470509.

22. Measurement and classification of humans and bots in internet chat / S. Gianvecchio, M. Xie, Z. Wu, H. Wang // USENIX security symposium. 2008. № 17. P. 155–170.

23. McDermott D. Artificial intelligence and consciousness // The Cambridge handbook of consciousness. Cambridge: Cambridge University Press, 2007. P. 117–150. ISBN: 113946406X, 9781139464062.

***Tkach Sergey,***
*Sociologist, Center for Applied Sociology, St Petersburg University, St. Petersburg, Russia*, s.tkach@spbu.ru

***Vorobyova Polina D.,***
*Sociologist, Center for Applied Sociology, St Petersburg University, St. Petersburg, Russia*, st098355@student.spbu.ru

***Rusakova Maya M.,***
*Candidate of Sociological Sciences, Director, Center for Applied Sociology, St Petersburg University, St. Petersburg, Russia,* m.rusakova@spbu.ru

**Experience of implementing discourse analysis and conceptual mapping of healthy eating communities**

The article presents the experience of implementing discourse analysis methods as interpreted by E. Laclau and C. Mouffe and the concept mapping method as interpreted by W. Trochim through the network analysis technique using the example of healthy eating. The result of the analysis is a graph that makes it possible to highlight the struggle of discourse agents for the meaning of key discursive signs, as well as the conceptual map of participants in an online discussion on controversial issues within the topic discussed in the article as an example. The empirical basis for the discourse analysis was 3 000 collected comments in four communities about healthy eating on the social networks VKontakte and Odnoklassniki. The modified versions of the methods of concept mapping and discourse analysis were adapted for the analysis of online discussions. Subsequent validation of the methods seems to be a promising direction for further research. The proposed designs have a number of limitations, which are discussed in the article.
*Keywords*: discourse analysis, concept mapping, graph analysis, sentiment analysis, natural language analysis

**References**

1. Williams H.T., McMurray J.R., Kurz T., Lambert, F. H. Network analysis reveals open forums and echo chambers in social media discussions of climate change, Lambert, *Global environmental change*, 2015, no. 32, p. 126−138. DOI: 10.1016/j.gloenvcha.2015.03.006.
2. McClellan C., Ali M.M., Mutter R., Kroutil L., Landwehr J. Using social media to monitor mental health discussions − evidence from Twitter,

J*ournal of the American Medical Informatics Association*, 2017, vol. 24, no. 3, p. 496–502. DOI: 10.1093/jamia/ocw133.
3. Trochim W.M., McLinden D. Introduction to a special issue on concept mapping, *Evaluation and program planning*, 2017, no. 60, p. 166–175. DOI: 10.1016/j.evalprogplan.2016.10.006.
4. Tumsky S.V. Concept mapping as an interdisciplinary method of analysis (in Russian), *Kognitivnye issledovanija jazyka (Cognitive language research),* 2014, no. 17, p. 182–188.
5. Trochim W.M. An introduction to concept mapping for planning and evaluation, *Evaluation and program planning*, 1989, vol. 12, no. 1, p. 1–16. DOI: 10.1016/0149-7189(89)90016-5.
6. Abramova N.V., Nikolaeva Yu.V. Constructing concept maps as a method for increasing the validity of evaluation research results (in Russian), *Sotsiologiya 4M (Sociology: methodology, methods, mathematical modeling),* 2006, no. 23, p. 83–99.
7. Kane M., Trochim W.M. *Concept mapping for planning and evaluation.* CA: Sage Publications, 2007. 216 p. ISBN: 1412940273, 9781412940276.
8. Lee S., Chun J. Conceptualizing the impacts of cyberbullying victimization among Korean male adolescents, *Children and Youth Services Review*, 2020, no. 117, art. 105275. DOI: 10.1016/j.childyouth.2020.105275.
9. Scourfield P. A Critical Reflection on the Involvement of "Experts by Experience" in Inspections, *The British Journal of Social Work*, 2010, vol. 40, no. 6, p. 1890–1907. DOI: 10.1093/bjsw/bcp119.
10. Jorgensen, M. W., Phillips L. *Discourse analysis as theory and method* (in Russian); 2nd ed., rev. Kharkov: Gumanitarnyj centr, 2008. 352 p. ISBN: 0-7619-7112-2.
11. Jones R.H., Chik A., Hafner C.A. *Discourse and digital practices: Doing discourse analysis in the digital age*. London: Taylor&Francis, 2015. 262 p. ISBN: 1317537009, 9781317537007.
12. Carpentier N., De Cleen B. Bringing discourse theory into media studies: The applicability of discourse theoretical analysis (DTA) for the study of media practises and discourses, *Journal of language and politics*, 2007, vol. 6, no. 2, p. 265–293. DOI:10.1075/jlp.6.2.08car.
13. Ignatieva O.A. Discourse analysis of political judgments in the context of digitalization (in Russian), *Politicheskaja jekspertiza: POLITJeKS*

*(Political expertise: POLITEX)*, 2021, vol. 17, no. 3, p. 259–272. DOI: 10.21638/spbu23.2021.303.

14. Haditaghi J., Hassasskhah J., Sorahi M.A. A network-based approach for discourse analysis from Laclau and Mouffe's perspectives, *Journal of Computer-Assisted Linguistic Research*, 2020, vol. 4, p. 1–22. DOI: 10.4995/jclr.2020.12105.
15. Bakumov P. An Alternative Model for the Operationalization of Discourse Theory of Laclau and Mouffe, *Laboratorium: Russian Review of Social Research,* 2022, vol. 14, no. 3, p. 119–134. DOI: 10.25285/2078-1938-2022-14-3-119-134.
16. Reimers N., Gurevych I. Sentence-bert: Sentence embeddings using siamese bert-networks, *Cornwall University* [site]. 27.08.2019. URL: https://arxiv.org/abs/1908.10084 (date of access: 05.07.2023).
17. Vatolin A.S., Smirnova E.Y., Shkarin S.S. Russian News Similarity Detection with SBERT: Pre-training and fine-tuning (in Russian), *Komp'juternaja Lingvistika i Intellektual'nye Tehnologii,* 2021, no. 20, p. 692–697. DOI: 10.28995/2075-7182-2021-20-692-697. EDN: NKSZTA.
18. GitHub – bureaucratic-labs/dostoevsky: Sentiment analysis library for Russian language, *GitHub* [site]. URL: https://github.com/bureaucratic-labs/dostoevsky (date of access: 05.07.2023).
19. Kostyashina E.A. Discursive organization of the world picture of a popular scientific medical journal (in Russian), *Vestnik Tomskogo gosudarstvennogo universiteta. Filologija* (*Bulletin of Tomsk State University). Philology*, 2010, vol. 3, no. 11, p. 41–47.
20. Bradshaw S., Howard P. Troops, trolls and troublemakers: A global inventory of organized social media manipulation, *Computational Propaganda Research Project*, 2017, no. 12, p. 1–37.
21. Prier J. *Commanding the trend: Social media as information warfare. In Information warfare in the age of cyber conflict*. London: Routledge, 2020. P. 88–113. ISBN: 9780429470509.
22. Gianvecchio S., Xie M., Wu Z., Wang, H. Measurement and classification of humans and bots in internet chat, *USENIX security symposium,* 2008, no. 17, p. 155–170.
23. McDermott D. *Artificial intelligence and consciousness, The Cambridge handbook of consciousness.* Cambridge: Cambridge University Press, 2007. P. 117–150. ISBN: 113946406X, 9781139464062.

DOI: 10.19181/4m.2023.32.1.5
EDN: CKAHLQ

О.Р. Чепьюк, О.Ю. Ангелова, А.Л. Сочков, Т.О. Подольская
(*Нижний Новгород*)

# ТИПОЛОГИЗАЦИЯ ПРОФЕССИОНАЛЬНЫХ ТРАЕКТОРИЙ ОДАРЕННЫХ ЛИЧНОСТЕЙ С ПОМОЩЬЮ НЕЙРОСЕТЕВОГО АНАЛИЗА[1]

На основе массива данных (100 биографий), сформированного авторами по результатам контент-анализа биографического материала о выдающихся ученых XIX и XX вв. в гуманитарной и естественно-научных сферах, проведена кластеризация профессиональных траекторий одаренных личностей. Методом кластеризации стал нейросетевой анализ на основе самоорганизующихся карт Кохонена. Сами траектории были сформированы в рамках поведенческой модели линейно-стадиального подхода в исследовании жизненных циклов. В рамках этого подхода карьера и профессиональная самореализация человека понимаются как

**Ольга Ростиславовна Чепьюк** – доктор философских наук, профессор кафедры управления человеческими ресурсами, Национальный исследовательский Нижегородский государственный университет им. Н.И. Лобачевского, Нижний Новгород, Россия. Email: chepyuko@yandex.ru.
**Ольга Юрьевна Ангелова** – кандидат экономических наук, доцент кафедры информационных технологий и инструментальных методов в экономике, Национальный исследовательский Нижегородский государственный университет им. Н.И. Лобачевского, Нижний Новгород, Россия. Email: oangelova@mail.ru.
**Андрей Львович Сочков** – кандидат технических наук, доцент кафедры информационных технологий и инструментальных методов в экономике, Национальный исследовательский Нижегородский государственный университет им. Н.И. Лобачевского, Нижний Новгород, Россия. Email: sochkov@iee.unn.ru.
**Татьяна Олеговна Подольская** – кандидат социологический наук, доцент кафедры управления человеческими ресурсами, Национальный исследовательский Нижегородский государственный университет им. Н.И. Лобачевского, Нижний Новгород, Россия. Email: podolskaya@iee.unn.ru.

последовательность этапов эволюции человека, фиксированных в порядке наступления. Каждый из этапов был закодирован, а биографии преобразованы в систему векторов. В свою очередь задача кластеризации заключалась в разбиении массива объектов из сотни векторов на типовые группы, имеющие несколько вещественных интервальных координат. Критериями качества кластеризации стали показатели минимальной суммы ошибок квантования, а также коэффициент силуэта. По итогам исследования были выделены и интерпретированы семь профессиональных траекторий одаренных личностей. Анализ траекторий проводился с точки зрения скорости достижения успеха (среднего возраста успеха) и тех факторов и условий жизненного пути, которые могли повлиять на более быстрое или медленное достижение профессиональных целей и самореализацию. На этом примере были показаны возможности и ограничения использования нейросетевого анализа для решения сходных исследовательских задач, в особенности когда требуется работать со сложными формами кластеров и находить их оптимальное число.

*Ключевые слова:* нейросетевой анализ, одаренность, одаренная личность, профессиональная траектория, машинное обучение, нейронная сеть, «социальный лифт», карты Кохонена

## Введение

Управление талантами и поддержка социальных лифтов декларируются как одна из важнейших задач для государства (макроуровень)[1], корпораций (мезоуровень) и отдельной личности (микроуровень). Изучение возможностей для самореализации одаренной личности находится на стыке наук: педагогики, психо-

[1] Создание в России «возможностей для самореализации и развития талантов» является одной из приоритетных целей развития Российской Федерации до 2030 г. (национальных целей). См.: Указ Президента РФ от 21 июля 2020 г. № 474 «О национальных целях развития Российской Федерации на период до 2030 года» // Официальное опубликование правовых актов [site]. URL: http://publication.pravo.gov.ru/document/0001202405070015 (дата обращения: 20.01.2023).

логии, социологии, философии, экономики. На фундаментальном уровне вопрос может быть сформулирован следующим образом: каким образом образование, рынок труда и в целом хозяйственный организм обеспечивают человека возможностями и условиями для развития его задатков и самореализации в профессиональном плане? На прикладном уровне задача заключается в выявлении повторяющихся паттернов (траекторий), описывающих социальные условия, в которых личность достигает успеха в кратчайшие сроки. Появление многочисленных структурированных источников биографической информации (WikiData, Pantheon), а также возможностей их автоматической обработки дает возможность проводить такие исследования: имеются значительные объемы цифровых следов, отражающих основные этапы жизненного пути человека[1]. Методологи отмечают, что ученые, по сравнению с коммерческими компаниями, предпринимают осторожные попытки в применении численных методов анализа биографий [1]. В то же время специалисты в сфере компьютерного анализа активно создают структурированные базы биографий [2; 3; 4; 5], а также инструменты для визуализации [6; 7].

Российская и советская научные школы традиционно фокусировались на изучении жизненного пути человека. Так, Л.С. Выготский делал акцент на процессах становления личности: человек – субъект жизни, а только затем – субъект поведения. В одаренности важно не столько наличие задатков, сколько их развитие и применение [8]. Исторические исследования одаренности, начиная с работ Ф. Гальтона в XIX в. и продолжая трехкольцевой моделью Дж. Рензулли, подчеркивают важность так называемых «земных» факторов. В России вклад в исследование одаренности внесли научные школы А.Н. Леонтьева [9], В.С. Мерлина, Я.А. Пономарева, Б.М. Теплова, В.Д. Шадрикова [10]. Особое внимание в контексте

[1] См., напр.: Wikidata [site]. URL: https://www.wikidata.org/wiki/Wikidata:Main_Page (date of access: 20.01.2023).

исследования заслуживают работы американского исследователя Д. Галенсона [11; 12]. Он выделил два паттерна успеха: *экспериментальные* инноваторы достигают результатов методом проб и ошибок, на закате профессиональной жизни; *концептуальные* ученые и деятели искусства совершают прорыв на самых ранних этапах карьеры.

Опираясь на традицию российской научной школы, главной задачей исследования стал поиск новых методов, которые бы позволили изучать процессы профессионального становления личности. Ключевым критерием стал срок достижения успеха у одаренного с учетом социальных условий его жизни (стимулов и препятствий к успеху).

## Обоснование метода

На методологическом уровне задача выявления устойчивых (повторяющихся) траекторий развития одаренной личности связана с вопросами кластеризации. Кластеризация используется в современной социологической науке для группировки объектов или наблюдений на основе их сходства. Этот метод помогает выявлять паттерны и структуры в данных, что может быть полезным при изучении траекторий одаренных. Тема кластеризации в социологическом исследовании является одной из часто обсуждаемых: неоднократно описаны как достоинства, так и недостатки различных методов, в том числе метода *k*-средних, анализа латентных профилей, метода пороговых значений [1; 13], иерархической кластеризации [14]. В нашем исследовании применялся метод кластеризации на основе нейросетевого анализа. В отличие от методов иерархической кластеризации, методы на основе нейросетей и машинного обучения могут автоматически извлекать более сложные паттерны в данных. В отличие от метода *k*-средних, они могут учитывать более сложные формы кластеров и не требуют задания числа кластеров. Среди недостатков следует

отметить, что методы на основе нейросетей могут иметь проблемы с выбросами и шумом, если не настроены правильно, они требуют большего объема данных и вычислительных ресурсов для обучения модели [15; 16].

Отметим, что кластеризация для определения типовых траекторий достижений успеха одаренных личностей является более сложной задачей, чем традиционно ставятся исследователями жизненного пути: например, их часто интересует, в каких высших школах работали Нобелевские лауреаты [17], насколько значим фактор мобильности в карьере [18]. Весьма популярными являются вычисления среднего «возраста успеха» [7; 12; 19; 20]. Отдельный интерес представляют биографии спортсменов: их профессиональные достижения легко поддаются оцифровке [21; 22]. Стоит отметить, что исследование социокультурного контекста хотя и декларируется, остается преимущественно частью теоретических размышлений на эту тему [23]. Налицо недостаток поведенческих моделей и новых методов социологического исследования, которые могли бы стать опорными для количественного анализа профессиональных траекторий.

## *Описание этапа сбора данных и контент-анализа*

В выборку исследования вошли биографии 100 выдающихся ученых, чей расцвет профессиональной карьеры приходился на XIX и XX вв. Квотирование выборки осуществлялось по сферам научного знания: 45 человек естественно-научного профиля, 55 человек – гуманитарного. Распределение квот в выборке было сделано на основе двух допущений. Во-первых, оно отражает текущее распределение исследовательских интересов в академическом сообществе, где гуманитарные науки часто включают междисциплинарные направления. Во-вторых, такое распределение позволяет исследовать специфические траектории развития

карьеры в гуманитарных науках, которые исторически меньше исследовались по сравнению с естественными. Это дает возможность оценить различия и сходства в траекториях одаренности. Отметим, что для соблюдения валидности в выборку были включены персоналии, входящие не менее чем в три рейтинга международных и российских издательств, в том числе: американского журнала «Time», «Британской энциклопедии», российской серии научно-популярных книг «Жизнь замечательных людей», «Большой советской энциклопедии», «Большой российской энциклопедии». Для справочной навигации по биографиям использовалась информация из русскоязычной Википедии.

Работа коллектива была разделена на два этапа. На первом этапе была поставлена задача определить время успеха одаренной личности. Годом успеха признавалась публикация (обнародование) первого из значимых (программных) произведений (совершенных открытий), признаваемых обществом ключевым достижением одаренной личности в соответствии с ее доминантой оригинальности. Например, для Д.И. Менделеева, несмотря на его разносторонние увлечения, таким произведением остается его «Таблица периодических элементов» (1869 г.). На втором этапе с помощью контент-анализа проводилась первичная обработка смысловых единиц биографий, которые характеризовали различные модели поведения одаренного относительно выбора профессионального пути. Эти условия (категории) были сгруппированы в соответствии с линейно-стадиальным подходом к описанию жизненного пути (рис. 1).

Из биографий выделялись такие кодовые фразы, которые могли охарактеризовать условия выбора. Так как выборка состояла из биографий ученых, значимыми условиями признавались те, в которых одаренный имел возможность получить образование и занять сильную профессиональную позицию. Образование и сфера карьерной реализации должны были соответствовать доминанте оригинальности ученого и профилю его будущих достижений.

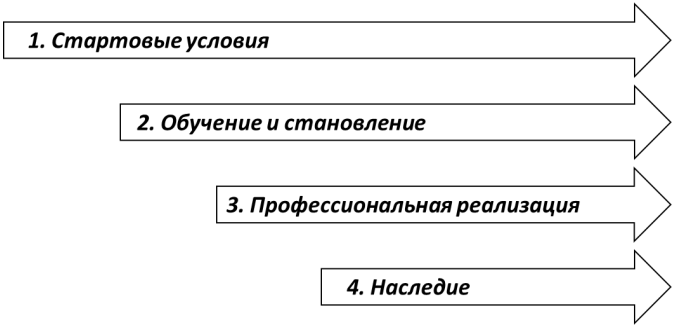



***Рис. 1. Линейно-стадиальная модель жизненного пути***

На первом этапе исследования мы определили, что основными факторами, влияющими на успех одаренной личности, являются условия, создаваемые семьей. Это заключение было сделано на основе анализа биографических данных, где особое внимание уделялось двум аспектам: уровню финансовой обеспеченности семьи и уровню образования и мотивации к обучению и развитию ближайшего круга, в первую очередь родителей. Мы провели систематический контент-анализ биографий, в ходе которого выявили и категоризировали эти факторы как ключевые. Было установлено, что сочетание этих условий создает благоприятную среду для одаренных личностей, позволяя им более свободно выбирать образовательные учреждения и направление профессиональной самореализации. Это, в свою очередь, способствует развитию их потенциала (рис. 2А). Например, из биографии Нильса Бора мы узнаем, что он «*родился в семье Христиана Бора, дважды кандидата на Нобелевскую премию, и Эллен Адлер, дочери весьма влиятельного банкира*». Такое содержание позволяет отнести условия его жизни к ситуации 1.3 (рис. 2А).

На втором этапе «Обучение и становление» к значимым условиям были отнесены наличие и профильность образования (см. рис. 2Б). Здесь учитывалось соответствие сферы достижений ученого (доминанты одаренности) направлению его образова-


***Рис. 2А. Матрица сочетания условий на стартовом этапе***

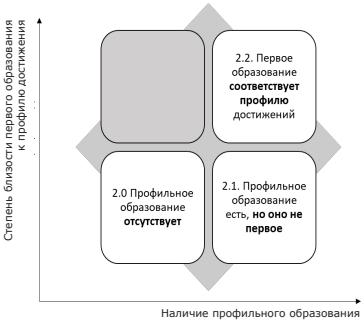



***Рис. 2Б. Матрица сочетания условий на этапе обучения***

ния – как первого (ранняя профессиональная ориентация), так и последующих (осознанный выбор для карьеры). Так, из биографии В. Даля мы узнаем, что он «*после нескольких лет службы на флоте поступил в Дерптский университет на медицинский факультет*». Однако доминанта оригинальности его достижений лежит в области лексикографии и лингвистики. Следовательно, условия на втором этапе могут быть отнесены к ситуации 2.0 (рис. 2Б). Иная ситуация была у Н. Бора. Он «*успешно изучал физику, химию и математику в том же вузе, где преподавал его отец*», поэтому его ситуация кодируется как 2.2 (рис. 2Б).

На третьем этапе – «Профессиональная самореализация» – условиями, определяющими модель поведения, стали степень инновационности сферы приложения усилий одаренного, а также баланс между ролью коллектива (в том числе сильного влияния наставника и научной школы) и личной оригинальной позицией одаренного (рис. 3).

Под степенью инновационности понимается не столько новизна результатов, сколько осознанный или случайный выбор сферы приложения творческих усилий одаренного. Например, из биографии В. Даля узнаем, что на его замыслы и творчество

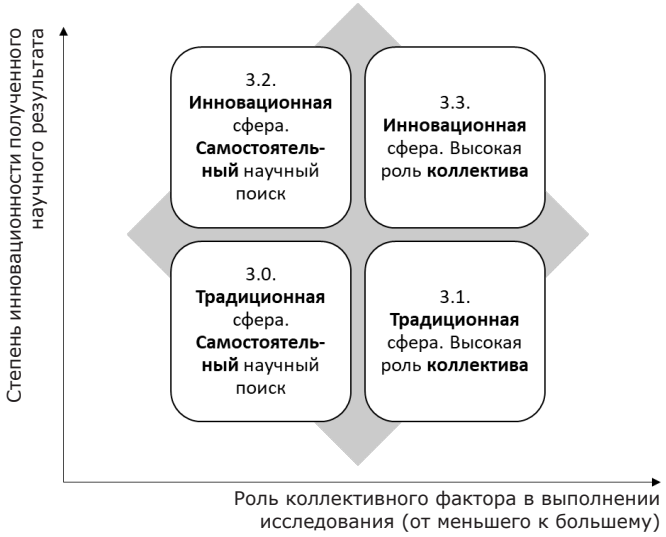



***Рис. 3. Матрица сочетания условий на этапе «Профессиональная реализация»***

серьезное влияние оказал А.С. Пушкин. Такие условия могут быть закодированы как ситуация 3.1 (рис. 3): работа в традиционной сфере при существенном влиянии наставника и его окружения.

Процедура кодировки биографических данных проводилась в два этапа. На первом этапе был осуществлен контент-анализ биографий, выделены данные по ключевым стадиям жизненного пути ученых. Эти данные включали как категориальные переменные (например, ФИО, сектор науки), так и количественные (скорость успеха в годах). На втором этапе была проведена кластеризация с использованием карт Кохонена, где каждый ученый был представлен вектором из шести координат. Эти координаты включали как номинальные, так и преобразованные в интервальные вещественные переменные (например, оценки стартовых условий, условий обучения и профессиональной реализации). Оценки этих факторов были определены экспертным путем на

основе данных первого этапа анализа. Результаты кластеризации использовались для анализа распределения ученых по различным категориям. Вопросы о неоднозначности данных и их репрезентативности решались привлечением нескольких экспертов для усреднения оценок и получения более объективных результатов. Отметим, что в исследовании на данном этапе не рассматривался заключительный этап жизненного цикла («Наследие», рис. 1). Он располагается за пределами жизни одаренной личности и, по сути, не имеет ограничений во времени.

## Метод нейросетевого кластерного (типологического) анализа

В рамках исследования стояла задача выявления типологии профессиональных траекторий одаренных личностей, основываясь на данных, собранных через контент-анализ биографий 100 выдающихся ученых. Целью было определить общие группы с аналогичными профессиональными путями и исследовать их уникальные черты. Для достижения этой цели мы выбрали метод, который уже успешно применялся в смежных областях социально-экономического исследования [24; 25; 26; 27], где он демонстрировал эффективность в типологизации сложных социальных феноменов. Такая задача формулируется следующим образом: как разбить на типовые группы массив объектов из сотни векторов, имеющих несколько вещественных интервальных координат. Отметим, что этот метод подходит для работы с большими объемами многомерных данных и не требует предварительных предположений о структуре данных. Это делает его удобным для анализа биографических данных, структура которых заранее неизвестна.

Для количественного описания траектории отдельно взятого ученого использована система из трех показателей, обоснованная выше поведенческой моделью и представленная вектором (x_j1,x_j2,x_j3). Координаты этого вектора соотносятся с этапами

жизненного пути одаренной личности, которые были представлены на рис. 1. Первая координата $x_{j1}$ отражает уровень стартовых условий ученого и может принимать числовые значения на вещественном отрезке (1; 4), причем единица соответствует трудным стартовым условиям, когда родители имеют невысокий уровень образования и финансового достатка, а 4 – самым благоприятным условиям, когда семья образована и имеет высокий уровень финансовой состоятельности.

Вторая координата $x_{j2}$ связана с этапом обучения и становления ученого, который проиллюстрирован на рис. 3. Эта координата в общем случае может принимать порядковые значения 1, 2 и 3 на вещественном отрезке (1; 3), причем единица соответствует случаю отсутствия профильного высшего образования в сфере успеха одаренной личности, а 3 – случаю, когда ученый добился успеха в сфере своего первого высшего профессионального образования, то есть, другими словами, рост числового значения от 1 до 3 отражает все более быструю и эффективную профессиональную ориентацию одаренной личности.

Третья координата $x_{j3}$ отражает условия третьего этапа профессиональной реализации ученого, который проиллюстрирован на рис. 4. Она может принимать порядковые значения на отрезке (1; 4), причем единица соответствует случаю самостоятельного поиска ученого по одному из традиционных направлений науки, а 4 – случаю его работы в рамках научной школы, разрабатывающей инновационные научные направления. Таким образом, рост численного значения этой координаты отражает уровень инновационности исследований одаренной личности и роль коллектива, командной работы в достижении успеха (в том числе наличие научной школы, наставника).

Кроме этих трех основных координат каждый вектор имеет три информационных показателя (эти показатели не участвуют в процессе дальнейшей кластеризации): ФИО – фамилия (имя, отчество) ученого; *SR* – скорость его успеха, которая является вещественным

показателем и исчисляется в годах до первого значимого достижения; сектор науки, в котором ученый добился успеха. Для иллюстрации векторизации профессиональных траекторий рассмотрим примеры, представленные в табл. 1. Аналогичным образом были «оцифрованы» биографии других ученых. Все векторы были сведены в единый массив данных, состоящий из 100 объектов.

*Таблица 1*

ПРИМЕРЫ ВЕКТОРОВ ПРОФЕССИОНАЛЬНЫХ ТРАЕКТОРИЙ УЧЕНЫХ

| ФИО | $x_{j1}$ | $x_{j2}$ | $x_{j3}$ | *SR* | Сектор науки |
|---|---|---|---|---|---|
| Бор Н. | 4,0 | 3,0 | 4,0 | 31 | ЕН |
| Даль В.И. | 2,0 | 1,0 | 2,0 | 29 | ГУМ |

Выявление групп однотипных профессиональных траекторий проводилось путем кластеризации полученного набора данных, причем число кластеров (типов) и их характеристики заранее не были известны.

Задачу кластеризации можно сформулировать следующим образом. Дано множество профессиональных траекторий ученых $S = \{s_1, s_2, \ldots s_i, \ldots s_n\}$, $n = 100$, каждая из которых представлена вектором $x_j = \{x_{j1}, x_{j2}, x_{j3}\}$, $j = 1, 2, \ldots, n$. Требуется построить множество кластеров $C$ и отображение $F$ такое, что $F : S \rightarrow C$, где

$$C = \{c_1, \ldots c_k, \ldots c_m\}, \rightarrow m = 2,3,4,5,6,7,8;$$

$c_k$ – кластер, содержащий однотипные траектории из множества $S$:

$$c_k = \{s_i, s_j | s_i \in S, s_j \in S \text{ и } d(s_i, s_j) < \sigma\}.$$

*Здесь* $\sigma$ – критерий близости траекторий, $d(s_i, s_j)$ – мера близости между траекториями. В случае $d(s_i, s_j) < \sigma$ траектории помещаются в один кластер, а если $d(s_i, s_j) \geq \sigma$, то они попадают в разные кластеры (типы). При этом характер распределения объектов в трехмерном пространстве вещественных координат заранее не известен.

Для решения такой задачи в общем случае можно применять классические алгоритмы машинного обучения (*k*-средних, DBSCAN или иерархическую кластеризацию), а также нейросетевые подходы (самоорганизующиеся карты или сети Кохонена). Классические алгоритмы не являются универсальными, эффективность их применения зависит от характера распределения объектов (векторов) массива данных в многомерном пространстве состояний. Сравнительный анализ алгоритмов и их недостатки неоднократно освещались в обзорных публикациях как российских [28], так и зарубежных [29; 30] исследователей. Как следует из предыдущих исследований, метод *k*-средних весьма критичен к выбору координат центроидов и их количеству. Он хорошо работает только в случае группировки объектов в пространстве в плотные сгустки. DBSCAN зависим от выбора корневых объектов, игнорирует выбросы (шумы) и подходит для выявления кластеров на базе компактных группировок объектов специфической (кольцевой или ленточной) конфигурации. Алгоритм иерархической кластеризации, объединяющий на каждом шаге в один кластер два ближайших объекта (кластера), критичен к выбору функции близости, а самое главное, не дает ответа на основной вопрос об оптимальном количестве кластеров (типов) объектов в изучаемом массиве данных. Таким образом, для применения того или иного классического алгоритма необходимо предварительное изучение распределения объектов в пространстве, что легко сделать для случая двухмерного пространства. Для более сложных случаев требуется применение методов понижения размерности пространств, что снижает точность анализа и повышает вероятность ошибок при выборе метода кластеризации.

В свою очередь, при использовании метода нейросетевого кластерного анализа происходит несколько разбиений, которые анализируются с точки зрения критериев качества кластеризации. В настоящем исследовании применялись два критерия. Первый определяется как минимум сумм ошибок квантования (подробнее об этом критерии можно найти в предыдущих работах членов

творческого коллектива [25; 31]). Критерий был рассчитан в программной среде Deductor (версия Deductor Academic 5.3). В ней же проводились все вычислительные эксперименты с искусственными нейронными сетями (ИНС). Первый критерий, таким образом, позволил настроить начальные условия процесса моделирования сети и вычислить средние и максимальные ошибки квантования, которые нужны для выбора лучшей карты. Второй критерий предполагает расчет коэффициентов силуэта [28, с. 46]. Выбирая разбиение с максимальным коэффициентом силуэта, можно определить оптимальное количество кластеров. Наилучшее разбиение, таким образом, подтверждается сразу двумя критериями.

Метод нейросетевого кластерного анализа не требует предварительной подготовки данных и изучения распределения объектов в пространстве. Применение самоорганизующихся карт Кохонена [32; 33], которые представляют собой специальные искусственные нейронные сети, позволяет в автоматическом режиме группировать объекты исходного датасета и формировать кластеры, состоящие из однотипных векторов. Для получения достоверного результата процесс кластеризации повторяется многократно с разными начальными условиями. После завершения вычислительных экспериментов выбирается лучшая карта с использованием критерия минимума сумм ошибок квантования.

Таким образом, в рассматриваемом случае кластеризации профессиональных траекторий ученых, представленных векторами в трехмерном пространстве вещественных координат, целесообразно использовать именно нейросетевой кластерный анализ, поскольку предварительно не известно количество типовых групп и не ясен характер распределения объектов в пространстве. Кроме того, в этом случае не требуется никакой предварительной подготовки данных.

## Ход исследования

Предварительный анализ позволил сделать несколько общих замечаний относительно всей выборки исследования. Средним

возрастом успеха по всей выборке стали 33,5 года, при этом персоналии гуманитарного профиля имели чуть более длинный путь, чем естественно-научного направления (рис. 4).

Рассмотрим распределение количества биографий в соответствии с заданной ранее матрицей классификации жизненных условий, в которой можно отразить либо низкие, либо высокие значения факторов (рис. 2А, 2Б, 3). Например, на рис. 5 сегмент 1.0 соответствует низкому уровню образования и мотивации к действию близкого круга одаренного и низкому уровню финансовой обеспеченности семьи. К сегменту 1.1 отнесем тех одаренных, у которых низкий уровень финансовой обеспеченности родителей сочетался с их высокой заинтересованностью в развитии доминанты оригинальности ребенка, либо в обратном сочетании (сегмент 1.1). В сегмент 1.3 вошли биографии ученых, для которых были созданы наиболее благоприятные стартовые условия.

Анализ данных, представленных в виде предложенной матрицы условий (рис. 5), позволяет предположить, что большая часть одаренных на старте жизненного пути либо имела удачное сочетание финансовой обеспеченности и образованности своего окружения (сегмент 1.3, рис. 5), либо не имела никакой финансовой и мотивационной поддержки от близкого окружения (сегмент 1.0, рис. 5).

По мере усложнения социального института науки добиться значительных успехов без качественного образования стало сложнее (рис. 2Б – рис. 6).

Если в XIX в. одаренные личности достигали успеха и без образования, и без ранней профориентации (сегменты 2.0 и 2.1, рис. 6), то в XX в. преимущество за теми из ученых, кто сразу определился с профильным образованием и сумел уклониться от «метаний».

Что касается коллективного фактора (рис. 3 – рис. 7), то ожидаемо его роль имела решающее значение для научной сферы (сегмент 3.2, рис. 7). Лишь немногим удался полностью самостоятельный профессиональный путь.

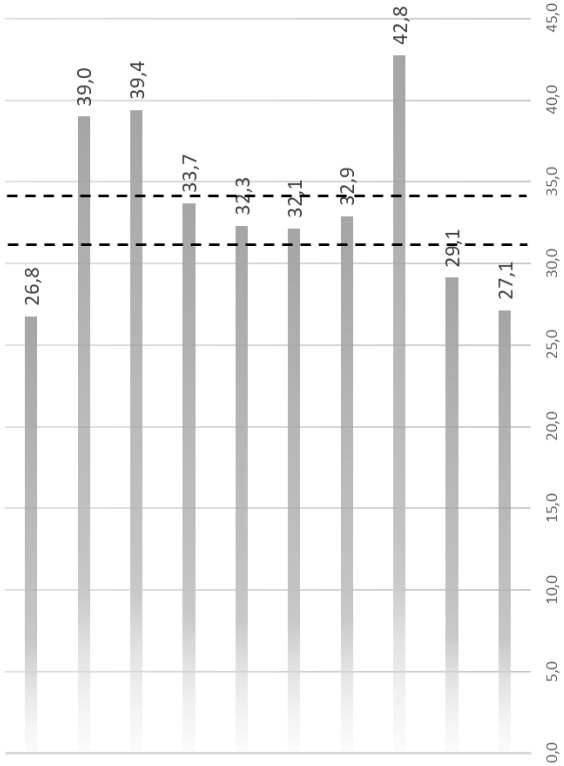



***Рис. 4. Средний «возраст успеха» представителей естественно-научного и гуманитарного профиля выборки ученых (100 чел., XIX, XX вв.), годы***

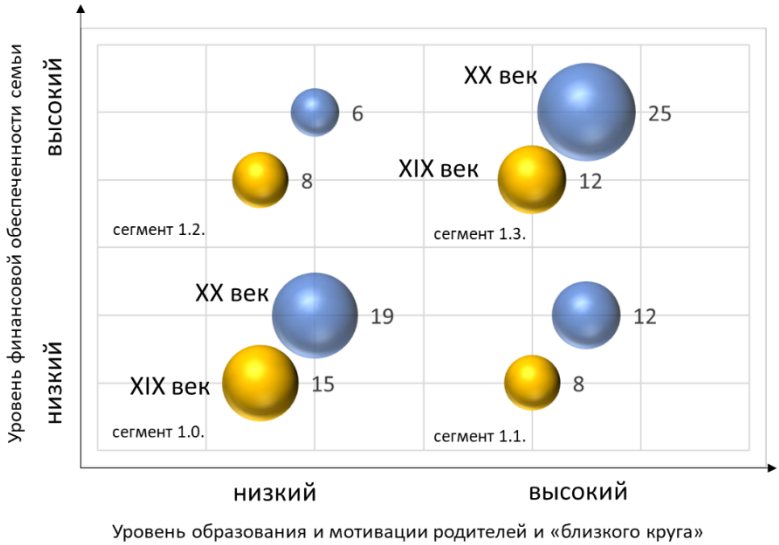



***Рис. 5. Матричный анализ по 4 группам стартовых условий***

*Примечание.* Здесь, на рис. 5, и далее на рис. 6–7 размер шара соответствует количеству биографий, которые *могут быть отнесены к тем или иным сегментам предложенной классификационной матрицы*, а цвет относится к веку: *желтый* – XIX *в.*, синий – XX *в.*

Для построения типологии профессиональных траекторий нейросетевой кластерный анализ проводился следующим образом. Весь массив векторов «оцифрованных» биографий ученых разбивался последовательно на 2, 3, 4, 5, 6, 7 и 8 кластеров. На каждом этапе было проведено по 12 вычислительных экспериментов с разными значениями гиперпараметров алгоритма самоорганизации нейронной сети. Заключительный этап продемонстрировал наличие кластера нулевой мощности, что сигнализировало о формировании семи стабильных групп и завершении всей процедуры. Далее, из всех синтезированных карт было выбрано лучшее разбиение на базе критерия минимума сумм ошибок квантования, вычисляемых в программе Deductor. Ошибки квантования характе-

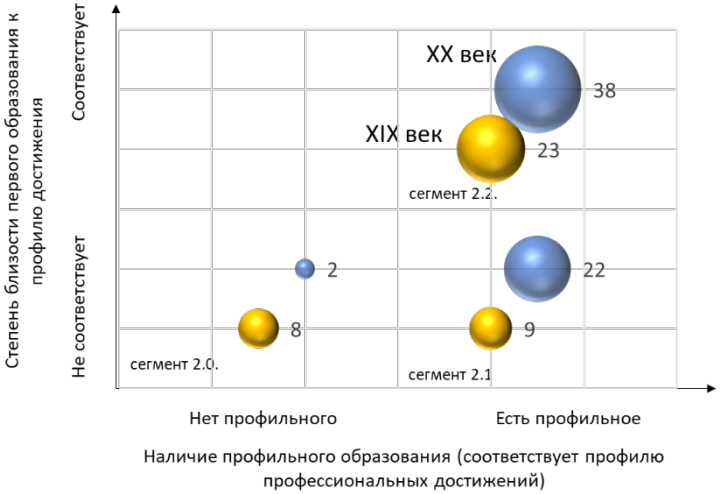



***Рис. 6. Матричный анализ по 3 группам этапа «Обучение и становление»***

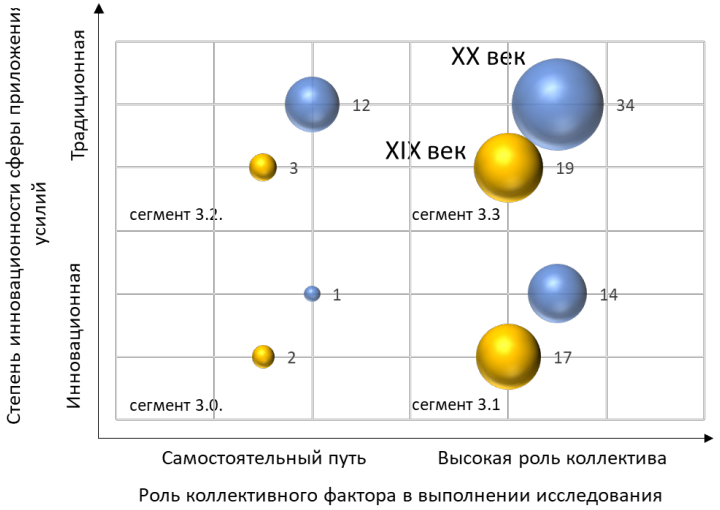



***Рис. 7. Матричный анализ по 4 группам этапа «Профессиональная реализация»***

ризуют отклонения конкретных векторов кластера от центра соответствующих ячеек карты Кохонена: чем меньше эти ошибки, тем качественнее кластеризация [24]. Сформированные кластеры были проанализированы с точки зрения их состава и характеристик.

## Обсуждение полученных результатов

Содержательный анализ составов полученных кластеров и их количественных характеристик, приведенных в табл. 2, позволил выявить характерные типы профессиональных траекторий.

*Таблица 2*

КОЛИЧЕСТВЕННЫЕ ХАРАКТЕРИСТИКИ КЛАСТЕРОВ

| №/цвет | МК | Характеристики кластера | $X_1$ | $X_2$ | $X_3$ | *SR* | ГУМ | ЕН |
|---|---|---|---|---|---|---|---|---|
| 1 Голубой | 9 | Среднее значение | 3,11 | 3 | 2,44 | 35,11 | 1 | 8 |
| | | Стандартное отклонение | 0,33 | 0 | 0,88 | | | |
| 2 Оранжевый | 13 | Среднее значение | 1,54 | 2,77 | 2,15 | 35,85 | 9 | 4 |
| | | Стандартное отклонение | 0,52 | 0,44 | 0,55 | | | |
| 3 Серый | 20 | Среднее значение | 3,75 | 1,7 | 3,65 | 32,2 | 15 | 5 |
| | | Стандартное отклонение | 0,44 | 0,47 | 0,49 | | | |
| 4 Желтый | 15 | Среднее значение | 3,67 | 3 | 4 | 31,53 | 6 | 9 |
| | | Стандартное отклонение | 0,49 | 0 | 0 | | | |
| 5 Синий | 20 | Среднее значение | 1,15 | 3 | 3,7 | 34,55 | 10 | 10 |
| | | Стандартное отклонение | 0,37 | 0 | 0,47 | | | |

*Окончание табл. 2*

| №/цвет | МК | Характеристики кластера | $X_1$ | $X_2$ | $X_3$ | *SR* | ГУМ | ЕН |
|---|---|---|---|---|---|---|---|---|
| 6 Зеленый | 15 | Среднее значение | 1,27 | 1,8 | 3,67 | 32,67 | 10 | 5 |
| | | Стандартное отклонение | 0,46 | 0,41 | 0,62 | | | |
| 7 Темно-синий | 8 | Среднее значение | 4 | 2,88 | 2,5 | 28,75 | 4 | 4 |
| | | Стандартное отклонение | 0 | 0,35 | 0,53 | | | |

*Примечание*. В этой таблице приняты следующие обозначения: №/цвет – номер кластера и цвет его линии на рис. 8; МК – мощность кластера (сумма числа ученых, чьи биографии отнесены к кластеру); *SR* – скорость успеха, лет; $X_1$; $X_2$; $X_3$ – вещественные интервальные переменные, отражающие уровни трех основных факторов стадиальной модели жизненного пути ученого (рис. 1); ГУМ – число ученых-гуманитариев, попавших в кластер; ЕН – число ученых-естествоиспытателей, попавших в кластер.

На основе табл. 2 построены лепестковые диаграммы. Они дают возможность визуализировать полученные кластеры в координатах, соответствующих факторам модели (семья, образование, профессиональное становление (реализация)). Цвета линий, представленных на рис. 8, соотнесены с кластерами в табл. 2 в графе №/цвет.

**Первый** кластер («Инноваторы традиций») характеризуется хорошими стартовыми условиями, точным самоопределением. Большинство представителей этой группы достигли успеха в естественно-научной сфере. Как правило, на жизненном пути одаренные из этого кластера отличались большой самостоятельностью (например, А. Эйнштейн или Э. Шрёдингер).

Во **втором** кластере («Одиссеи науки») − преимущество за гуманитариями. Представители этого кластера имели непростые стартовые условия, но отличались быстрым самоопределением

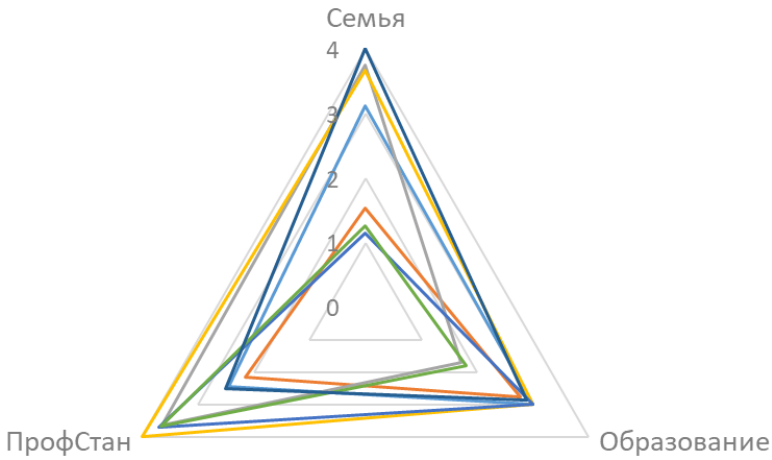



***Рис. 8. Лепестковые диаграммы кластеров***

на этапе обучения. Однако сфера приложения их усилий не отличается высокой степенью инновационности и высокой научной коллаборацией. Здесь мы найдем много «одиночек», самостоятельно сформировавшихся как профессионалы в науке. Средняя скорость успеха – почти 36 лет, что существенно выше среднего значения по всей выборке. Например, в этот кластер попала биография Ф. Ницше.

Особенностью представителей **третьего** кластера («Пионеры») является длинный этап профессионального самоопределения, связанный с поиском сферы приложения своих талантов. Здесь перевес представителей гуманитарных наук. Однако в отличие от второго кластера, представители данного кластера брались за исследования нового направления. Поэтому, как правило, они сами становились основателями научных школ. Учитывая, что у вошедших в этот кластер ученых была хорошая стартовая семейная поддержка, им удалось добиться успеха, хоть путь самоопределения у многих был непростой. Например, сюда была отнесена биография Ч. Дарвина. Он происходил из обеспеченной

и уважаемой семьи, что дало ему определенную творческую и академическую свободу. Другим примером может быть биография М. Вебера: он происходил из семьи успешного юриста и политика.

Попавших в **четвертый** кластер можно характеризовать как «круглых отличников», «Моцартов науки». У них хороший старт и точное самоопределение, сфера применения их усилий отличается высокой инновационностью. Не случайно средний возраст успеха в этом кластере 31,5 года. Большинство представителей этого кластера работали в сфере точных и естественных наук. Например, в этот кластер попадают биографии Н. Бора и М. Борна.

В **пятый** кластер («Стойкие экспериментаторы») попали ученые, чьи биографии отличались самыми сложными стартовыми условиями, которые им приходилось восполнять «удачей» и трудолюбием на втором и третьем этапах. Здесь много экспериментаторов, например – Д. Менделеев, А. Макаренко. Профессиональный путь таких людей требует времени и большого трудолюбия, но достижения имеют инновационное значение для отрасли.

В **шестом** кластере сосредоточены «Загадочные таланты»: их путь представляется самым неопределенным с точки зрения условий для успеха. Для них характерны сложные стартовые условия и низкая скорость самоопределения в период обучения и становления. Вторая учеба, или второстепенная работа, «мешала» сосредоточиться на доминанте оригинальности. При этом их включение в научную коммуникацию ничем не уступает «экспериментаторскому» пути. Изучение биографий этого кластера дает основание для гипотезы о длинном поиске «себя» в связи с разносторонними талантами (задатками). Например, к этому кластеру относится Л. Пастер, который успел поработать репетитором, администратором в сфере образования, и даже размышлял о карьере художника. Однако страсть к химии в итоге победила другие интересы и занятия. В биографии К.Д. Ушинского встречается история о случайно найденном архиве инспектора Гатчинского сиротского института, Е.О. Гугеля, в котором, как писал сам

Ушинский, он нашел «полное собрание педагогических книг».

Представители **седьмого** кластера («Прорывные мастера») отличаются самыми лучшими стартовыми условиями и самым коротким временем до первого успеха – всего 28,75 года. Это, скорее всего, связано с рано проявленной гениальностью (Галенсон), высокой оригинальностью работ еще на ранних стадиях самоопределения. Здесь поровну представителей обоих научных направлений – и «физиков», и «лириков». Сюда были отнесены биографии М. Планка, Л. Ландау, а из гуманитариев, например, Ж.-П. Сартра.

Сделаем несколько обобщающих замечаний.

1. Высокая скорость достижения успеха в науке в первую очередь зависит от стартовых условий (семейного благосостояния вкупе с образованностью семьи). 3-й, 4-й и особенно 7-й кластеры показывают сильную связь между факторами первой группы и «средним возрастом» успеха.

2. Правильный выбор научной школы (мирового научного центра) может компенсировать «низкий старт». В этом результаты исследования согласуются с выводом о важной роли мобильности в профессиональной карьере [17;18].

3. Представители естественно-научной сферы, как правило, быстрее достигали успеха, чем гуманитарии. Их отличают более высокие показатели самоопределения на этапе «Образование и становление». Это может быть обусловлено недостаточной развитостью мировых центров гуманитарной науки, сильной фрагментацией гуманитарного научного знания, слабой преемственностью школ и наставничества.

Одним из основных ограничений исследования является его ориентированность на конкретную выборку выдающихся ученых XIX и XX вв. Выбранное временное окно может не учитывать факторы и изменения, которые могли бы повлиять на профессиональные траектории в современных условиях. Ограничением является также допущение о роли формального образования в успехе

одаренной личности: в течение XIX–XX вв. институт образования претерпевал существенные изменения. Другим ограничением является качество данных. Результаты контент-анализа зависят от доступности и точности биографических материалов, а также выбранных источников информации. Метод нейросетевого анализа с самоорганизующимися картами Кохонена представляет собой новый для этой сферы исследовательский подход, однако следует помнить, что результаты могут быть зависимы от параметров и начальных условий.

## Выводы

В исследовании с помощью нейросетевого анализа на выборке в 100 биографий выдающихся личностей из науки XIX и XX вв. были выявлены и интерпретированы семь траекторий. Было обосновано, что для поставленной задачи целесообразно использовать именно нейросетевой кластерный анализ, поскольку предварительно не известно количество типовых групп и не ясен характер распределения объектов в пространстве. Кроме того, в этом случае не требуется никакой предварительной подготовки данных.

Анализ траекторий проводился с точки зрения скорости достижения успеха (среднего возраста успеха) и тех факторов и условий жизненного пути, которые могли повлиять на более быстрое или медленное достижение профессиональных целей и самореализацию одаренной личности. На выборке исследования были сформулированы гипотезы, открывающие возможности для дальнейших исследований с использованием разработанной методологии контент-анализа, а также развития опыта применения нейросетевого анализа. Ограничения метода заключаются в повышенных требованиях, которые предъявляются к вычислительным мощностям при проведении вычислительных экспериментов, и в возможных проблемах с шумом и выбросами. Однако в контексте поставленной задачи поиска конечного числа типовых про-

фессиональных траекторий одаренных личностей нейросетевой анализ показал свои лучшие стороны, так как позволил работать со сложными формами кластеров и сосредоточиться на поиске их оптимального числа.

## ЛИТЕРАТУРА

1. *Mayer K., Pfeffer J.* Computational Social Science // Schlüsselwerke der Netzwerkforschung / Ed. by B. Holzer, C. Stegbauer. Wiesbaden: Springer VS, 2019. P. 721–723. DOI: 10.1007/978-3-658-21742-6_77.

2. *Beytía P., Schobin J.* Networked pantheon: a relational database of globally famous people: Social and behavioural sciences // Research Data Journal for the Humanities and Social Sciences. 2020. Vol. 5, № 1. P. 50–65. DOI: 10.17632/twvsjygw3m.1.

3. *Chisholm A., Radford W., Hachey B.* Learning to generate one-sentence biographies from Wikidata // Cornwall University [site]. 21.02.2017. URL: arXiv preprint arXiv:1702.06235. (дата обращения: 01.02.2023).

4. *Reznik I., Shatalov V.* Hidden revolution of human priorities: An analysis of biographical data from Wikipedia // Journal of Informetrics. 2016. Vol. 10, № 1. P. 124–131. DOI: 10.1016/j.joi.2015.12.002.

5. *Samoilenko A., Yasseri T.* The distorted mirror of Wikipedia: a quantitative analysis of Wikipedia coverage of academics // EPJ Data Science. 2014. Vol. 3, № 1. P. 1–11. DOI: 10.1140/epjds20.

6. Beyond One-Dimensional Portraits: A Synoptic Approach to the Visual Analysis of Biography Data / F. Windhager, M. Schlögl, M. Kaiser [et al.] // Conference: Biographical Data in a Digital World 2017 (BD 2017). Volume: CEUR Vol-2119. Linz, Austria, 2017. P. 67–75.

7. *Collison P., Nielsen M.* Science is getting less bang for its buck // The Atlantic [site]. 16.11.2018. URL: https://www.theatlantic.com/science/archive/2018/11/diminishing-returns-science/575665/ (дата обращения: 19.01.2023).

8. *Sternberg R.J.* Identification for utilization, not merely possession, of gifts: What matters is not gifts but rather deployment of gifts // Gifted Education International. 2022. Vol. 38, № 3. P. 354–361. DOI: 10.1177/02614294211013345.

9. *Леонтьев Д.А., Лебедева А.А., Костенко В.Ю.* Траектории личностного развития: реконструкция взглядов Л.С. Выготского // Вопросы образования. 2017. № 2. С. 98–112. DOI: 10.17323/1814-9545-2017-2-98-112. EDN: YUPYIX.

10. *Шадриков В.Д.* Отношение понятий «жизнь», «поведение», «деятельность» // Мир психологии. 2020. № 2(102). С. 57–65. EDN: BCDJMW.

11. *Galenson D.W.* Old Masters and Young Geniuses: The Two Life Cycles of Human Creativity // Journal of Applied Economics. 2009. Vol. 12, № 1. P. 1–9. DOI:

10.1016/S1514-0326(09)60002-7.

12. *Weinberg B.A., Galenson D.W.* Correction to: Creative Careers: The Life Cycles of Nobel Laureates in Economics // De Economist. 2019. Vol. 167, № 3. P. 241–241. DOI: 10.1007/s10645-019-09342-0.

13. *Кольцова О.Ю., Маслинский К.А.* Выявление тематической структуры российской блогосферы: автоматические методы анализа текстов // Социология: методология, методы, математическое моделирование (Социология: 4М). 2013. № 36. С. 113–139. EDN: RCFOWJ.

14. *Ким А.В., Мальцева Д.В., Щеглова Т.Е.* Блокмоделинг для анализа социальных структур: пример изучения структуры сообщества петербургских социологов // Социология: методология, методы, математическое моделирование (Социология: 4М). 2021. № 53. С. 7–38. DOI: 10.19181/4m.2021.53.1. EDN: HYNUSK.

15. The effects of outliers' data on neural network performance / A. Khamis, Z. Ismail, K. Haron, A. Tarmizi // Journal of Applied Sciences. 2005. Vol. 5, № 8. P. 1394–1398. DOI: 10.3923/jas.2005.1394.1398.

16. *Mijwel M.M.* Artificial neural networks advantages and disadvantages // Mesopotamian Journal of Big Data. 2021. Vol. 2021. P. 29–31. DOI: 10.58496/mjbd/2021/006.

17. *Schlagberger E.M., Bornmann L., Bauer J.* At what institutions did Nobel laureates do their prize-winning work? An analysis of biographical information on Nobel laureates from 1994 to 2014 // Scientometrics. 2016. Vol. 109, № 2. P. 723–767. DOI: 10.1007/s11192-016-2059-2.

18. *Lucchini L., Tonelli S., Lepri B.* Following the footsteps of giants: modeling the mobility of historically notable individuals using Wikipedia // EPJ Data Science. 2019. № 8. P. 36. DOI: 10.1140/epjds/s13688-019-0215-7.

19. *Cui H., Wu L., Evans J.A.* Aging Scientists and Slowed Advance // Cornwall University [site]. 08.02.2022. URL: arXiv preprint arXiv:2202.04044. 2022 (дата обращения: 20.02.2023).

20. *Jones B.F., Reedy E.J., Weinberg B.A.* Age and scientific genius // The Wiley handbook of genius / Ed. by D. K. Simonton. Hoboken: John Wiley & Sons, Ltd, 2014. P. 422–450. DOI: 10.1002/9781118367377.ch20.

21. Developmental biographies of Olympic super-elite and elite athletes: A multidisciplinary pattern recognition analysis / A. Güllich, L. Hardy, L. Kuncheva [et al.] // Journal of Expertise. 2019. Vol. 2 (1). P. 23–46.

22. *Letiagina E., Perova V., Orlova E.* Neural network analysis of the development of physical education and sports in Russia as an economic factor of country security // Proceedings of the 4th International Conference on Innovations in Sports, Tourism and Instructional Science (ICISTIS 2019). Atlantis Press. 2019. № 11. P. 138–142. DOI: 10.2991/icistis-19.2019.37.

23. *Subotnik R.F., Olszewski-Kubilius P., Worrell F.C.* Environmental factors and personal characteristics interact to yield high performance in domains // Frontiers in

Psychology. 2019. № 10:2804. P. 1–8. DOI: 10.3389/fpsyg.2019.02804.

24. *Трифонов Ю.В., Сочков А.Л., Соловьев А.Е.* Оценка экономического потенциала регионов РФ на основе методологии нейросетевого кластерного анализа // Вестник Нижегородского университета им. Н.И. Лобачевского. Серия: Социальные науки. 2021. № 3(63). С. 38–47. DOI: 10.52452/18115942_2021_3_38. EDN: SIQALB.

25. *Трифонов Ю.В., Сочков А.Л., Куликова А.В.* Построение и реализация моделей интеллектуальных цифровых коммуникаций в социально-политических сферах // Экономика и предпринимательство. 2021. № 8(133). С. 1087–1095. DOI: 10.34925/EIP.2021.133.8.209. EDN: UGMSUS.

26. *Carboni O.A., Russu P.* Assessing Regional Wellbeing in Italy: An Application of Malmquist–DEA and Self-organizing Map Neural Clustering // Social Indicators Research. 2015. Vol. 122, № 3. P. 677–700. DOI: 10.1007/s11205-014-0722-7.

27. Regional disaster risk assessment of China based on self-organizing map: Clustering, visualization and ranking / N. Chen, L. Chen, Y. Ma, A. Chen // International Journal of Disaster Risk Reduction. 2019. № 33. P. 196–206. DOI: 10.1016/j.ijdrr.2019.101085.

28. *Абдурахманова Э.М.* Исследование структур обобщенных групп, выделяемых разными методами, на примере результатов исследования СТАРТ // Социология: методология, методы, математическое моделирование (Социология: 4М). 2020. № 50–51. С. 37–63. EDN: MEPMGD.

29. *Gülagiz F.K., Sahin S.* Comparison of hierarchical and non-hierarchical clustering algorithms // International Journal of Computer Engineering and Information Technology. 2017. Vol. 9, № 1. P. 6–14.

30. *Musdholifah A., Hashim S.Z.M., Zaiton S.* Cluster analysis on high-dimensional data: A comparison of density-based clustering algorithms // Australian Journal of Basic and Applied Sciences. 2013. Vol. 7, № 2. P. 380–389.

31. Research of the innovative development of the Russian Federation regions and its impact on the eco-friendliness of the economy based on neural network cluster analysis for the purpose of economic security / S. Yashin, Y. Trifonov, A. Sochkov [et al.] // E3S Web of Conferences. 2021. Vol. 291. P. 1–10. DOI: 10.1051/e3sconf/202129103008.

32. *Kohonen T.* The self-organizing map // Proceedings of the IEEE. Vol. 78, № 9. P. 1464–1480. DOI: 10.1109/5.58325.

33. Engineering applications of the self-organizing map / T. Kohonen, E. Oja, O. Simula [et al.] // Proceedings of the IEEE. Vol. 84, № 10. P. 1358–1384. DOI:10.1109/5.537105.

***Chepyuk Olga R.,***
*Doctor of Philosophy, Professor of the Department of Human Resource Management, National Research Lobachevsky State University, Nizhny Novgorod, Russia,* chepyuko@yandex.ru

***Angelova Olga Yu.,***
*Candidate of Economic Sciences, Associate Professor of the Department of Information Technologies and Instrumental Methods in Economics, National Research Lobachevsky State University, Nizhny Novgorod, Russia*, oangelova@mail.ru

***Sochkov Andrey L.,***
*Candidate of Technical Sciences, Associate Professor of the Department of Information Technologies and Instrumental Methods in Economics, National Research Lobachevsky State University, Nizhny Novgorod, Russia*, sochkov@iee.unn.ru

***Podolskaya Tatyana O.,***
*Candidate of Sociological Sciences, Associate Professor of the Department of Human Resource Management, National Research Lobachevsky State University, Nizhny Novgorod, Russia*, podolskaya@iee.unn.ru

**Typology of professional trajectories of gifted individuals using neural network analysis**

Based on a data set (100 biographies) created by the authors through content analysis of biographical material about outstanding scientists of the 19th and 20th centuries in the humanities and natural sciences, the clustering of professional trajectories of gifted individuals was carried out. Neural network analysis based on self-organizing Kohonen maps was used as a clustering method. The professional trajectories were formed within the framework of the behavioral model of the linear-stage approach to studying life cycles. Within this approach, career and professional self-realization are understood as a sequence of evolutionary stages fixed in their order of occurrence. Each stage was encoded, and the biographies were transformed into a vector system. In turn, the task of clustering consisted in dividing a hundred vectors into typical groups with several real-valued coordinates. The criteria for the quality of clustering were the minimum sum of quantization errors and the silhouette coefficient. As a result of the study, seven professional trajectories

of gifted individuals were identified and interpreted. The analysis of trajectories was carried out from the point of view of the speed of success (average age of success) and those factors and conditions of the life path that could affect either rapid or slow achievement of professional goals and self-realization. This example demonstrates the possibilities and limitations of using neural network analysis for solving similar research tasks, especially when working with complex cluster forms and finding their optimal number.
*Keywords*: neural network analysis, giftedness, gifted personality, professional trajectory, machine learning, neural network, Kohonen maps

**References**

1. Mayer K., Pfeffer J. *Computational Social Science*. Wiesbaden: Springer VS, 2019, P. 721–723. DOI: 10.1007/978-3-658-21742-6_77.
2. Beytía P., Schobin J. Networked pantheon: a relational database of globally famous people: Social and behavioural sciences, *Research Data Journal for the Humanities and Social Sciences,* 2020, vol. 5, no. 1, p. 50–65. DOI: 10.17632/twvsjygw3m.1.
3. Chisholm A., Radford W., Hachey B. Learning to generate one-sentence biographies from Wikidata, *Cornwall University* [site]. 21.02.2017, URL: arXiv preprint arXiv:1702.06235. (date of the application: 01.02.2023).
4. Reznik I., Shatalov V. Hidden revolution of human priorities: An analysis of biographical data from Wikipedia, *Journal of informetrics,* 2016, vol. 10, no. 1, p. 124–131. DOI: 10.1016/j.joi.2015.12.002.
5. Samoilenko A., Yasseri T. The distorted mirror of Wikipedia: a quantitative analysis of Wikipedia coverage of academics, *EPJ data science*, 2014, vol. 3, no. 1, p. 1–11. DOI: 10.1140/epjds20.
6. Windhager F., Schlögl M., Kaiser M. et al. Beyond One-Dimensional Portraits: A Synoptic Approach to the Visual Analysis of Biography Data, *Conference: Biographical Data in a Digital World 2017* (BD 2017), Vol.: CEUR Vol-2119, Linz, Austria, 2017. P. 67–75.
7. Collison P., Nielsen M. Science is getting less bang for its buck, *The Atlantic* [site]. 16.11.2018. URL: https://www.theatlantic.com/science/archive/2018/11/diminishing-returns-science/575665/ (date of access: 19.01.2023).
8. Sternberg R.J. Identification for utilization, not merely possession, of gifts: What matters is not gifts but rather deployment of gifts,

*Gifted Education International*, 2022, vol. 38, no. 3, p. 354–361. DOI: 10.1177/02614294211013345.
9. Leontyev D.A., Lebedeva A.A., Kostenko V.Yu. Trajectories of personal development: reconstruction of the views of L.S. Vygotsky (in Russian), *Issues of education*, 2017, no. 2, p. 98–112. DOI: 10.17323/1814-9545-2017-2-98-112.
10. Shadrikov V.D. The relationship of the concepts "life", "behavior", "activity" (in Russian), *World of Psychology*, 2020, vol. 2, no. 102, p. 57–65.
11. Galenson D.W. Old Masters and Young Geniuses: The Two Life Cycles of Human Creativity, *Journal of Applied Economics,* 2009, vol. 12, no. 1, p. 1–9. DOI: 10.1016/S1514-0326(09)60002-7.
12. Weinberg B.A., Galenson D.W. Correction to: Creative Careers: The Life Cycles of Nobel Laureates in Economics, *De Economist*, 2019, vol. 167, no. 3, p. 241–241. DOI: 10.1007/s10645-019-09342-0.
13. Koltsova O.Yu., Maslinsky K.A. Revealing the thematic structure of the Russian blogosphere: automatic methods of text analysis (in Russian), *Sotsiologiya 4M (Sociology: methodology, methods, mathematical modeling),* 2013, no. 36, p. 113–139.
14. Kim A.V., Maltseva D.V., Shcheglova T.E. Block modeling for the analysis of social structures: an example of studying the structure of a community of St. Petersburg sociologists (in Russian), *Sotsiologiya 4M (Sociology: methodology, methods, mathematical modeling),* 2021, no. 53, p. 7–38. DOI: 10.19181/4m.2021.53.1.
15. Khamis A., Ismail Z., Haron K., Tarmizi A. The effects of outliers' data on neural network performance, *Journal of Applied Sciences*, 2005, vol. 5, no. 8, p. 1394–1398. DOI: 10.3923/jas.2005.1394.1398.
16. Mijwel M.M. Artificial neural networks advantages and disadvantages, *Mesopotamian Journal of Big Data,* 2021, vol. 2021, p. 29–31. DOI: 10.58496/mjbd/2021/006.
17. Schlagberger E.M., Bornmann L., Bauer J. At what institutions did Nobel laureates do their prize-winning work? An analysis of biographical information on Nobel laureates from 1994 to 2014, *Scientometrics*, 2016, vol. 109, no. 2, p. 723–767. DOI: 10.1007/s11192-016-2059-2.
18. Lucchini L., Tonelli S., Lepri B. Following the footsteps of giants: modeling the mobility of historically notable individuals using Wikipedia, *EPJ Data Science,* 2019, no. 8, p. 36. DOI: 10.1140/epjds/s13688-019-0215-7.

19. Cui H., Wu L., Evans J.A. Aging Scientists and Slowed Advance, *Cornwall University [site]*. 08.02.2022. URL: arXiv preprint arXiv:2202.04044, 2022 (date of access: 20.02.2023).
20. Jones B.F., Reedy E.J., Weinberg B.A. Age and scientific genius, *The Wiley handbook of genius*, Hoboken: John Wiley & Sons, Ltd, 2014, p. 422–450. DOI: 10.1002/9781118367377.ch20.
21. Gullich A., Hardy L., Kuncheva L. et al. Developmental biographies of Olympic super-elite and elite athletes: A multidisciplinary pattern recognition analysis, *Journal of Expertise*, 2019, vol. 2, no. 1, p. 23–46.
22. Letiagina E., Perova V., Orlova E. Neural network analysis of the development of physical education and sports in Russia as an economic factor of country security, *Proceedings of the 4th International Conference on Innovations in Sports, Tourism and Instructional Science (ICISTIS 2019)*. Atlantis Press, 2019, № 11, P. 138–142. DOI: 10.2991/icistis-19.2019.37.
23. Subotnik R.F., Olszewski-Kubilius P., Worrell F.C. Environmental factors and personal characteristics interact to yield high performance in domains, *Frontiers in Psychology*, 2019, no. 10:2804, P. 1–8. DOI: 10.3389/fpsyg.2019.02804.
24. Trifonov Yu.V., Sochkov A.L., Soloviev A.E. Assessment of the economic potential of the regions of the Russian Federation based on the methodology of neural network cluster analysis (in Russian), *Bulletin of the Lobachevsky University. Series: Social Sciences,* 2021, 3 (63), p. 38–47. DOI: 10.52452/18115942_2021_3_38.
25. Trifonov Yu.V., Sochkov A.L., Kulikova A.V. Construction and implementation of models of intelligent digital communications in socio-political spheres (in Russian), *Economics and Entrepreneurship*, 2021, 8 (133), p. 1087–1095. DOI: 10.34925/EIP.2021.133.8.209.
26. Carboni O.A., Russu P. Assessing Regional Wellbeing in Italy: An Application of Malmquist–DEA and Self-organizing Map Neural Clustering, *Social Indicators Research,* 2015, vol. 122, no. 3, p. 677–700. DOI: 10.1007/s11205-014-0722-7.
27. Chen N., Chen L., Ma Y. , Chen A. Regional disaster risk assessment of China based on self-organizing map: Clustering, visualization and ranking, *International Journal of Disaster Risk Reduction*, 2019, no. 33, p. 196–206. DOI: 10.1016/j.ijdrr.2019.101085.

28. Abdurakhmanova E.M. Study of the structures of generalized groups identified by different methods, using the results of the START study as an example (in Russian), *Sotsiologiya 4M (Sociology: methodology, methods, mathematical modeling)*, 2020, no. 50–51, p. 37–63.
29. Gülagiz F.K., Sahin S. Comparison of hierarchical and non-hierarchical clustering algorithms, *International Journal of Computer Engineering and Information Technology*, 2017, vol. 9, no. 1, P. 6–14.
30. Musdholifah A., Hashim S.Z.M., Zaiton S. Cluster analysis on high-dimensional data: A comparison of density-based clustering algorithms, *Australian Journal of Basic and Applied Sciences*, 2013, vol. 7, no. 2, p. 380–389.
31. Yashin S., Trifonov Y. , Sochkov A. et al. Research of the innovative development of the Russian Federation regions and its impact on the eco-friendliness of the economy based on neural network cluster analysis for the purpose of economic security, *E3S Web of Conferences*. 2021, vol. 291, p. 1–10. DOI: 10.1051/e3sconf/202129103008.
32. Kohonen T. The self-organizing map, *Proceedings of the IEEE,* vol. 78, no. 9, p. 1464–1480. DOI: 10.1109/5.58325.
33. .Kohonen T., Oja E. , Simula O. et al. Engineering applications of the self-organizing map, *Proceedings of the IEEE*, vol. 84, no. 10, p. 1358–1384. DOI:10.1109/5.537105.

# *К СВЕДЕНИЮ АВТОРОВ*

## *О журнале*

«Социология: методология, методы и математическое моделирование» (Социология: 4М) – специализированное издание, посвященное проблемам методологии и методов социологических исследований, вопросам сбора, измерения и анализа социологических данных, построению математических моделей социальных процессов.

Редакция журнала отдает предпочтение статьям, которые вносят вклад в развитие социологической методологии, проясняя существующие в этой области проблемы и предлагая новые решения. Одновременно в журнале публикуются аналитические обзоры по социологическим методам, статьи, в которых делается акцент на опыте применения методов сбора и анализа данных для решения актуальных социологических задач.

Основные рубрики журнала:

- общие вопросы методологии и методики исследований;
- методологические проблемы социологической теории;
- статистические методы и анализ данных;
- теория и методы измерения, теория и история методов;
- процедуры сбора данных;
- качество социологических данных, онлайн-опросы;
- качественные методы в социологии;
- методический эксперимент.

Критерии соответствия рукописей тематике журнала не являются жесткими, вопрос о целесообразности публикации статьи решается в каждом случае индивидуально.

## *Периодичность выхода и доступ к номерам*

Журнал выходит два раза в год. Полнотекстовые версии статей размещаются в свободном доступе на официальном сайте журнала после выхода номера.

## Порядок рассмотрения и рецензирования

После получения рукописи статьи редакция принимает решение о соответствии её содержания профилю журнала и о целесообразности передачи рукописи рецензентам. Причины отрицательного решения могут включать в себя в том числе отсутствие результатов проверки математических моделей на оригинальных эмпирических данных, недостаточное или вызывающее сомнение в их достоверности описание источников эмпирических данных, несоответствие современному состоянию исследований по проблеме, а также отсутствие научной новизны. В случае положительного решения статья передается на рецензирование. Решающим для принятия или отклонения рукописи становятся отзывы независимых рецензентов, назначаемых редакцией. Все статьи, направляемые в адрес редакции, проходят обязательную процедуру рецензирования одним экспертом. В случае необходимости редакция назначает второго рецензента.

Процедура рецензирования анонимна и для авторов, и для рецензентов. Рецензент получает рукопись статьи без указания имени и аффилиации авторов. Редакция не сообщает авторам статей фамилии рецензентов и не обсуждает их квалификацию. Рецензенты отбираются из числа специалистов в данной тематической области. Редакция сообщает о результатах рецензирования автору статьи посредством электронной почты в течение трех месяцев после ее получения; в случае отсутствия отзывов к этому моменту редакция сообщает о новых сроках рассмотрения.

Редакция журнала предоставляет авторам право ответить на замечания рецензента по существу и прояснить собственную позицию.

Журнал публикует оригинальные исследовательские работы, которые не публиковались прежде (за исключением электронных препринтов и тезисов). Передавая в редакцию рукопись, автор принимает на себя обязательство не публиковать ее ни полностью, ни частично в каком бы то ни было ином издании без согласования с редакцией журнала.

Плата за публикацию рукописей не взимается.

## *Оформление статьи*

Редакция журнала принимает статьи объемом до 40 тыс. знаков, включая пробелы (1 авт. лист). Материалы должны быть переданы в редакцию в электронном носителе (предпочтительно – посредством электронной почты).

Текст, включая примечания и библиографический список, должен соответствовать стандартам.

Шрифт – Times New Roman

Размер шрифта – 12

Межстрочный интервал – 1,5

Выравнивание – по ширине

Поля страницы: 2 см со всех сторон

Рисунки, схемы и таблицы должны быть такого же формата, что и текст, и снабжаться сквозной нумерацией.

Формулы и обозначения должны быть набраны в редакторе формул *Microsoft Equation*.

Комплект статьи включает, кроме основного текста, аннотацию, 8−10 ключевых слов с пометкой «Ключевые слова», справку об авторе (авторах) с указанием фамилии, имени и отчества, места работы, должности, ученой степени и звания, полного почтового домашнего адреса, номеров телефонов и адреса электронной почты.

Сопроводительное письмо к рукописи должно содержать описание научной новизны и краткое обоснование, почему статья может представлять интерес для читателей «Социологии: 4М».

Также в этом письме автор должен подтвердить, что представленная статья носит характер оригинального исследования, которое прежде не публиковалось нигде (кроме препринтов и тезисов конференций) и не находится на рассмотрении ни в каком другом издании.

## Требования академической этики

Редакционная политика журнала предполагает соблюдение всеми сторонами, участвующим в процессе подготовки статей (авторами, рецензентами и редакцией), требований публикационной этики, обеспечивающих беспристрастную и конфиденциальную оценку рукописей, отсутствие плагиата или незаконного присвоения результатов. Редакция выражает готовность публиковать сообщения о найденных ошибках и о фактах нарушения авторами рукописей публикационной этики.

## Сведения о статье на английском языке

Статья может быть принята к публикации только при наличии следующей информации на английском языке: автор, заглавие, данные об аффилиациях автора (наименования организаций, электронный адрес автора, ответственного за корреспонденцию), аннотация, ключевые слова. В качестве английских наименований организаций рекомендуется использовать названия, индексируемые в зарубежных базах научного цитирования (например, *Web of Science* или *Scopus*).

Аннотация на английском языке может быть расширенной, т.е. более полной по сравнению с аннотацией на русском языке. Аннотация должна укладываться в объем от 100 до 250 слов.

## Список использованных источников

Все источники, упомянутые в тексте, должны сопровождаться библиографическими ссылками. Автор обязан указать источники всех приводимых в статье цитат, цифр и иной информации. За точность (правильность) цитат в статье, а также цифр и иной информации, ответственность несет автор.

Ссылки на источники оформляются в виде пристатейного библиографического списка и нумеруются в порядке следования с указанием по тексту в квадратных скобках порядкового номера ссылки цифрой: [1], [7].

Одновременная ссылка на несколько номеров дается в одних скобках: [3; 7; 11; 12; 13], [3, с. 5; 7, с. 8−14]. Ссылка на неопубликованные работы не допускается. Библиографические описания изданий оформляются в соответствии с государственным стандартом и примерами, приведенными ниже.

При оформлении библиографических описаний обязательно указание DOI и EDN (при их наличии). Для монографий требуется указание ISBN.

Примеры библиографических описаний:

1. *Дюркгейм Э.* Моральное воспитание / Пер. с фр. А.Б. Гофмана. М.: НИУ ВШЭ, 2021. 456 с. ISBN 978-5-7598-2530-2. DOI: 10.17323/978-5-7598-2530-2. EDN: QOJUNL.

2. *Климова А. М., Артамонов Г. А., Чмель К. Ш.* Измерение политического знания: разработка и апробация шкалы в России // Социология: методология, методы, математическое моделирование (Социология:4М). 2021, № 52. С. 61–94. DOI: 10.19181/4m.2021.52.3. EDN: ARWFHV.

3. *Сорокин П.А.* Дальняя дорога: автобиография. М.: Терра, 1992. 303 с. ISBN 5-239-01378-0.

4. *Inglehart R., Baker W.E.* Modernization, cultural change, and the persistence of traditional values // American sociological review. 2000. Vol. 65, № 1. P. 19–51. DOI: 10.2307/2657288. EDN: GSHGFR.

5. *Glänzel W., Schubert A.* Analysing scientific networks through co-authorship // Handbook of quantitative science and technology research. Springer: Dordrecht, 2004. P. 257–276. ISBN 978-1-4020-2702-4. DOI: 10.1007/1-4020-2755-9_12.

## Использованные источники в романском алфавите (латинице)

В случае принятия рукописи к публикации авторы по запросу редакции обязаны предоставить транслитерированный в латинице полный список литературы к своей статье.

Основные требования:

- названия цитируемых русскоязычных публикаций следует давать в виде перевода на английский с пометкой в скобках, что речь идет о работе на русском языке (in Russian);
- для переводных работ указывается исходное название источника на языке публикации, выходные данные – в транслитерированном виде;

- названия источников (журналов), а также фамилии авторов желательно давать в том виде, в каком они индексируются в зарубежных базах научного цитирования (например, *Web of Science* или *Scopus*);
- название источника может сопровождаться его переводом на английский язык, например: Sotsiologiya 4M (Sociology: methodology, methods, mathematical modeling);
- при составлении списка источников недопустимо использовать российский ГОСТ, в частности запрещено в качестве разделительных знаков использовать «//» и «–»;
- название источника выделяется курсивом;
- указывается DOI источника (при наличии), но не указывается ISBN и EDN.

Примеры библиографических описаний на английском языке:

1. Durkheim É. *L'éducation morale* (transl., in Russian). Moscow: HSE University, 2021. 456 p. DOI: 10.17323/978-5-7598-2530-2.
2. Klimova A., Artamonov G., Chmel K. Measuring political knowledge: development and testing the scale in Russia (in Russian), *Sotsiologiya 4M* (*Sociology: methodology, methods, mathematical modeling*), 2021, no. 52, p. 61–94. DOI: 10.19181/4m.2021.52.3.
3. Sorokin P. *A long journey* (transl., in Russian). Moscow: Terra, 1992. 303 p.
4. Inglehart R., Baker W.E. Modernization, cultural change, and the persistence of traditional values, *American sociological review*, 2000, vol. 65, no. 1, p. 19–51. DOI: 10.2307/2657288.
5. Glänzel W., Schubert A. "Analysing scientific networks through co-authorship", in: *Handbook of quantitative science and technology Research*, ed. by H.F. Moed, W. Glänzel, U. Schmoch. Springer, Dordrecht, 2004, p. 257–276. DOI: 10.1007/1-4020-2755-9_12.

# INFORMATION FOR AUTHORS

## About the journal

*"Sociology: Methodology, Methods, Mathematical Modeling" (Sociology: 4M)* is a peer-reviewed scholarly journal presenting pioneering work on problems of methodology and methods of sociological research, on the collection, measurement and analysis of social science data, on building mathematical models of social processes.

*Sociology: 4M* publishes articles that contribute to the development of sociological methodology, clarify the existing problems in this area and offer new solutions. In addition, it publishes analytical reviews of sociological methods.

Main topics:

- general issues of methodology and research methods ;
- statistical methods and data analysis;
- theory and methods of measurement, theory and history of methods;
- data collection procedures;
- the quality of social science data, online surveys;
- qualitative methods in sociology;
- social science experiments.

The research topics are not limited by this list; the decision on publication is made in each case individually.

## Publication frequency and access

The journal is published twice a year. Full-text versions of articles are available for open access on the journal's website.

## The review procedure

All papers submitted to *Sociology: 4M* are screened by the editors for general suitability. The reasons for the negative decision may include the absence of mathematical models verification based on original empirical data, insufficient description of the empirical data sources, inconsistency with current state of research on the problem, and the lack of scientific novelty. If the decision is positive the article is sent to formal review. The results of

this review are crucial to the acceptance or rejection of the manuscripts. All papers meeting basic editorial criteria are reviewed at least by one expert. If necessary, editors appoint a second reviewer for the paper.

The review procedure of is anonymous for both authors and reviewers. The experts receive manuscripts without any indication of the name and affiliation of the authors. Editors do not reveal the reviewers' names to authors and not discuss their qualification. Reviewers are selected from experts in the subject area. The editorial team will contact the authors by email with the results of peer-review within three months after submission; if the reviews have not been received by this date, a new target date is announced.

The editorial team provides authors the right to respond to comments of the reviewers and to clarify their own position.

The journal publishes only original research papers which have not been published before (except for electronic preprints and theses). By submitting a manuscript the author agrees not to publish it in whole or in part in any form without the editors' permission.

There are no publication fees in *Sociology: 4M*.

## *Article guidelines*

The total length of the manuscript shall not exceed 40.000 characters (including spaces). Materials should be submitted via e-mail.

The text including notes and bibliography must conform to the following standards:

- Font – Times New Roman;
- Font size – 12;
- Line spacing – 1.5;
- Justified text alignment;
- Page margins: 2 cm on all sides.

Formulas and symbols should be typed in *Microsoft Equation*.

The author should also provide an abstract, 8-10 keywords, author's name, affiliation and position, contact information.

The cover letter accompanying the manuscript should contain a description of scientific novelty and a brief justification of why the article may be of interest to «Sociology: 4M» readers.

In this letter the author must confirm that the article is based on an original study, has not been published anywhere before (except as a preprint or in conference abstracts) and is not under consideration in any other journal.

## Publication ethics statement

The editorial policy requires compliance with the requirements of publication ethics by all parties involved in the preparation of the article (authors, reviewers and editors), that provides the confidential review of manuscripts, absence of plagiarism or misappropriation of the results. The editorial board expresses its readiness to publish error reports and information about violations of publication ethics by authors.

## Information about the article in English

Authors should provide the following information in English: authors' names, titles, authors' affiliation (names of organizations, e-mail of the author responsible for correspondence), abstract and keywords. We recommend using the English organizations' names from such citation indexes as Web of Science and Scopus.

We also recommend supplying an extended English abstract (up to 250 words).

## List of references

Authors should provide references to all sources mentioned in the text. The author must indicate sources of all citations, numbers, and other information. Authors are responsible for the accuracy of quotes as well as numbers and other information.

References to unpublished works are not permitted. Bibliographic descriptions of publications should be made in accordance with Russian state technical standards (GOST).

## Transliteration of references

If the manuscript is accepted the authors must provide transliterated references within two weeks upon acceptance for publication. Basic requirements are as follows:

- publication titles should be given in English with a note that it is a work in Russian;
- for translated works (originally published in one of Romance languages) the title should be given in original language, the source – in transliterated form;
- we recommend to provide the author names and source titles (journals) in the form they are indexed in such citation databases as Web of Science and Scopus.
- the source title may be accompanied by a translation into English, for example: Sotsiologiya 4M (Sociology: methodology, methods, mathematical modeling);
- source titles should be given in italics.

**Социология: методология, методы, математическое моделирование**

Сетевое научное издание

Зарегистрирован Федеральной службой по надзору в сфере связи, информационных технологий и массовых коммуникаций (Роскомнадзор) Эл № ФС77-85872 от 4 сентября 2023 г.

ISSN 2949-463X

---

**Учредители**

Федеральное государственное бюджетное учреждение науки
Федеральный научно-исследовательский социологический центр
Российской академии наук (ФНИСЦ РАН)
Адрес: 117218, Москва, ул. Кржижановского, д. 24/35, к. 5
Сайт: https://www.fnisc.ru. Телефон: 8 499 125-00-79

Общественная организация «Российское общество социологов»
Адрес: 117218, Москва, ул. Кржижановского, д. 24/35, к. 5
Сайт: https://www.ssa-rss.ru. Телефон: 8 499 719-09-71

---

Главный редактор – Девятко И. Ф.

Журнал «Социология: методология, методы, математическое моделирование» включен в базу РИНЦ, перечень ВАК

Журнал открытого доступа. Доступ к контенту журнала бесплатный. Плата за публикацию с авторов не взимается

Адрес редакции: 117218, Москва, ул. Кржижановского, д. 24/35, к. 5
Электронная почта редакции: sociology.4m@gmail.com
Телефон редакции: 8 499 391-02-80.
Официальный сайт журнала: https://www.soc4m.ru

2023. № 56. Дата выхода в свет 17.07.2024