Санкт-Петербургский государственный университет

На правах рукописи

Краснов Федор Владимирович

МЕТОДОЛОГИЯ ПОСТРОЕНИЯ ЦИФРОВОГО ДВОЙНИКА НАУЧНО-ТЕХНИЧЕСКОГО ЦЕНТРА В НЕФТЕГАЗОВОЙ ОТРАСЛИ

Научная специальность 05.13.01 — «Системный анализ, управление и обработка информации (технические науки)»

ДИССЕРТАЦИЯ
на соискание учёной степени
доктора технических наук

Научный консультант:
доктор технических наук, доцент
Дегтярев Александр Борисович

Санкт-Петербург — 2020

Оглавление

Стр.

Стр.

Стр.

Стр.

Общая характеристика работы

Качественный скачок в структуре и динамике развития производительных сил обеспечивается деятельностью отраслевых научно-технических центров (НТЦ). Количество НТЦ в энергетической отрасли растет из года в год, а по мере исчерпания запасов легко добываемой нефти роль научной составляющей в ее добыче возрастает. Поэтому эффективность деятельности НТЦ является ключевой характеристикой, нуждающейся в оценке и планировании.

Рассмотренные в данной работе вопросы, касающиеся методов оценки НТЦ, позволяют определить перечень наблюдаемых характеристик, обеспечивающих достоверную оценку научно-технических центров и позволяющих как сравнивать их, так и выстраивать математические модели для сценарного планирования их эффективности. Традиционно НТЦ создавались по лекалам российских научно-проектных институтов, которые оценивали запасы месторождений нефти и газа, ставили их на государственный баланс и формировали проектные документы на разработку месторождений. В задачи таких институтов также входила разработка и внедрение новых технологий и материалов, но часто проявлялась их главная уязвимость - обособленность от бизнеса.

Изначально, после получения разрешительных документов в Центральной комиссии по разработке ме-

сторождений, НТЦ отходил в сторону, и в дело вступали производственники. Современные НТЦ представляют собой научно-проектную структуру, которая полностью интегрирована в производство. Оценка деятельности таких НТЦ нуждается в пересмотре.

Таким образом, объектом исследования данной работы являются результаты научной деятельности НТЦ. Предметом исследования являются методы измерения, оценки и планирования результатов деятельности.

Цели работы

1. Комплексный анализ, диагностика и моделирование социальных процессов в организационной среде,
2. Синтез путей решения проблем обратной связи при прогнозировании развития науки по определенным приоритетным направлениям, что имеет большое значение для хозяйственного развития научно-технологического комплекса России,
3. Разработка методов сбора и анализа результатов деятельности НТЦ для создания и обучения модели эффективности НТЦ с использованием алгоритмов машинного обучения и практик работы с «больши-

ми данными». Построение прогнозов о результатах деятельности НТЦ.

Задачи исследования

1. Определение и разработка методов сбора результатов научной деятельности НТЦ
2. Определение результатов научной деятельности НТЦ, влияющих на научную эффективность НТЦ
3. Разработка интегральных показателей эффективности научной деятельности НТЦ
4. Выбор класса моделей для интегральных показателей эффективности научной деятельности НТЦ
5. Создание модели на основе алгоритмов машинного обучения
6. Апробация модели на действующих НТЦ
7. Построение прогнозов "что-если"на основе созданных моделей.

Научная новизна работы состоит в комплексном подходе к разработке и исследованию алгоритмов сбора цифровых артефактов научной деятельности НТЦ, создании интегральных показателей научной эффективности с целью моделирования на основе алгоритмов машинного обучения, применяемые для выделения приоритетных

направлений финансирования и развития человеческого капитала в энергетической отрасли.

1. Предложена формализация процесса самоорганизации команд для достижения определённой цели – написания научных статей,
2. Разработан детальный алгоритм образования соавторств,
3. Исследована временная зависимость структуры соавторств,
4. Создана модель для прогнозирования соавторств,
5. Создана модель укрупненного виденья научных направления развития НТЦ на основе публичных данных о публикационной активности сотрудников,
6. Создана модель движения персонала в организации и модель выполнения наукоёмких заданий,
7. Разработан математический аппарат построения графов соавторства на основе двудольного графа,
8. Создана общая математическая модель НТЦ,
9. Проведен эксперимент по многоагентному симулированию, в котором в качестве агентов выступали научные сотрудники лабораторий, взаимодействующие друг с другом и производящие в качестве результата своей работы научные статьи.

Научные результаты, полученные в работе, нашли применение в практике научно-методической работы по сопровождению добычи углеводородов.

Положения, выносимые на защиту:

1. Обобщение алгоритмов сбора цифровых артефактов деятельности НТЦ.
2. Методология построения интегральных показателей эффективности научной деятельности НТЦ.
3. Комплексная модель научной эффективности НТЦ.
4. Алгоритм управления технологической стратегией НТЦ.
5. Результаты численного моделирования научной эффективности НТЦ.

Методы исследования базируются на теориях компьютерной лингвистики, искусственного интеллекта, социальных сетей, информационного поиска, графов, машинного обучения, имитационного моделирования, системной динамики, теории моделирования интеллектуальных процессов и на математической статистике.

Обоснованность и достоверность полученных результатов обеспечивается корректностью применяемого математического аппарата, строгостью утверждений и наложенных ограничений, результатами комплексных исследований с использованием компьютерного моделирования.

По теме исследования опубликовано более **60** работ; **32** из них опубликованы в рецензируемых научных изданиях, рекомендованных ВАК. Доклады автора опубликованы

в **20** сборниках из списка Web of Science и Scopus. Получены **4** свидетельства о регистрации программ для ЭВМ.

Глава 1. Введение

1.1 Цифровые двойники

Конкуренция побуждает бизнес к принятию концепции цифровых двойников. В каких-то отраслях цифровые двойники уже нашли свое место: есть цифровые двойники заводов и городов. Но в отраслях, где основным продуктом являются знания, цифровые двойники еще не столь востребованы. Нет сомнений в том, что научные организации имеют внутренний стержень, корпоративную культуру, ценности, которые и позволяют им выполнять уникальные научно-исследовательские работы. Моделируя эти внутренние, скрытые особенности организации можно получить уникальные инструменты для прогнозирования и управления технической стратегией. Цифровой двойник научно-технического центра можно рассматривать как особый тип модели системы, которая выявляет такие скрытые явления, как появление новых научных направлений, образование научных школ и степень творческого выгорания и усталости коллектива.

Прототипирование, как инженерная дисциплина, существует уже более 30 лет и на первый взгляд занимается тем же. Но в отличие от прототипа цифровой двойник не

ставит задач по быстрой реализации базовой функциональности для анализа работы системы в целом.

Таким образом, чтобы понять преимущества цифровых двойников нужно понять те новые возможности, которые они предоставляет. Идя от частного к общему, автор выбрали для исследования научно-технический центр и рассмотрели подходы к построению цифрового двойника, а затем обобщили эти подходы. В результате получена актуальная постановка исследовательских гипотез, которые нужно проверить прежде чем создавать цифровой двойник организации или ее части, нацеленной на производство новых знаний [1].

Концепция цифровых двойников (ЦД) не нова. Цифровые двойники относятся к направлению по цифровому представлению физических систем, и на протяжении более 30 лет команды разработчиков продуктов и процессов использовали 3D-рендеринг моделей автоматизированного проектирования, моделей активов и моделирования процессов для обеспечения и проверки технологичности. Роскосмос, например, десятилетиями проводил сложные симуляции космических кораблей, а центр управления полётами дублирует большинство процессов.

В настоящее время совместное влияние сразу нескольких факторов, побудило выдвинуть концепцию цифрового двойника на передний план как прорывную тенденцию, которая будет иметь все более широкое и глубокое влияние на экономику в течение следующих пяти лет. Фактически,

Gartner прогнозирует, что к 2021 году половина крупных промышленных компаний будет использовать цифровых двойников, в результате чего эффективность этих организаций увеличится на 10%.

Одним из факторов, повлиявших на становление концепции цифровых двойников является лавинообразный рост узлов в сети Интернет. Представление в Интернет началось для компаний с создания корпоративного сайта и пришло к пониманию цифровой экосистемы предприятия, которая как бы выворачивает часть бизнес процессов компании наизнанку. Короткая максима такова: «Если этого нет в Интернет, этого не существует». Так развитие цифрового маркетинга в Интернет привело к необходимости продления потребления продукта до вовлечения потребителя в создание новых продуктов. Сугубо внутренний процесс разработки продукта теперь представлен в полностью прозрачном виде и готов для потребления. Отчасти этот подход продлевают и краудфандинговые платформы, в которых будущие потребители еще и вкладывают денежные средства в создание продукта.

Нефтегазовая индустрия впитывает в себя все новые тренды и концепция ЦД не исключение. Вот несколько исследований 2018 года по цифровым двойникам в нефтяной отрасли:

– Цифровые двойники ускоряют бурение [2; 3] и упрощают мониторинг процесса бурения [4],

- Позволяют точнее управлять жизненным циклом месторождения[5],[6],
- Позволяют лучше координировать действия инженеров, платформы [7] и окружающей среды [8],
- Позволяют ускорить строительство [9],
- Упрощают контроль качества при строительстве [10].

Научно-технические центры (НТЦ) в нефтегазовой отрасли имеют свою историю развития и отличаются от R&D в других отраслях высокой степенью вовлеченности в производство. Особенности НТЦ в нефтегазовой отрасли можно проследить с помощью таких факторов, влияющих на спрос их услуг, как цена на нефть, темп роста экономики, налоговая политика государства, объёмы добычи. Помимо общего тренда на увеличение наукоемкости добычи углеводородов, в связи с исчерпанием запасов "простой нефти", рост цены на нефть в США, например, приводит к увеличению количества заявок на патенты в нефтегазовой отрасли.

1.1.1 Научно-технический центр, как объект исследования

Совсем упрощенно можно считать, что НТЦ состоит из исследователей и процессов, среди которых наиболее важен процесс НИР. Проблема объективной оценки эффек-

тивности НИР находится в центре внимания исследователей уже давно, и это, в первую очередь, связано с вопросами финансирования, как бюджетного, так и в рамках грантов.

Эффективность организации – очень сложный и многогранный концепт. На него оказывают влияние различные факторы. Одним из важных предвестников рыночного успеха научно-исследовательской организации является хорошо развитая коммуникация и кооперация между сотрудниками. Многие теоретические и практические исследования демонстрируют связь между продуктивностью организации и структурой коммуникации её сотрудников, например, см. [11; 12].

Исследование социальной структуры организаций и профессиональных сообществ становится одним из главных направлений прикладного анализа социальных сетей. В сфере общественных связей и управления глубоко изучаются модели коммуникаций внутри организаций, организация рассматривается как социальный объект. Начало этим исследованием положено в работе C.H.Cooley "Социальная организация"[13].

Научная публикация – главный артефакт для оценки эффективность научно-исследовательской работы. Процесс публикации является достаточно длительным: начиная с зарождения исследовательской идеи, проведения эксперимента и заканчивая публикацией работы. Организационные условия для исследователей могу по-разному влиять на производительность процесса публикаций.

Исследование [14] показало, что за последнее десятилетие есть чётко выраженная тенденция учёных объединяться в группы соавторов для публикации статей. Отсюда можно сделать вывод, о том, что наиболее важным фактором, положительно влияющим на публикацию работ, является объединение исследователей в команды.

В свою очередь, командообразование тоже бывает успешным и неуспешным; оно также поддаётся изучению, в результате которого можно выделить условия успешного командообразования. Задача поиска оптимальных параметров команды соавторов для наиболее продуктивного написания научных статей относится к классу задач оптимизации. Традиционно исследователи обращают внимание на следующие параметры, имеющие значение для продуктивного научного творчества:

- Размер команды
- Ментальные модели сообщества
- Компетенции сотрудников (дополняющие и гомофильные)
- Слабые связи между учёными

Научная кооперация между индивидами со схожими характеристиками более вероятна, однако уровень этой схожести тоже важен. В работе [15] было показано, что социальная схожесть более чем по одному показателю приводит к тому, что люди с меньшей вероятностью будут формировать между собой взаимоотношения. Авторы данного исследования объясняют этот наблюдаемый эффект тем,

что слишком схожие по многим характеристикам люди, как правило, не могут привнести что-то новое и конструктивное в личные отношения или же в команду. Для продуктивного сотрудничества необходима не только схожесть интересов, но также и различный профессиональный и жизненный опыт, позволяющий предложить многомерные подходы к решению общей задачи.

Одним из результатов такого сотрудничества является научная статья и в самой ее простой форме - это текст, который можно анализировать различными автоматизированными средствами.

Анализ текста иногда называют *Text Mining*. Суть этого процесса в превращении данных (текста) в высококачественную информацию способную приносить знания. Важным моментом является то, что при получении этих знаний человеческие затраты должны быть минимальны.

Полученные из текста знания становятся основой для принятия управленческих решений в организационной среде. Отдельным процессом рассматривается получение текста для исследования, иногда называемое созданием корпуса текстов. Описываемые явления, процессы и закономерности находят свое отражение в текстах при помощи специалистов-авторов, а процесс анализа текста специалистом-читателем делает обратное: на основе текстов составляет информацию о реальной природе вещей. Многомодовым подходом к анализу текстов является принятие во внимание сопутствующей основному тексту информации. Например,

модами могут стать название журнала, номер выпуска, должности соавторов научной статьи.

Формально анализ текста производится в следующей последовательности:

1. анализ языка текста
2. анализ содержания текста
3. получение информации об авторе текста
4. вывод определённых переменных, характеризующих природу вещей в тексте

Вместе с текстом можно анализировать авторов как социальные атомы в производственном процессе, обладающие различными связями. В книге [16] отмечается, что базисом для анализа социальных сетей является теория социометрии, основоположником которой принято считать J.L.Moreno [17]. Социометрия изучает взаимоположения социальных атомов в группах. Социограммой по Морено является графическое отображение социального выбора членов социальной группы. В рассматриваемой модели социализации - процессе создания и публикации научной работы, социальным выбором может быть выбор лидера, дружеские отношения между членами научного коллектива разработчиков и/или соавторов, выполнение совместных задач. Социограмма представляет граф, состоящий из вершин ребер.

Граф соавторства является частным случаем социальной сети. Одним из первых исследований графа соавторства является работа [18], сделанная в 1973 году. С этого време-

ни исследования научной деятельности при помощи графов соавторства не прекращались и обрели статус проверенного инструмента анализа. Например, в недавнем исследовании [19] предпринята попытка предсказания будущих научных исследований на основе графа соавторства, а в работе [20] построен глобальный граф соавторства на основе Google Scholar, который содержит более 400 тысяч вершин. Оба исследования проведены в 2017 году и учитывают новейшие достижения в данной области.

Таким образом, мы выделили три основных направления исследования НТЦ, которые помогут создать цифровой образ организации - это командообразование, анализ текста и анализ графа соавторства.

1.1.2 Вопросы к цифровому двойнику научно-технического центра

Рассмотрим построение цифрового двойника НТЦ как парадигму моделирования. Есть объект, который изменяется как по внутренней структуре, так и по внешним проявлениям. Законы, по которым происходят внутренние изменения объекта нам не известны. Но мы можем наблюдать внешние проявления этих изменений: количество и качество выполняемых НИР. Тогда нас будут интересовать следующие вопросы:

1. В какой степени научная статья отражает проведенную НИР? Можно ли судить о качестве НИР по опубликованным научным исследованиям?
2. Каковы социальные механизмы объединения исследователей для проведения НИР? Какие виды компетенций и в какой степени влияют на такое объединение?
3. Как зависит время проведения НИР от количества участвующих исследователей? Существуют ли естественные ограничения на количество и состав исследовательских групп и на чем они основаны?
4. Каковы эвристические алгоритмы поведения исследователей по отношению к издательствам и программным комитетам конференций? Существуют ли базовые стратегии поведения? Если возможность идентификации и имитации базовых стратегий?
5. Применимы ли подходы time management (“управление временем”) к НИР? Насколько эффективно рассмотрение научно-исследовательской деятельности как проектной деятельности?
6. Какова модель зрелости научно-исследовательской организации [21] в части проведения НИР? В какой степени возможно определение степени зрелости научно-исследовательской организации на основе анализа публикуемых ею научных статей?

7. Какова структура процессов, составляющих научно-исследовательскую деятельность? Насколько применим процессный подход к изучению научно-исследовательской деятельности? Есть показатели научно-исследовательской деятельности, отражающие характерную структуру составляющих ее процессов?

В рамках этого направления исследования можно сформулировать две взаимодополняющие постановки задачи: прямую и обратную.

– *Изучение деятельности НТЦ по внешним проявлениям.* К внешним проявлениям относятся цифровые артефакты деятельности организации - это опубликованные научные статьи, материалы конференций, информационные сайты в сети Интернет и новости о компании.
– *Изучение НТЦ изнутри.* К исследованиям в этом направлении относятся моделирование научной деятельности, эффективность производственных процессов, самоорганизации малых творческих коллективов и модели персонала научной организации.

Архитектура методического каркаса для изучения прямой и обратной задач представлена на рисунке 1.1.

Разделение НТЦ на составляющие подсистемы позволяет разрабатывать модели подсистем с использованием наиболее эффективных математических методов. Авторы считают уместным применение Байесовских методов для

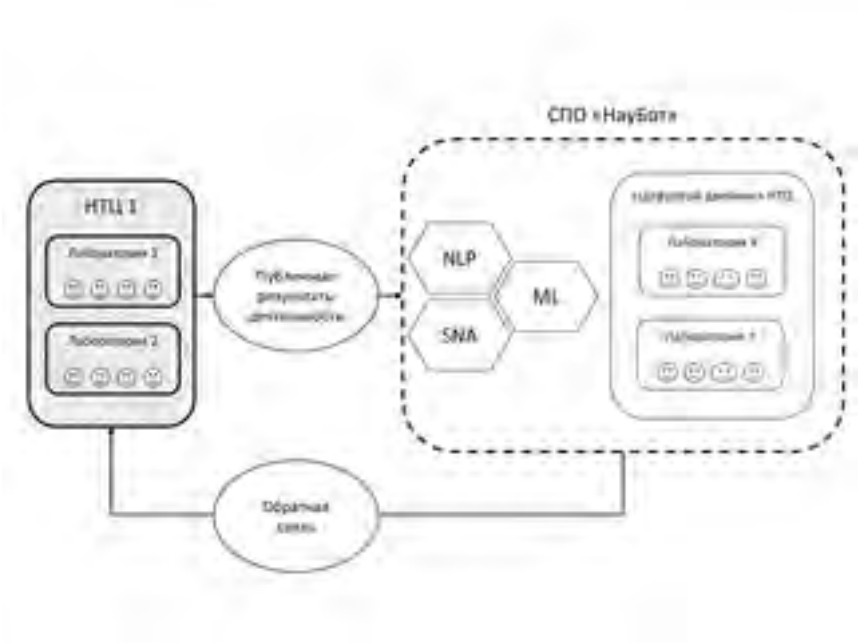


Рисунок 1.1 — Методический каркас исследования.

определения параметров цифрового двойника НТЦ при решении обратной задачи.

Пусть дана функция $\Phi(x)$ и нам нужно найти x при котором она достигает максимума $\Phi(x) \longrightarrow \max_x$. Добавим условие, при котором расчет каждого значения $\Phi(x)$ – это ресурсоемкая задача. Такое условие встречается, например, в следующих случаях:

– x - это географические координаты скважины, а $\Phi(x)$ - это количество нефти, которое можно добыть, пробурив скважину с координатами x. В

таком случае одно значение $\Phi(x)$ стоит миллионы рублей;

– x - это гиперпараметры искусственной нейронной сети глубокого обучения, $\Phi(x)$ - это целевая метрика точности предсказания. В этом случае одно значение $\Phi(x)$ будет занимать месяцы работы;

При решении прямой задачи моделирования наиболее продуктивным подходом представляется имитационное моделирование сложных систем. Допустим, что в отраслевой научно-исследовательской организации Ω работают лаборатории λ_i , где $i \in (1 \dots N_\lambda)$. Обозначим множество лабораторий $\Lambda = \{\lambda_i, \dots, \lambda_{N_\Lambda}\}$. В лабораториях работают научные сотрудники $A = \{a_i, \dots, a_{N_A}\}$. Обозначим множество тематик t_i, где $i \in (1, \dots, N_T)$, по которым организация Ω ведет НИР как $T = \{t_1, \dots, t_{N_T}\}$. Тогда деятельность организации Ω по выполнению НИР может быть описана следующими компонентами :

$$\mathbb{M}_\Omega = \left\{ S, \Xi, \Psi, E \right\}, \text{ где } S = \{\Lambda, A, T, P, X\} \tag{1.1}$$

– $\Xi = \{\xi_1, \dots, \xi_{N_\Xi}\}$ – множество связей между субъектами,

– $\Psi = \{\psi_1, \dots, \psi_{N_\Psi}\}$ – множество действий субъектов,

– $P = \{\rho_1, \dots, \rho_{N_P}\}$ – множество научных работ,

– $X = \{\chi_1, \dots, \chi_{N_X}\}$ – множество научных журналов и конференций.

Задавая априорные распределения для таких событий как возникновение научной идеи исследования, встречи со-

авторов, подачи статьи на конференцию, можно определять с помощью имитационных расчетов результат работы НТЦ.

В настоящем исследовании автор сформировали общую расширенную постановку целей и методов для исследования научно-технической деятельности с помощью методологии цифрового двойника. Как показал автор, методика цифрового двойника может быть применена для:

- комплексного анализа, диагностики и моделирования социальных процессов в организационной среде,
- поиска путей решения проблем обратной связи при прогнозировании путей развития науки по определенным приоритетным направлениям,
- построения прогнозов о результатах деятельности НТЦ.

Из вышеизложенного следует, что методология ЦД для изучения НТЦ представляет иерархию моделей. Автор показал, что для решения прямой задачи моделирования НТЦ необходимо разрабатывать модели персонала, модели выполнения интеллектуальных заданий и модели образования малых команд. В свою очередь, для решения обратной задачи моделирования НТЦ необходимо строить модели научного текста, модели соавторства и модели публикации научных результатов. Каждая из этих моделей требует отдельной проработки внутренних механизмов и механизмов взаимодействия между собой.

Объединение перечисленных моделей с помощью имитационного многоагентного моделирования и являет-

ся цифровым двойником НТЦ, который поможет повысить эффективность работы организации и позволит улучшить управление технической стратегией нефтяного холдинга.

1.2 Возникновение Научно-технических центров

Ряд исследований последнего времени демонстрируют уверенную корреляцию между ростом цен на нефть и объемом капиталовложений в перспективные исследования и разработку новых технологий в нефтяной отрасли. Оптимальным для инновационных инвестиций диапазоном цен на нефть можно признать в современных условиях диапазон в 60-70 USD за баррель. При значениях цены в районе 50-55 и меньше USD за баррель нефтедобывающая отрасль попадает в режим выживания с соответствующей жесткой оптимизацией всех расходов. При цене более 80 USD за баррель возникает известный эффект эйфории с предпочтением вложений прибыли в иные секторы экономики с предполагаемой быстрой отдачей, в частности в спекулятивные финансовые инструменты и рынки. Ситуация несколько отличается для сектора Downstream, поскольку дорогое сырье стимулирует потребность в более глубокой его переработке. Однако в настоящее время в традиционных процессах нефтепереработки достигнут определенный

технологический предел, а внедрение новых процессов требует преодоления известного психологического барьера со стороны владельцев нефтеперерабатывающих производств. Резкие колебания цен на нефть и вызванные ими потенциальные решения картелей (например, ОПЕК) создают общий нервозный фон в отрасли, который не способствует инновационным финансовым инвестициям. Таким образом, финансовые вложения в разработку и развитие новых технологий носят импульсный во времени характер, привязанный к колебаниям цен на нефть. В то же время разработка, апробация и внедрение новых технологий требуют времени существенно большего, чем длительность спекулятивного делового цикла рынка углеводородного сырья. Более того, многие технологии стадии старт-ап или даже более зрелые потребуют для своей доработки и индустриального внедрения дополнительных средств. При этом не каждый пик инвестиционно-инновационной активности принесет средства в бюджет разработки данной конкретной технологии. Технологических идей все еще достаточно много, также имеет место конкурентная борьба научных групп и направлений за выделяемые средства. Инвесторы по причинам психологического и поведенческого характера могут вложить очередной транш инвестиций в какие-либо новые проекты вместо проектов, находящихся в стадии активной разработки, но еще не продемонстрировавших с точки зрения менеджмента свою практическую эффективность. На основании изложенного можно сделать вывод, что канди-

датами на выживание являются технологические проекты, которые могут быть доведены на средства первого инвестиционного транша как минимум до стадии feasibility, а лучше до стадии pilot plant.

Несколько иная ситуация в газовой отрасли. Газ является более дешевым сырьем, процесс его добычи и транспортировки в известном смысле более технологичен, а рынок более стабилен за счет больших постоянных объемов спроса со стороны систем производства электроэнергии, бытового и промышленного отопления, получения высокопотенциального технологического тепла и известных отраслей газохимии. Однако эти же перечисленные факторы одновременно и ограничивают инновационную активность и инвестиционно-инновационную привлекательность в газовой отрасли. Развитие газохимии в плане новых технологий переработки газа является привлекательным с теоретической точки зрения и может составить в перспективе достойную конкуренцию ряду традиционных направлений нефтехимии. Однако на практике технология энергетически выгодной конверсии метана все еще не разработана, а существующая технология через паровой или паро-кислородный реформинг может конкурировать по затратам с нефтепереработкой только при ценах на нефть от 90 USD за баррель. Что касается процессов переработки высших углеводородов, то они в известной степени развиты и сырьем для них является детандерный отбор (разделение) природного газа на фракции. Однако сырьем

для этой же группы технологий могут служить и попутные газы нефтепереработки, прежде всего этилен, каковые процессы реализованы на многих нефтеперерабатывающих производствах.

Отдельно следует отметить перспективную роль угля при использовании его как в качестве топлива, так и химического сырья. Ключевым в обоих сферах является промышленное внедрение эффективных технологий газификации и пиролиза с полным циклом кондиционирования и очистки получаемого продукта. Несмотря на все конъюктурные перипетии текущей ситуации на рынке углеводородного сырья использование угля остается важным в долгосрочной перспективе для таких индустриально-развитых стран как США, Германия, Китай, ЮАР, РФ. Украина и Казахстан. Мы здесь не случайно отнесли РФ, Украину и Казахстан к индустриально-развитым державам, хотя кто-то и может сказать, что такое отнесение имеет условный характер. Действительно, указанные государства находятся на экономическом перепутье, но все еще обладают как достаточно мощным промышленным потенциалом так и сырьевыми возможностями. От взвешенной инвестиционной и инновационной-технологической политики этих государств, и в первую очередь в топливно-энергетическом секторе экономики, зависит, войдут ли они в клуб ведущих мировых экономических игроков или и далее будут подвержены дезинтеграционным и деградационным процессам.

Помимо краткосрочных и долгосрочных экономических тенденций на отрасль добычи и переработки углеводородного сырья, и в частности на ее инновационно-технологический сектор, оказывают существенно влияние социальные и политические факторы. Так для США, равно как и для транснациональных корпораций на углеводородном рынке, актуальны экологические проблемы, которые можно подразделить на локальные (воздействие на окружающую среду в местах непосредственной добычи и переработки углеводородного сырья) и глобальные (парниковый эффект, загрязнение мирового океана, загрязнение подземных вод, в частности при применении новых популярных технологий добычи сланцевой нефти и сланцевого газа). Для стран Восточной и Западной Европы актуальна политическая проблема зависимости от поставок газа из Российской Федерации и поиска альтернативных источников топлива и химического сырья.

Следовательно эффективность деятельности по разработке, развитию и внедрению в производство новых технологий в углеводородной отрасли должна оцениваться на основе многокритериального подхода, в котором должна быть учтена конъюктурная (определяющая текущие инвестиции), экономическая долгосрочная, собственно технологическая и социально-политическая составляющие. Это требует привлечения методов многофакторного анализа с использованием новейших алгоритмов из области Data Mining, Big Data Analysis [22], neuroscience [23], мето-

дов машинного обучения, поиска, систематизации и анализа цифровых артефактов деятельности научно-технических центров и лабораторий, семантического и компьютерного лингвистического анализа текстов [24] и т.п.

Для Российской Федерации имеется список специфических проблем, которые могут быть отнесены как к социально-политической, так и к технологической сфере.

Для обзора указанных проблем обратимся к краткой истории нефтяной и газовой отрасли в РФ. Временем начала индустриальной добычи нефти считается вторая половина девятнадцатого века, однако с незапамятных времен нефть добывалась открытым способом в местах ее выхода на поверхность и использовалась проживающими в тех местностях людьми в различных целях, которые жили в разных уголках мира, где нефть просачивалась на поверхность. Согласно письменным источникам в России племена, проживавшие на территории Тимано-Печерского района, в частности по берегам реки Ухты собирали нефть с поверхности водоемов и использовали ее в качестве смазки, а также для медицинских целей. Нефть из этого региона была впервые доставлена в Москву в 1597 году. 1684 годом датируется донесение об обнаружении нефти начальника Иркутского острога Леонтия Кислянского. В 1703 году в первом выпуске газеты “Ведомости” было напечатано сообщение об обнаружении нефти на реке Сок в Поволжском регионе. Позднее появились сообщения о добычи нефти местными жителями на Северном Кавказе.

Местные жители добывали нефть с помощью ведер из скважин глубиной 1-2 метра. Использование нефти носило в основном медицинский характер. О проявлениях нефти и газа на западном побережье Каспийского моря еще в 10-ом веке сообщали арабские путешественники и историком еще в десятом веке. Согласно данным итальянского историка и путешественника Марко Поло люди в этом регионе использовали нефть в медицинских целях и в религиозных целях. С четырнадцатого века нефть с побережья Каспия поставлялась в страны Среднего Востока.

Первая попытка организации нефтеперерабатывающего производства может быть отнесена к 1745 году, когда уроженец Архангельска Федор Прядунов получил разрешение добывать нефть на реке Ухте в уже упомянутом ранее Тимано-Печерском районе. Прядунов также создал нефтеперегонный куб и ряд продуктов нефтеперегонки поставлял в Москву. Однако указанная технология не получила дальнейшего развития, поскольку в течение всего XVIII века практическое применение нефти и продуктов из нее оставалось крайне узким. Не изменилась существенно данная ситуация и в первой половине XIX века. Тем не менее 1823 годом датируется ввод в эксплуатацию нефтеперегонного завода братьев Дубининых, сырьем для которого служила нефть открытого Вознесенского месторождения недалеко от города Моздок.

Расширении Российской Империи на прикаспийский регион в начале XIX века и присоединение Северного Кав-

каза обозначили эти два региона как основные по части нефти. Первая в мире нефтяная скважина разведочного характера была пробурена на Биби-Айбатском месторождении Апшеронского полуострова (неподалеку от Баку) в 1847 году, что более чем на десять лет опередило старт нефтяной индустрии в США. Однако первая полноценная эксплуатационная скважина близкая по своему устройству к современным скважинам была введена в строй на Кубани на р. Кудако в 1864 году.

1849 год по праву можно считать поворотным в мировой нефтяной индустрии, т.к. канадский геолог Абрахам Геснер получил в этом году из нефти керосин как стабильный продукт с воспроизводимыми свойствами. В 1853 году львовские аптекари Иван Лукасевич и Ян Зех изобрели безопасную керосиновую лампу, что ознаменовало начало эры широкого потребления нефти.

Нефтеперерабатывающий завод прямого действия для производства керосина был запущен в Баку в 1863 году под руководством инженера Давда Меликова. Несколькими годами позже им же был основан нефтеперерабатывающий завод в городе Грозном.

Тем временем в США в 1859 году в штате Пенсильвания пробурена первая скважина и начинается добыча нефти. Нефтяной промысел стремительно развивается и нефть транспортируют в стандартных деревянных бочках емкостью 42 галлона или 168 литров, изначально предназначенных для транспортировки соленой сельди. Так

появляется мера объема нефти 1 баррель, равная 42-м галлонам. В 1865 году для транспортировки нефти от нефтяных скважин на железнодорожную станцию Миллер Фарм Стэйшн был построен первый в мире нефтепровод с пропускной способностью 2500 баррелей в сутки. Этот узел послужил также прообразом нефтеналивных транспортных терминалов и кустовой (цветковой) схемы объединения нефтяных потоков с нескольких близкорасположенных скважин перед транспортировкой нефти по магистральному нефтепроводу.

В 1870 году Рокфеллер основал компанию Standard Oil, доля которой в нефтедобыче США менее чем за 10 лет выросла с 10% до 90%, что привело к введению в действие антимонопольного закона впервые в мире.

Интересно, что в 1871 году в России родился Иван Михайлович Губкин (1871-1939) – один из основоположников и создателей геологии нефти как отдельного раздела общей геологии. Губкин внес практически неоценимый вклад в развитие нефтяной отрасли России, и сегодня его имя присвоено Российскому государственному университету нефти и газа.

В России в районе города Баку первый нефтепровод был пущен в эксплуатацию в 1878 году. В отличии от США он соединил скважины с нефтеперерабатывающим заводом. А еще в 1877 году Россия впервые в мире освоила использование нефтеналивных судов (танкеров) для транспортировки нефти.

Поначалу государство в России было монополистом в нефтяной отрасли, однако к концу седьмого десятилетия 19-го века к нефтедобыче были допущены иностранные компании. На апшеронском полуострове была обнаружена большая концентрация месторождений с легко извлекаемыми запасами нефти, однако транспортировка нефти и продуктов нефтепереработки конечному потребителю была совершенно не налажена. Одним из ключевых достижений братьев Нобелей и семейства Ротшильдов в России явилось именно объединение нефтедобычи, нефтепереработки и транспортировки нефти и нефтепродуктов конечным потребителям в рамках единых коммерческих компаний. Именно в России в 1874 году появилась первая вертикально интегрированная нефтяная компания - "Бакинское нефтяное общество". Нефтяная промышленность России демонстрировала в этот период существенный рост и к началу XX века доля России в полном объеме мировой нефтедобычи составляла около 30%. Интересно, что компания Шелл Транспорт энд Трейдинг, вошедшая позже в состав Роял Датч-Шелл, на первом этапе своей деятельности осуществляла перевозку бакинской нефти из России в Западную Европу.

Процессы нефтедобычи и нефтепереработки не остались вне сферы интересов российской науки того времени. Среди российских ученых, внесших в клад в нефтяную науку и практику можно отметить химика Зелинского, математиков и механиков, Л.С. Лейбензона, И.П. Москалькова,

И.А. Чарного, В.Н. Щелкачева, Я.И. Хургина и многих других ныне признанных классиков.

Основой нефтяной науки стали достижения органической химии, а также аппарат теоретической механики, механики грунтов и горных пород, гидромеханики. Был развит и достиг высокого совершенства, аппарат дифференциальных уравнений в частных производных, описывающих перенос флюидов в пористых средах на основе феноменологических представлений, таких как закон Дарси.

Большую роль в становлении науки о нефти в России сыграл Дмитрий Иванович Менделеев. В начале 90-х годов основная доля научных интересов ученого была связана с вопросами нефтехимии и нефтепереработки. Так Менделеев предложил способ непрерывной дробленой перегонки нефти, аналитические методы определения состава продуктов перегонки нефти, предложил использование селективных растворителей. Он неустанно доказывал необходимость использования всех фракций нефти, включая тяжелые. Им было предложено использование в осветительных лампах вместо керосина солярового масла. Также он способствовал строительству в городе Рыбинске, благодаря чему вместо ежегодного убытка в размере около 100 000 рублей в ценах того времени (затраты на покупку смазочных масел) Россия вскоре приобрела несколько миллионов рублей ежегодно от экспорта таких смазочных масел.

Менделеев выступал против преимущественного использования продуктов нефти в топках паровых котлов.

“Топить можно и ассигнациями”, – писал он в одной из своих экономических статей, обосновывая целесообразность использования нефти в качестве химического сырья, а угля – в качестве топлива.

Еще в 1881 году Менделеев предложил изучить возможность термической глубокой переработки нефти путем пропускания ее через трубы с температурой 300–400 градусов Цельсия. Он предполагал, что такой термической переработки следует подвергать и тяжелые остатки нефтеперегонки, с целью получения из них дополнительного количества годных продуктов. Эти идеи были тем более важны, что российская нефть была более плотной по сравнению с американской и от ее перегонки оставалось больше тяжелых масел и иных остатков. Менделеев был сторонником абиогенной концепции нефтеобразования посредством взаимодействия раскаленных карбидов железа и никеля с водой в ранние геологические эпохи Земли.

Большое внимание Менделеев уделял рациональной организации производственного цикла добычи и переработки нефти. Он предложил размещать нефтеперегонные заводы не только вблизи скважин (месторождений), но и на берегах Волги, где в то время была большая концентрация промышленного производства. С его участием был основан один из старейших в России нефтеперерабатывающий завод в Ярославле.

Известная полемика с Нобелем, который был сторонником широкого использования нефти как топлива,

а также зачастую отдавал распоряжения попросту выливать отогнанный бензин, т.к. для него в то время не находилось еще достаточных применений, демонстрирует как раз противостояние научно-технологической и экономико-конъюнктурной концепций при оценке эффективности производства, о чем говорилось в первых разделах данной главы.

Менделеев выступал за строительство нефтепровода и керосинопровода Баку-Батуми. Он писал: “С нефтепроводом спрос сырой нефти возрастает, и цены на нее урегулируются, потому что явятся новые места сбыта, а потому явятся и новые буровые в самом Баку и других местах Кавказа, чего и должно желать”.

Изобретение в 90-х годах XIX века двигателей внутреннего сгорания, в частности дизельного двигателя, и зарождение автомобильной индустрии еще более повысило спрос на нефть и привело к развитию технологий более глубокой переработки нефти. Наряду с керосином появились такие фракции как бензин и лигроин. Остатки нефтепереработки получили использование в качестве смазочных масел в машинах и механизмах.

Однако драматические события в России, связанные с Первой Мировой войной и революцией 1917 года привели к падению нефтедобычи и потере Россией главенствующего положения на рынке углеводородного сырья. Если в 1913 году в России было добыто более 9 млн тонн нефти, то в 1920 эта цифра уменьшалась более, чем на 40

%. Страны Антанты пытались отделить нефтеносные районы от территории Советской республики, но в конечном счете потерпели поражение. В результате в 1920 году братья Нобель продали значительную часть своих российских активов компании Стандард Ойл из Нью-Джерси. Позже данная компания стала основой компании Экссон. Стандард Ойл выступила против решений советского правительства о национализации нефтяных месторождений и отказывалась от дальнейшего сотрудничества с советской властью. Напротив, нью-йоркские нефтяные компании (впоследствии преобразованные в компанию Мобаил) продолжили осуществление инвестиций в российскую нефтяную отрасль, так что к 1923 году экспорт нефти и нефтепродуктов из России снова достиг дореволюционного уровня.

Таким образом уже в 20-е годы сформировалась частичная зависимость российской (советской) отрасли нефтедобычи от западных капиталов и западных технологий. Для устранения этой негативной ситуации советское правительство приняло в частности решение об интенсивной подготовке собственных кадров в области нефтяного инженерного дела и нефтяной геологоразведки.

Огромную роль в реализации этой программы сыграл Иваан Михайлович Губкин — организатор советской нефтяной геологии, академик АН СССР (1929), вице-президент АН СССР (1936), председатель Азербайджанского филиала Академии Наук СССР (1936 — 1939), лауреат премии им. В.И.Ленина (1931). Депутат Верховного Совета

СССР 1-го созыва (1937). В отличии от Д.И.Менделеева И.М.Губкин выступал сторонником теории биогенного образования нефти. Он в частности писал: “Мы полагаем, что нефтеобразование, начавшись с разложения жиров в биогенном иле до его погребения, продолжалось и после его погребения при активном содействии анаэробных бактерий во весь период диагенетического изменения породы.” К сожалению теория нефтеобразования И.М.Губкина осталась неизвестной в рамках мировой науки, поскольку труды Губкина в то время не были переведены на иностранные языки.

В 1930 году под руководством И.М.Губкина вышел учебник “Учение о нефти”, по словам самого Губкина “излагающий главнейшие вопросы нефтеведения”. Основой учебника послужил курс лекций самого Губкина, однако были достаточно широко использованы и материалы других авторов. Так А.И.Косыгин был автором раздела “Основные приёмы разведки нефтяных месторождений”, а геофизик А.И.Заборовский написал главу “Элементы геофизических методов разведки”.

Александр Игнатьевич Заборовский доктор физико-математических наук, геофизик. Он был одним из основателей советской школы геологоразведочной геофизики и разработчиком программы подготовки специалистов в вузах по данному направлению. Заборовский – автор монографии “Геофизические методы разведки”, которая использовалась в учебных заведениях СССР в качестве учебника по прикладной геофизике.

В 1919—1926 годах Заборовский магнитометрическими работами на Курской Магнитной Аномалии. Работал он в одной команде с П.П.Лазаревым, А.Д.Архангельским, И.М.Губкиным, О.Ю.Шмидтом и другими видными российскими учеными того времени. В результате деятельности этой группы на территории курской магнитной Аномалии были выявлены значительные скопления железистых кварцитов, причем по выполненным оценкам общее количество железа в данном месторождении превосходило суммарные запасы железа, разведанные к тому времени в Европе.

В 1926 году Заборовский разработал ряд геофизических методов, основанных на данных сейсморазведки. С 1929 года он читал учебные курсы по геологоразведочной геофизике в МГУ, а с 1930 года возглавил созданные им факультет и кафедру геологоразведочной геофизики в Московском государственном геологоразведочном университете. В периоды с 1944 по 1949 и с 1954 по 1968 годы Заборовский также руководил кафедрой геофизических методов геологического факультета МГУ.

Даже на этих двух примерах деятелей советской геологической науки мы видим, что в 30-е, 40-е и 50-е годы наряду с практическими достижениями и теоретическими разработками большое внимание уделялось подготовке квалифицированных кадров для отрасли.

Вплоть до начала Второй Мировой войны каспийский регион и Северный Кавказ оставались основными районами нефтедобычи и нефтяной промышленности. Одной из

основных стратегических задач командования нацистской Германии был захват этих нефтеносных районов. Известно, что Германия не располагает собственными запасами нефти, поэтому Гитлер вступил в войну с бензином, производимым из ацетилена, который в свою очередь получался сложным и дорогостоящим способом электродугового пиролиза угля в среде инертных газов. После войны добыча нефти в прикаспийском регионе снова выросла и достигла в 1951 году рекордного уровня в 850 000 баррелей в день. Помимо собственно нефтедобычи Баку стал индустриальным центром по производству оборудования для нефтедобычи и нефтехимии в масштабах всего СССР. Однако советское правительство начало целенаправленные работы по поиску новых месторождений, прежде всего в Волго-Уральском регионе, первичная геологоразведка в котором была проведена еще в 30-е годы. Преимуществами месторождений данного региона была их малая геологическая сложность и близость к узлам транспортной инфраструктуры. С середины 50-х годов добыча с месторождений Волго-Уральского региона составляла до 40 % от общего объема нефтедобычи в СССР для того периода времени. Добываемая нефть направлялась для переработки на новые заводы. Интересен факт, что один из крупнейших в мире для того времени Омский нефтеперерабатывающий завод, который был пущен в эксплуатацию в 1955 году, будучи расположен на территории западной Сибири, которая сама является нефтеносным районом, первоначально использовал сырье с месторожде-

ний Волжского региона. Однако волжская нефть уступала по своим свойствам бакинской и северокавказской. Это стимулировало новый виток исследований в нефтехимии и нефтепереработке.

В 30-е годы проводились поиски нефтегазовых месторождений на Елшано-Курдюмской газоносной площади в Саратовской области. В 1941 году в районе поселка Елшанка под Саратовом была пробурена первая газовая скважина с суточной продуктивностью 800 тыс.куб.м газа. В июне 1942 года была пробурена еще одна скважина, которая, как и первая, оказалась высокопродуктивной, что позволило специалистам сделать заключение об открытии месторождения с промышленными запасами природного газа. Эти даты можно считать датами рождения газовой индустрии СССР (России). Получаемый из скважин газ с 1942 года направлялся для снабжения Саратовской ГРЭС, для чего в октябре 1942 года был построен газопровод “Елшанка — Саратов” протяженностью 16 км. До начала добычи природного газа в СССР в качестве горючего технологического газа на производствах использовался светильный газ, получаемый конверсией раскаленного угля водяным паром. Природный газ оказался гораздо более технологичным и менее токсичным, чем светильный газ. В состав которого входит моноксид углерода CO. Следом за ГРЭС и на других предприятиях Саратова началось использование природного газа для получения технологического тепла и для отопления помещений.

В 1943 году около поселка Курдюм в Саратовской области было обнаружено еще одно месторождение с дебитом 1 млн куб.м газа в сутки, а в 1944 году в регионе выявлены значительные запасы газа — 6 млрд куб.м. В конце 1944 года Государственный Комитет Обороны СССР принял решение о строительстве 843-километрового газопровода “Саратов — Москва” для обеспечения газом промышленности и населения столицы.

На строительстве объекта ежесуточно работали до 30 тыс.человек. Десятки заводов машиностроения, приборостроения, тяжелого машиностроения, электропромышленности и других отраслей изготовили почти 9 тыс. наименований различного оборудования и материалов, необходимых газопроводу. Газопровод стал опытным полигоном, на котором отрабатывались новые технологии. Здесь был впервые применен поточно-скоростной метод ведения линейных работ, испытывались строительные механизмы и приспособления для трассовых операций, газосварочные агрегаты, была на практике проверена сварка встык тонкостенных труб высокого давления с толщиной стенки 6,25 мм.

Становление газовой отрасли СССР (России) было отмечено в дальнейшем такими вехами как строительство и ввод в эксплуатацию газотранспортной системы “Средняя Азия — Центр”, которая соединила газовые месторождения Туркмении, Казахстана, Узбекистана с промышленно развитыми районами центральной России, строительство Орен-

бургского газоперерабатывающего завода. В конце 70-х годов строительство газопровода Уренгой-Помар-Ужгород положило начало экспорта российского газа в Западную Европу.

Для научно-технического и инженерного обеспечения газовой отрасли были созданы научно-технические центры. Такие из них как ВНИИгаз, ВНИПИгазопереработка и др. до сих пор являются действующими организациями.

Химическая переработка природного газа в основном связана с процессами получения метанола и азотной кислоты. Соответствующие технологии, включая катализаторы для всех стадий процессов разрабатывались в частности Новомосковске (институт НИАП, ныне входит в компанию "Алвиго").

В 50-е и 60-е годы продолжилась геологоразведка и ввод в эксплуатацию нефтяных месторождений Европейского Севера СССР (республика Коми, Тимано-Печорский бассейн). Началось строительство транспортной системы нефтепроводов. Рост нефетдобычи открыл для СССР возможности увеличения экспорта и упрочнения позиции на международном рынке. В 60-х годах СССР занял второе место среди экспортеров нефти в мире, потеснив Венесуэлу. Уже в то время наметилась непреодоленная до сих пор в современной России негативная тенденция экспорта преимущественно сырой нефти вместо создающих добавленную стоимость продуктов нефтепереработки. Демпинговые цены на нефть, установленные СССР на мировом рынке

привели в конечном счете к конфликту между западными нефтедобывающими компаниями и правительствами стран Ближнего Востока, где были расположены основные месторождения нефти, используемые Западом в то время. Правительства ближневосточных стран для урегулирования этого круга вопросов создали Организации Стран Производителей Нефти (ОПЕК). Арабо-израильский конфликт 1972 года еще более обострил ситуацию. СССР выступил на стороне арабских стран, не в последнюю очередь руководствуясь соображениями удержания доминирующих позиций на нефтяном рынке. Перебои с поставками нефти в западные страны привели к началу добычи нефти Великобританией и Норвегией на шельфе Северного моря.

К этому периоду времени относится также расцвет советской нефтяной науки во всех трех секторах – Upstream, Midstream, Downstream. Как известно, согласно принятой на западе классификации полный производственный цикл добычи и переработки нефти делится на три части - Upstream, Midstream, Downstream. К Upstream относятся процессы нефтедобычи и, говоря более широко, все технологические процессы, связанные с эксплуатацией месторождений. К Midstream относятся процессы подготовки нефти к транспортировке и собственно транспортировки. Технологические процессы Midstream охватывают эксплуатацию трубопроводной транспортной системы для транспортировки нефти. Процессы Downstream связаны с переработкой нефти на НПЗ (нефтеперерабатывающих заводах). Целевой

фокус настоящей работы сконцентрирован на технологиях Upstream.

Геофизические модели месторождений и нефтеносных пластов и газо-гидродинамические модели процессов нефтедобычи получили интенсивное развитие в совместных трудах Московской и Махачхалинской школ математической физики. Можно упомянуть, например, труды Холодова А.С. и Магомедова К.М. с сотрудниками в области численного решения многомерных нелинейных уравнений газодинамики и гидродинамики гиперболического типа. Сибирское отделение АН СССР, прежде всего Институт катализа им.Борескова, становится одним из центров каталитической химии и ее применений в нефтепереработке. Отраслевые научные центры в Ярославле, Стерлитамаке и Нижнем Новгороде начинают интенсивные работы в области катализаторов на основе искусственных цеолитов. Осваивается широкомасштабное промышленное производство самих цеолитов. Общими и специализированными вопросами в отрасли, включая вопросы эксплуатации и диагностики нефтепроводов (и газопроводов) занимаются отраслевые центры в Краснодаре, Саратове, Уфе, ВНИИГАЗ в Москве и т.п. В перспективе планируется распространить геологоразведку и на дно морских шельфов. Такие организации как “Южморгео”, “Союзморгео” и другие приступают к разведке на шельфах Черного, Охотского, Японского морей. Организуются морские экспедиции для

разведки в южно-китайском море в рамках сотрудничества с республикой Вьетнам.

В то время окончательно прояснилась необходимость организации выделенных научно-технических центров (НТЦ) в отрасли, а не только проектных институтов. Именно в НТЦ накапливался интеллектуальный потенциал и формировался банк интеллектуальных ценностей, которые затем могли быть использованы для технологического развития и перевооружения всей отрасли в целом. К недостаткам этого периода можно отнести то, что достижения отраслевой науки оставались зачастую ограниченными не только отраслью, но даже территориальным управлением. Достижения же академической науки, в частности в области нефтепереработки и нефтехимии, не внедрялись, поскольку отраслевые объединения не имели для этого необходимых стимулов.

Тем временем СССР приступил к освоению месторождений Западной Сибири. Высокий уровень добычи, определяемый большими объемами (запасами) отдельных месторождений и относительно небольшие затраты на добычу, явился одним из ключевых факторов наметившегося упадка нефтяной отрасли СССР. На волне успеха были фактически сокращены как затраты на геологоразведку и разработку новых месторождений, так и на совершенствование технологий нефтедобычи и нефтепереработки. Руководствуясь приоритетом максимизации объемов нефтедобычи в краткосрочной, а не в долгосрочной перспективе

советские плановые органы стимулировали производственные объединения добывать как можно больше нефти с уже освоенных месторождений без учета последствий для состояния месторождений. На каждом освоенном месторождении бурилось чрезмерное количество скважин, а в нефтеносный пласт закачивалось очень большое количество воды. В результате к середине 70-х годов прошлого века СССР столкнулся с резким падением отдачи эксплуатируемых скважин в западной Сибири. Правительству СССР удалось приостановить этот негативный процесс путем больших капиталовложений в геологоразведку и ввод в эксплуатацию новых месторождений, однако это дало лишь отсрочку ввиду провала в разработке и внедрении новых технологий по всему технологическому циклу. Как ни странно, но именно в этот период в нефтедобывающей отрасли СССР возникли и получили развитие новые перспективные идеи автоматизации процессов бурения и нефтедобычи, в частности:

- автоматизация процесса бурения, автоматический контроль параметров бурильной установки, потребляемой мощности, сопротивления породы и т.п.;
- прогнозирование и предотвращение аварийных режимов и поломок,оптимизация распределения трудовых и материально-технических ресурсов при проведении ремонтных работ;
- автоматизированная диагностика, контроль потребляемой мощности и предсказание аварийных

режимов для глубинных штанговых насосов инефтеперекачивающих насосов и др.

Следующее падение добычи пришлось на период с 1982 по 1986 год и благодаря политическому кризису и распаду СССР плавно перетекло в упадок нефтяной отрасли 90-х годов. Дезинтеграционные процессы вызвали резкое падение спроса на нефть на внутреннем рынке, помимо этого потребители нефти внутри страны зачастую не могли своевременно оплачивать потребляемое сырье. Возможности экспорта нефти оставались ограниченными, помимо этого финансовые потоки от экспортной продажи проходили через руки финансовых монополистов и около-криминальных структур, так что добывающим компаниям доставалось минимальное количество вырученных от продажи средств. Результатом всех указанных негативных процессов стало дальнейшее падение добычи нефти, которое остановилось лишь в 1997 году.

Этот сложный период в истории нефтяной отрасли России отмечен рядом перспективных разработок в области каталитического крекинга нефти, внедренных в том числе и в производственные процессы на российских НПЗ. Видимо такой интерес бизнеса к технологиям был вызван тем, что НПЗ, будучи расположенными в конце технологической цепочки, в значительной мере ощутили на себе негативное влияние указанных выше деструктивных экономических и политических процессов.

Запасы нефти в России по оценке независимых экспертных организаций все еще довольно значительные. Так остаточные запасы в регионе Западной Сибири оцениваются в более чем 150 миллиардов баррелей (более 20 миллиардов тонн), при этом уровень добычи может быть увеличен в два-три раза по сравнению с текущим. Однако нефтедобыча осложняется тяжелыми геологическими условиями, поскольку месторождения в данном регионе имеют как правило несколько нефтеносных пластов.

Все это потребует вложения средств как в геологоразведку новых и априорное уточнение профилей уже эксплуатируемых месторождений, так и в совершенствование технологий нефтедобычи, включая автоматизацию, а также разработку и использование комплексных цифровых моделей процессов добычи с непосредственной их адаптацией к эксплуатации конкретных месторождений.

Оценка для Европейского Севера России (Тимано-Печорский бассейн) составляет девять миллиардов баррелей (1,25 миллиарда тонн). Указанные запасы относятся преимущественно к трудноизвлекаемым, нефть по своим качественным свойствам относится к тяжелой нефти. Помимо этого развитие нефтедобычи в данном регионе осложняется суровыми климатическими условиями и деградацией транспортной системы времен СССР. Тем не менее потенциал указанного региона, равно как и Волжско-Уральского региона, оценка для которого составляет величину, сравнимую

с оценкой для Тимано-Печерского бассейна, не стоит сбрасывать со счетов.

Оценка остаточных запасов региона Восточной Сибири составляет три миллиарда баррелей (0,45 миллиарда тонн), однако имеющихся данных геологоразведки недостаточно для более точных оценок, в результате чего реальные запасы нефти могут быть в несколько раз больше. Разработка нефтяных месторождений в этом регионе затруднена как геологическими причинами, так и удаленностью месторождений от рынков сбыта и слабый уровень развития транспортной инфраструктуры в регионе.

В последнее время как правительство России, так и западные нефтяные корпорации проявляют интерес к месторождениям на морских шельфах – в Карском море и вблизи острова Сахалин. Разработка этих месторождений сдерживается высокой капиталоемкостью, однако к положительным факторам следует отнести возможность непосредственной транспортировки продуктов добычи морем с помощью танкеров.

История развития нефтяной отрасли России (СССР) в значительной степени укладывается в тенденции, описанные в первых разделах данной главы, с той лишь разницей, что в условиях плановой экономики предиктор рыночной цены на нефть должен быть заменен на прибыль от нефтедобычи (доходы за вычетом затрат). В наиболее благоприятные периоды правительство не уделяло достаточно внимания развитию и внедрению новых технологий, а

в критические периоды ставка зачастую делалась на импорт готовых технологий из-за рубежа. Зависимость от иностранных технологий для нефтяной отрасли России остается критической и сейчас. Так сообщение 30 октября 2014 в Financial Times о выходе иностранных нефтяных компаний из российских проектов повергло российских чиновников и руководителей компаний в уныние и пессимизм. Действительно в 2014 году компания ExxonMobil закрыла 10 совместных предприятий с компанией “Роснефть”. Другие западные компании (как корпорации Shell и Total, так и компании среднего уровня, специализирующиеся на сервисном обслуживании оборудования и инжиниринговом обеспечении) также минимизируют свою деятельность в России. По мнению экспертов указанные тенденции создают дополнительные препятствия в первую очередь для разработки и освоения новых месторождений.

Отчасти ответам на вызовы данной ситуации было посвящено интервью журнала “Нефть и газ – Евразия” с Генеральным директором “Газпромнефть НТЦ” Марсом Магнавиевичем Хасановым.

ООО “Газпромнефть Научно-Технический Центр” образовано 30 октября 2007 года. Предприятие было создано с целью повышения эффективности разработки месторождений и развития минерально-сырьевой базы ПАО “Газпром нефть”. Основными направлениями деятельности Научно-Технического Центра являются проектирование, анализ и мониторинг разработки нефтяных месторождений и геоло-

горазведочных работ, геологическое и гидродинамическое моделирование, технологическая поддержка и оперативный контроль бурения. В сферу ответственности НТЦ входят: создание и ведение корпоративной базы геолого-промысловой информации, управление процессом извлечения нефти из недр с использованием постоянно действующих геолого-технологических моделей, планирование и организация опытно-промышленных работ по внедрению новых технологий в добыче нефти. Также ООО “Газпромнефть НТЦ” выполняет комплекс работ по разработке, экспертизе и защите проектной документации для выполнения лицензионных обязательств, осуществляет планирование, анализ и сопровождение геологоразведочных работ, ведет обучение и переподготовку специалистов ПАО “Газпром нефть”.

По словам господина Хасанова одним из приоритетов ООО “Газпромнефть НТЦ” является взаимодействие с ведущими российскими вузами и привлечение к сотрудничеству молодых специалистов. Так силами ООО “Газпромнефть НТЦ” создан лабораторный центр в Санкт-Петербургском государственном горном университете, а в РГУ нефти и газа им. И.М.Губкина при участии ООО “Газпромнефть НТЦ” открыта кафедра геологии углеводородных систем, совместно организованная ВУЗом и “Газпромнефть НТЦ”. Также открыта специализация “Нефтяной инжиниринг” в Московском физико-техническом институте. Научно-технический центр учредил именные стипендии для успешно осваиваю-

щих программу и участвующих в научных исследованиях аспирантов и магистрантов.

Вместе с тем генеральный директор ООО “Газпромнефть НТЦ” отмечает, что “сегодня на рынке доступны все технологии, можно купить любую из них. Конкурентным преимуществом нефтяной компании в современном мире является не наличие собственных технологий, а умение правильно выбирать и применять эти технологии, все время совершенствовать свой уровень. Успешные компании отличаются от остальных тем, что правильно применяют технологии, используют их потенциал на 100% и вовремя меняют.” Также господин Хасанов отмечает: “Что касается технологий, зачастую НТЦ является проектным офисом для создания конвейера по их внедрению, определению технологических вызовов и их ранжированию, внедрению технологий в производство по проектному принципу.” Таким образом в концептуальном плане позиция Генерального директора ООО “Газпромнефть НТЦ” соответствует подходам советского правительства в 20-е годы предыдущего века – импорт и адаптация технологий и взращивание собственных кадров.

При этом в условиях современной России внедрение передовых зарубежных технологий осложнено разрывом в технологических укладах. В целях обеспечения реальной эффективности внедрение должно носить комплексный, адресный, проектно-ориентированный и проблемно-инициированный характер. И в этом процессе роль научно-тех-

нических центров нефтегазовой отрасли не должна быть недооценена.

Как уже отмечалось ранее, научно-технические центры (НТЦ) отрасли по существу призваны выполнять роль центров компетенции, совмещая ответственность за геологоразведку, оценку запасов, первичную идентификацию параметров вновь осваиваемых месторождений, обустройство и ввод месторождений в эксплуатацию, мониторинг, контроль и управление процессами нефтедобычи на месторождениях с целью максимизации КИН, оптимизацию капитальных затрат и операционных расходов, выбор оборудования и технологий, внедрение новых технологий и формирование и реализацию программ испытаний новых технологий с распространением полученного опыта в других производственных подразделениях компании.

Концентрация интеллектуальных ценностей, функционально-ориентированных знаний, высокопроизводительных вычислительных ресурсов и квалифицированных кадров в рамках НТЦ позволяет обслуживать практически в режиме реального времени большое количество удаленных друг от друга территориально месторождений.

При этом уже внедряются и используются на базе НТЦ и системы управления процессами разработки, бурения и нефтедобычи в режиме реального времени с удаленным доступом к оборудованию и измерительно-сенсорным системам месторождений. Такие системы позволят при разработке и эксплуатации геологически сложных

месторождений оперативно привлекать весь потенциал геологического, гидродинамического и 3D-моделирования, имеющийся в распоряжении специалистов НТЦ, в том числе в виде компьютерных информационно-аналитических инструментов, специализированного прикладного программного обеспечения, баз данных и знаний, совмещенных при этом с экспертными системами с элементами искусственного интеллекта, включая нейросетевые технологии [25] и элементы машинного обучения.

Типовые цепочки возникновения интеллектуальных ценностей в отрасли суть следующие. В первой типовой цепочке инициирующим фактором является проблема, возникающая непосредственно при разработке или эксплуатации конкретного месторождения. Однако проблема должна стать по настоящему типовой, т.е. характерной для нескольких месторождений, или для одного крупного месторождения и оказывать существенное влияние на процесс добычи нефти, чтобы менеджмент компании принял решение о заказе соответствующих исследований в каком-либо НТЦ, после выполнения которого и возникают интеллектуальные ценности, готовые для последующего применения и на иных месторождениях и в иных ситуациях. Другим вариантом является ситуация, когда в ходе заказанных нефтедобывающей компанией типовых проектных и сервисных инжиниринговых работ по разработке, обустройству и эксплуатации конкретного месторождения специалисты НТЦ делают предсказательный прогноз на основе апостери-

орных моделей [26], построенных на основании имеющихся в распоряжении НТЦ данных [27] и знаний, и рекомендуют Заказчику предпринять те или иные превентивные меры и в упреждающем порядке провести необходимые инженерно-технические и геолого-технические мероприятия для обеспечения в дальнейшем устойчивой эксплуатации месторождения с высоким КИН.

Во второй схеме инициатором исследования и разработки и внедрения технологии является руководство компании. Как правило речь здесь идет о вводе в эксплуатацию участков, признанных нерентабельными в рамках применения используемых технологий нефтедобычи, например, малопродуктивных участков с низкой проницаемостью, трещиноватых коллекторов, низко-дебитных скважин, требующих в частности для своего освоения и эксплуатации интеллектуального адаптивного управления процессом добычи.

В оба описанных потока формирования интеллектуальных ценностей вместо заказа исследования и разработки в НТЦ может быть встроена покупка уже существующей технологии на внешнем рынке. Однако и в этом случае роль НТЦ очень важна в плане адаптации технологии, ее внедрения на конкретном месторождении, сбора и анализа первичных данных об использовании технологии и возникающих в связи с этим проблемах.

Также очевидно, что трансфер отдельно взятой технологии может быть затруднен либо вообще невозможен по

причинам кардинальных различий в технологических укладах российской и зарубежной нефтедобывающих отраслей. Поэтому необходимо прежде всего обеспечить целенаправленный поэтапный трансфер оптимальной технологической среды, а для этого прежде всего необходимо внедрение современных концепций организации бизнес-процессов и неуклонное следование основным современным трендам. Только в рамках обновленного концептуального понимания экспликация отдельных технологических процессов не останется обособленными вкраплениями, не растворится со временем, а послужит зародышами новой фазы, вокруг которых начнется эффективная кристаллизация нового технологического уклада. Действительно, большую роль в этом играют процессы стандартизации и инновационного комплексного технологического обучения персонала, его подготовки и переподготовки. Именно, направить неизбежное влияние человеческого фактора при внедрении новых инновационных технологий в нужное русло возможно лишь путем взращивания проектной и корпоративной культуры нового уровня [28].

Помимо этого, внедрение практически любой современной технологии должно сопровождаться изменениями и внедрением новых технологий ИТ-сфере, информационной среде компании, системном и прикладном программном обеспечении. Концептуальные приоритеты здесь хорошо известны и четко определены, это:

- всеобъемлющая диджитализация (цифровизация) нефтяной отрасли от “Цифрового месторождения” до “Цифровой электронной нефтедобывающей компании”;
- применение для управления процессами систем искусственного интеллекта с элементами нейросетевых технологий и алгоритмов машинного обучения;
- широкое внедрение концепций и методов Big Data, включая облачные технологии, аналитические инструменты и специализированное программное обеспечение.

По существу речь идет о внедрении концепции Индустрии 4.0 в нефтедобывающее производство (Uppstream). Для этого необходимо создание условий широкого распространения цифровой культуры, а также обеспечения прямой заинтересованности в успешной цифровой трансформации со стороны сотрудников компании всех уровней и специализаций, но прежде всего высшего руководства компании. Необходимо всесторонне вовлекать в производственный процесс представителей “цифрового поколения” - специалистов по Big Data, нейросетям, кластерному анализу, методам машинного обучения, а также уже действующих сотрудников компании, проявляющих лояльность к динамичной цифровой экосистеме и других сотрудников, которые могут спокойно работать в динамичной цифровой экосистеме. После этого могут быть предприняты меры по скорейшему отмиранию старых “нецифровых” подходов.

Как хорошо известно, интеллектуальное цифровое нефтегазовое месторождение — это система автоматического управления операциями по добыче нефти и газа, предусматривающая непрерывную оптимизацию интегральной модели месторождения и модели управления добычей. Ввиду сложности и нечеткой определенности геологических моделей (как части интегральной модели) построить полностью автоматическое управление нефтедобычей в обозримый период времени представляется невозможным, но при этом возможно использовать данный эталон для формирования целей для программ по снижению человеческого фактора в процессах управления жизненным циклом месторождений.

Интеллектуальное цифровое месторождение — класс систем управления активами (производственными фондами) нефтедобывающих предприятий, построенных на базе формализованной, интегральной модели актива, обрабатываемой автоматизированной системой управления, гарантирующей оптимальное управление на всех уровнях предприятия при контроле целей задаваемых владельцами актива. Термин основан на понятии интеллектуального управления. Аналогом данного термина являются Цифровое нефтяное месторождение (Digital Oil Field), интегрированное управление операциями (Integrated Operation) на месторождении. Частным понятием данного термина является “интеллектуальная скважина”.

Необходимыми условиями существования интеллектуального месторождения является:

- формализованность информационной модели месторождения;
- аппарат управления;
- максимально точные интерфейсы обратной связи (датчики, связь);
- интерфейсы для оптимизации процессов, моделей и критериев.

Для обеспечения целостности управления месторождением, интегральная информационная модель актива должна включать и объединить все аспекты имеющихся знаний об активе, включая такие субмодели, как:

- геологическая модель;
- географическая модель;
- технологическая модель;
- логистическая модель цепочек поставок;
- экономическая модель;
- финансовая модель.

Внедрение интеллектуального цифрового нефтяного месторождения базируется на открытых стандартах ISO 15926, ISA-95, ISA-88 и т.д.

Интеллектуальное цифровое месторождение включает в себя несколько контуров управления, прежде всего:

- Операционный контур, который обеспечивает контроль над эффективностью процессов управления операциями на месторождение (добыча, контроль и

управление режимами работы и состояния оборудования, вспомогательные процессы и т.д.);

– Моделирующий контур, который обеспечивает динамическое развитие модели управления при изменяющихся внешних (контекст) и внутренних (контент) условиях.

Однако процесс диджитализация (цифровизации) наталкивается на ряд препятствий организационно-административного и бихейвориально-психологического характера.

Рассмотрим, например, практику, принятую при разработке месторождений типичную для российской компании ОАО “НК Роснефть”. Представители компании отмечают, что применение современных информационных технологий является одним из ключевых факторов для своевременного выполнения и обеспечения высокой надежности плановой программы бурения скважин в режиме постоянно увеличивающихся объемов бурения как вертикальных скважин так (и тем более) горизонтальных стволов. Для решения этой задачи в компании разрабатывается и внедряется в практику пакет специализированных информационных систем.

Основу плановой программы бурения компании составляют утвержденные проектные решения. Указанные проектные решения претерпевают изменения и подвергаются уточнениям по результатам бурения новых скважин для каждого конкретного месторождения. Для разработки проектных решений специалисты ОАО “НК “Роснефть”

применяют достаточно большой набор специализированных программных продуктов для обеспечения производственных и бизнес-процессов компании, причем это продукты как собственной разработки, так и коммерческие, приобретаемые на рынке. Это в частности пакеты обеспечения как коммерческого, так и собственной разработки, в частности пакеты геологического моделирования Schlumberger Petrel и гидродинамического моделирования Schlumberger Eclipse. Однако для учета результатов реализации программы бурения и иных мероприятий, выполненных на месторождениях компании в течение предыдущего операционного периода с целью уточнения рейтинга эксплуатируемых скважин геологические службы компании используют программный комплекс “Геология и Добыча” (ПК “ГИД”), разработанный российским независимым НТЦ ООО “РН-УфаНИПИнефть”. Далее на этапе реализации программы бурения с целью постоянного мониторинга и внесения необходимых корректировок службы геологического сопровождение в дополнение к ПК “ГиД” используют технологическую информационную систему “Добыча”, разработанную в самой компании ОАО “НК Роснефть” и предоставляющую возможности хранения и оперативного доступа к актуальным технологическим данным.

Перечисленные программные средства и информационные системы могут быть отнесены к “цифровому месторождению”. Однако состоялось ли оно в полной мере в компании ОАО “НК Роснефть”? Думается, что нет. Дей-

ствительно, явно видна разрозненность и разобщенность информационных и программных ресурсов на этапах проектирования и эксплуатации. Информационная связь между двумя этапами осуществляется посредством утвержденных проектных документов, представленных если и не в бумажной, то в устаревшей нединамичной компьютерной форме, явно несоответствующей требованиям цифровизации. Также отсутствует и единая информационная платформа, позволяющая осуществлять обмен данными между специалистами, ответственными за различные стадии и этапы работ. Следует тем не менее отметить, что налицо зародыш комплексного подхода, что может в будущем привести к успеху диджитализации (цифровизации) месторождений компании ОАО “НК Роснефть”. Более того, в компании развиваются и элементы “цифровой электронной нефтяной компании”. С целью снижения капитальных вложений в создание качественного нефтегазодобывающего фонда скважин в ОАО “НК “Роснефть” создается единое информационное пространство для контроля и управления процессами строительства и обустройства скважин на базе корпоративной информационной системы “Контроль и управление строительством скважин” (КиУСС), основными элементами которой являются информационный блок “Удаленный мониторинг бурения” и программные комплексы, обеспечивающие обработку поступающей из УМБ информации. Интегрирующим элементом всей информационной системы является база данных строительства скважин.

Данные (текущие геологические и технологические параметры), регистрируемые в процессе строительства скважин, поступают в указанную базу данных в режиме реального времени.

Помимо этого, в ОАО “НК “Роснефть” в 2007 г. впервые в отечественном нефтегазовом секторе разработано программное обеспечение “Горизонт” для геологического сопровождения бурения (геонавигации) горизонтальных скважин и боковых горизонтальных стволов. Ввиду отсутствия на рынке ранее коммерческих программных продуктов соответствующего профиля ранее приходилось привлекать для выполнения таких работ крупные нефтесервисные компании. Указанное программное обеспечение также без сомнения может быть отнесено к “цифровому (диджитализированному) месторождению”.

Зададимся также вопросом, почему в рассмотренном примере предпочтение перед зарубежными коммерческими программными продуктами отдается продукту, разработанному российским НТЦ. Специалисты НТЦ ООО “РН-УфаНИПИнефть” так отвечают на этот вопрос: существующие коммерческие программные продукты для геологического и гидродинамического моделирования нефтеносных пластов не учитывают динамику так называемых случайных полей, присутствие которых характерно для процессов нефтедобычи в условиях Российской Федерации. Вместо этого используется искусственно осредненное параметрическое представление, конечная точность и эффективность при-

менения которого зависит от специалиста, выполняющего расчет. В то же время программные продукты НТЦ ООО “РН-УфаНИПИнефть” содержат в своем составе модули аналитических расчетов, основанных на теории перколяции и современных моделях случайных полей.

Важность учета фактора случайности подтверждается и другими перспективными работами по учету случайно меняющихся зависимостей между проницаемостью и пористостью пласта. В качестве предикторов для установления таких зависимостей для конкретного месторождения помимо данных геофизического исследования скважин предложено использовать разбиение на зоны с примерно одинаковыми условиями осадконакопления. В рамках используемой модели считается, что статистические закономерности для пористости и проницаемости для каждой из таких зон и для их отдельных частей одинаковы.

Таким образом мы видим, что цифровизация (дижитализация) сталкивается в российских условиях не только с устоявшимися административно-организационными и бихейвиоральными аспектами деятельности в нефтедобывающей отрасли, но и с физикой, причиной чего является существование случайных полей. Это означает, что физика и модели геофизических сред, хотя они и не могут в полной мере быть отнесены к “цифровому” этапу и “эпохе Big Data”, все же имеют шанс на выживание при переходе к индустрии 4.0. Вместе с физикой имеют шанс на выживание,

по крайней мере на начальном этапе, и российские НТЦ в нефтедобывающей отрасли.

По указанной причине анализ деятельности российских НТЦ в нефтяной отрасли по-прежнему представляет большой интерес, в том числе и в разрезе разрабатываемых этими НТЦ инновационно-технологических решений. Чтобы быть проведенным в полной мере, такой анализ требует разработки интегральных критериев эффективности деятельности НТЦ.

Но прежде чем перейти к этому вопросу, рассмотрим некоторые аспекты применения Big Data в нефтедобывающей отрасли на современном этапе.

Следует отметить, что ситуация с Big Data в нефтедобывающей отрасли не является устоявшейся. Вместе с тем Big Data уже здесь, рядом с нами, - их появление неразрывно связано с цифровизацией (диджитализацией) нефтяных месторождений. Так при разработке компанией "ЛУКОЙЛ" иракского месторождения Западная Курна-2 учет всех особенностей бизнес-процессов компании "ЛУКОЙЛ" как оператора месторождения, а также положений нормативных и законодательных актов, действующих на территории республики Ирак, при разработке и внедрении системы автоматизации на базе комплексного MES-решения со сбором, обработкой и хранением технологической информации в реальном времени с помощью PISystem компании OSIsoft, учетом и распределением углеводородов посредством системы Energy Components компании Tieto,

поддержкой работы лаборатории посредством информационно-аналитической платформы STARLIMS и визуализации технологической информации на основе портала XHQ компании Siemens привел к тому, что распределенная система управления (DCS) месторождения генерирует более 20 тыс. сигналов управления и наблюдения за рабочими процессами. Адекватный анализ такого большого количества информации возможен лишь на основе планомерного и целенаправленного использования как идеологической концепции, так и аналитических инструментов Big Data.

Однако необходимо учитывать, что подход Big Data характеризуется экспоненциальным ростом числа измерительных операций и соответствующих им данных. При этом как сами данные, так и алгоритмы их обработки, реализованные в виде информационно-аналитических компьютерных систем с применением сетевых и облачных технологий, наравне с кадровыми ресурсами, технологическими “ноу-хау” и капиталом становятся одним из основных активов индустриальных, в том числе нефтедобывающих компаний. Аналитические инструменты Data Mining являются одновременно ключевыми инструментами достижения конкурентного преимущества на рынке. Важны также и системы сбора, первичной обработки, хранения и обеспечения безопасности информации. Неудивительно, что многие эксперты и аналитики говорят сейчас: “ мы осознали, что вокруг нас терабайты информации, и теперь нам надо понять, что нам с этими терабайтами делать”.

Возможно, что выход состоит отходе от традиционных компьютерных архитектур в сторону нейроморфных вычислительно-аналитических систем, снабженных алгоритмами глубокого (глубинного) машинного обучения. Также необходимым является использование методов вероятностного программирования на основе Баесовского вывода, поскольку зачастую для большой доли сенсоров, обеспечивающих мониторинг цифрового месторождения, характерна ситуация, когда случайный разброс наблюдаемых значений измеряемых величин оказывается сопоставимым с ценой деления средства измерения. Причиной этого является попытка управления сложными технологическими процессами путем контроля все большего числа степеней свободы сложных распределенных систем. Соответствующее увеличение количества сигналов от измерительных устройств с одновременным требованием повышения точности измерений приводит к уменьшению отношения полезный сигнал/шум для большой доли измеряемых величин и параметров. Нередки случаи, когда дальнейшее увеличение точности измерения некой технологической величины (параметра) невозможно по причине достижения физико-технологического предела для данного способа измерений или же слишком затратно. При этом повышение отношения полезный сигнал/шум для такого типа измерений, проводимых в реальных условиях, также либо технологически невозможно, либо очень затратно. В результате увеличение общего количества измеряемых параметров системы или

технологического процесса не ведет, начиная с некоторого предельного значения, к увеличению точности данных мониторинга и качества контроля и управления. Выход видится в разработке и использовании иерархических адаптивных регуляторов, основанных на нечеткой логике, а также нейросетевых систем, в том числе с алгоритмами глубинного многослойного обучения с элементами формирования абстрактных кластеров данных внутри нейросети (технологии глубинного обучения и нейроморфного компьютинга). Говоря условно, нейроморфная вычислительная сеть, снабженная специализированным алгоритмом настройки и обучения, способна в перспективе поглотить поток Big Data, генерируемый сенсорами и датчиками цифрового месторождения и подвергнуть этот поток многокритериальному анализу, отделяя существенные данные от несущественных. В этом же ключе следует рассматривать и перспективы machine to machine communication, когда вооруженные нейропроцессорами устройства вдоль нефтетранспортной магистрали от скважины до узла нефтеподготовки (насосы, клапаны и т.п.) обмениваются между собой данными с целью оптимизации технологических режимов и предиктивного прогнозирования неблагоприятных, внештатных и потенциально аварийных ситуаций. Такой подход полностью соответствует концепциям “Интернета вещей” и “цифровой индустрии 4.0”.

Другой подход к утилизации Big Data, генерируемых цифровым (диджитализированным) месторождением состо-

ит в применении к потоку данных архивации на основе выявленного для данного потока универсального кода, основанного на теории энтропии информации Шеннона и предикторах Лапласа и Кричевского.

Такого рода идеи отсылают нас к работам позднего советского периода (конца 70-х – начала 80-х годов прошлого века) в области автоматизации нефтегазовой отрасли. Тогда были, например, предложены подходы непрерывной диагностики бурового оборудования или глубинных штанговых насосов на основе непрерывного ваттметрирования, а также алгоритмы энерго-эффективного управления оборудованием за счет адаптивного регулирования частоты вращения электрических приводов оборудования различных типов. Разумеется, в современном нефтедобывающем и буровом оборудовании многие эти принципы уже реализованы в различных вариантах. Так, в компании “Газпромнефть” для обеспечения высокоточного бурения горизонтальных стволов с проведением многостадийных ГРП [29] был создан Центр геологического сопровождения бурения [30], специалисты которого в режиме “он-лайн” управляют бурением в условиях удаленного доступа с использованием роторно-управляемых систем. В том числе, подобным образом новейшая “цифровая” технология встречается и объединяется в условиях реального месторождения с электро-механистической технологией. Другим примером может служить разработанный НПК “УралНефтьСервис” гидравлический привод “Герон”, предназначенный для приведения

в действие одного или нескольких глубинных скважинных насосов. Новаторская идея совмещения инвертора с устройством останова привода позволяет отказаться от использования тормозных резисторов и существенно уменьшить время останова гидравлического привода. При этом рекуперированная энергия направляется обратно в электрическую сеть или другим потребителям, что позволяет экономить до 40% от стандартного энергопотребления.

Как же сочетаются приведенные электро-механистические примеры с идеями эффективной утилизации потоков Big Data, например, с использованием универсального кода? Одним из вариантов является пред-процессинг данных непосредственно на устройстве их получения или в рамках минимального технологического контура, в который входит данное устройство (оборудование). При этом преобразованные сигналы должны анализироваться на верхних уровнях АСУ ТП месторождения в рамках соответствующих иерархически организованных моделей работы оборудования и протекания процессов, используемых для мониторинга и контроля процесса бурения или нефтедобычи. При этом сигнал о превышении рабочей нагрузки глубинного штангового или перекачивающего насоса может означать как наступление предаварийной ситуации в работе самого оборудования, так и существенное изменение физико-химических свойств перекачиваемой нефти, а следовательно должен быть проанализирован на более высоких уровнях АСУ ТП и соотнесен в рамках используемой системы моделей с

сигналами от иных устройств, измерительных приборов и оборудования с целью идентификации вызвавших изменения причин и принятия в конечном счете необходимых операционных решений.

Нельзя обойти вниманием и суперкомпьютерные вычислительные технологии и их применение в нефтегазовой отрасли [31—33]. В последние 20 лет и по настоящее время зарубежные страны совместно с нефтяными и сервисными компаниями прилагают значительные усилия по стимулированию научно-исследовательских и прикладных работ по перспективному развитию и эффективному внедрению высокопроизводительных информационно-вычислительных технологий для решения вычислительных задач при поисках, разведке и разработке месторождений углеводородного сырья. В результате этой деятельности зарубежные нефтяные и сервисные компании получили конкурентные преимущества и смогли существенно потеснить российские компании на рынке нефтесервисных услуг, включая производство, продажу и сопровождение программного обеспечения, производство и применение супервычислителей (высокопроизводительных вычислительных систем), что привело к технологической зависимости российских организаций, высокому уровню затрат на сопровождение, отставанию в научно-техническом развитии, росту угрозы информационной безопасности и, в конечном итоге, угрозе полной потери перспективного высокотехнологичного рынка производства, продажи и сопровождения сложной

научно-технической продукции и информационно-вычислительных услуг. В то же время в России, в последние 20 лет, существенно сократились разработка и выпуск отечественных программно-аппаратных комплексов и программного обеспечения, ориентированного на решение задач поиска, разведки и разработки месторождений. Отчетливо проявилось отставание в развитии научных исследований, создании программных продуктов, качестве подготовки специалистов от уровня, достигнутого зарубежными странами. По ряду направлений происходит практически полное замещение отечественного оборудования и технологий импортными продуктами. На отечественном рынке более 80% компьютерных технологий высокого уровня для решения геолого-геофизических задач является импортной продукцией. При использовании зарубежных информационно-вычислительных технологий в области геофизики и геологии неизбежно возникают предпосылки для утечки важной информации о национальных недрах и стратегически важных ресурсах. Такая ситуация при определенных внешнеполитических обстоятельствах может весьма пагубно повлиять на энергетическую безопасность России. Это не означает, что государству следует с помощью организационных или экономических рычагов ограничить доступ на отечественный рынок высокоэффективных компьютерных технологий в этой сфере из-за рубежа. Вместо этого следует стимулировать и инициировать создание и внедрение качественных отечественных программных продуктов,

способных эффективно конкурировать с аналогичными зарубежными разработками. Для этого есть достаточно веские экономические основания. Так, например, распространенные пакеты программ ведущих зарубежных компаний (Petrel, Eclipse, Roxar), используемые для сейсмического [34] и геолого-гидродинамического моделирования [35] обходятся примерно в 4,5 млн. рублей на одно рабочее место плюс 1 млн.рублей ежегодно уплачивается за поддержку. При этом в ряде пакетов программ существует ограничение на использование количества узлов кластерной вычислительной системы. За возможность использования каждого следующего узла надо заплатить около 12 тысяч долларов. Таким образом, для нефтегазовой компании средней величины, в которой задействовано порядка 100 геологических рабочих мести порядка 200 потенциальных пользователей при геофизических работах, и в половину меньше для гидродинамического моделирования, при необходимости полного использования вычислительного кластера в 400 узлов, начальная стоимость затрат составляет не менее 1350 млн. рублей плюс затраты на ежегодную поддержку (экспертная оценка). При этом отечественный программный пакет tNavigator для расчета гидродинамических моделей производства компании “Рок Флоу Динамикс” поставляется без ограничения на количество задействованных вычислительных узлов, стоит в несколько раз дешевле и считает в несколько раз быстрее.

Задачи развития научных исследований, создания и внедрения в практику наиболее эффективных информационных технологий и препятствия, стоящие на этом пути, были отмечены в Энергетической стратегии России до 2030 г. Осознание важности развития суперкомпьютерных технологий и алгоритмов и программных средств высокопроизводительных вычислений для модернизации и инновационного развития различных отраслей народного хозяйства осознано в России на государственном уровне и нашло свое отражение в решениях Комиссии при Президенте Российской Федерации по модернизации и технологическому развитию, решениях Совета Безопасности, отражено в Стратегии развития геологической отрасли до 2030 года и Энергетической стратегии России до 2030 года, Государственной программе "Информационное общество на 2011 – 2020 гг.", Проекте "Создание системы подготовки высококвалифицированных кадров в области суперкомпьютерных технологий и специализированного программного обеспечения" Комиссии Президента РФ по модернизации и технологическому развитию экономики России, выступлениях специалистов, разработчиков суперкомпьютеров, программного обеспечения для нефтегазовой отрасли, представителей нефтегазовых и сервисных компаний сконцентрированных в решении Первой конференции "Суперкомпьютеры в нефтегазовой отрасли".

Обратимся к критериям для оценки эффективности деятельности НТЦ в нефтегазовой сфере Российской Фе-

дерации. Определение и практическое использование таких критериев как раз и является одной из целей настоящей работы. Прежде всего рассмотрим вопрос о том, каких типов НТЦ представлены в России для деятельности в нефтегазовой сфере. Их можно разделить на следующие группы:

- научно-технические центры при крупных российских нефтегазовых компаниях, таких как “Газпром”, “Роснефть”, ЛУКОЙЛ, ТНК-ВР, “Газпром нефть”, “Сургутнефтегаз” и др.;
- государственные научно-технические центры;
- независимые российские научно-технические центры;
- российские научно-технические подразделения иностранных сервисных компаний, например, департамент DCS в “Шлюмберже”.

Согласно сказанному в первых разделах текущей главы, деятельность НТЦ в нефтегазовой отрасли может быть охарактеризована с двух позиций:

- с точки зрения текущей конъюктурной ситуации;
- с точки зрения долгосрочных перспектив.

Оба этих аспекта имеют в своем составе как экономическую, так и технологическую составляющие. Рассмотрим вначале краткосрочный аспект. Как уже отмечалось ранее, экономическая составляющая при оценке эффективности деятельности НТЦ в краткосрочном плане определяется динамикой прибыли компании от продажи нефти, газа и продуктов их переработки. При этом с точки зрения НТЦ

как коммерческой организации его внутренняя структура должна быть оптимизирована под эффективное выполнение заказов, а багаж опыта и компетенций должен соответствовать текущим запросам рынка. Не последнюю роль играют тут и PR-технологии. Действительно, анализ выпусков журнала “Нефтегазовое хозяйство”, проведенный с участием автора настоящей работы, показывает, что в последние годы в статьях появилось большое количество “цифровых” терминов (тематик) таких "расхожих мемов, как “данные”, “метод”, “система”, “исследование”, “датчик”, “стандарт”, “схема”, в противоположность присутствовавшим в более ранних выпусках журнала “Нефтегазовое хозяйство” “трубопровод”, “нефтепровод”, “труба”, “специалист”, “геология”, “месторождение”, “технология”, “территория”, “скважина”, “НПЗ”. Действительно ли в нефтегазовой отрасли России растет интерес к использованию информационных технологий, а применение интеллектуальных методов анализа данных становится все более востребованным в нефтяном секторе экономики? Несомненно, однако статьи в отраслевом журнале, пестрящие подобными “модными” терминами также отражают и конъюктурный всплеск. Превратится ли этот всплеск в долгосрочный технологический тренд, покажет развитие ситуации во времени. Таким образом, экономическая конъюктура определяет текущую рекламную компанию в деятельности НТЦ, реализуемую посредством публикаций в открытых изданиях. Разумеется, далеко не все публикации в отраслевых научных изданиях носят спе-

кулятивный и конъюктурно-рекламный характер. Да и в рекламе посредством научных публикаций вовсе нет ничего плохого. Так регулярно появляющиеся в периодической печати публикации специалистов уже упомянутого ранее НТЦ ООО “РН-УфаНИПИнефть”, в том числе по теме случайных полей, имеют с одной стороны своей целью рекламу разработоок и программных продуктов компании, но с другой стороны служат просветительской цели, и в популярной форме разъясняют широкому кругу специалистов нефтедобывающей отрасли тему влияния случайных полей на процессы нефтедобычи и что с этим можно сделать. Что в свою очередь (в перспективе) опять же расширяет круг потенциальных потребителей продукции НТЦ ООО “РН-УфаНИПИнефть” среди специалистов отрасли.

Перейдем к рассмотрению оценки деятельности НТЦ с точки зрения долгосрочных перспектив. Экономическая составляющая тут может быть охарактеризована с помощью стандартных интегральных показателей экономического анализа деятельности предприятия, таких, например, как интегральные финансовые показатели за операционный период деятельности или удельные финансовые показатели в расчете на одного сотрудника. Интересны результаты исследования компании “Делойт”, которая оценила деятельность 33 НТЦ, действующих в нефтедобывающем секторе российской экономики, в числе которых представлены все перечисленные выше типы, т.е. государственные НТЦ, входящие в состав крупных нефтегазовых компаний, входящие

в состав крупных нефтесервисных зарубежных компаний и независимые (Рис.1.2).

Рисунок 1.2 — Структура предложения на рынке научно-исследовательских работ в 2009 году.Источник: официальные данные компаний, СПАРК, анализ “Делойта”

Итак, мы видим, что несмотря на определенную фактическую ориентацию отрасли на импортные технологии, доля НТЦ зарубежных нефтесервисных компаний невысока, в то время как как доля НТЦ при крупных российских вертикально интегрированных нефтегазовых компаниях превышают суммарную долю для НТЦ всех остальных типов. Это не означает, что отрасль использует отечественные технологии, а означает лишь, что крупные игроки отрасли предпочитают осваивать и адаптировать импортные технологии своими собственными силами. Учитывая текущую структуру рынка разработки технологий в нефтегазовой сфере России, пристальное внимание следует как раз таки уделить деятельности независимых НТЦ, причем с учетом времени их функционирования на рынке. Так можно выявить скрытые технологические тренды и актуальные производственно-технологические запросы в нефтедобывающей отрасли России. Разумеется, если НТЦ достаточно молод, например, функционирует на рынке не более 5 лет, то это вовсе не означает, что такая компания не может демонстрировать высокую эффективность, но при этом все же очевидно, что деятельность достаточно молодых компаний на рынке требует дополнительного анализа с точки зрения оценки эффективности и выявления перспективных трендов.

Перейдем к описанию критериев эффективности деятельности НТЦ в долгосрочном плане, связанных с собственно производимой НТЦ продукцией в виде новых

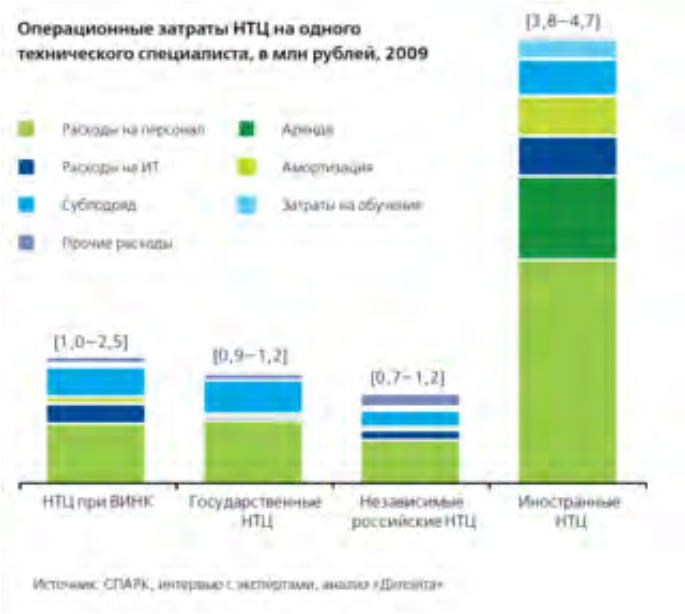


Рисунок 1.3 — Операционные затраты НТЦ на одного технического специалиста, в млн рублей, 2009

технологий. Для этого рассмотрим подход, основанный на анализе цифровых артефактов деятельности НТЦ, в первую очередь различного рода документов в электронном формате, отражающих результаты деятельности НТЦ и доступных для анализа из открытых источников. К такого рода цифровым артефактам могут быть отнесены:

– документы отраслевых электронных библиотек;
– патенты;
– статьи в отраслевых научных изданиях;
– другие материалы и документы, относящиеся к деятельности нефтегазовой отрасли, имеющиеся в

открытом доступе, в первую очередь в сети интернет

Все указанные документы могут быть подвергнуты компьютерному анализу, в первую очередь высоко-перспективным методом тематического моделирования, суть которого состоит в использовании би-кластеризации, то есть одновременной кластеризации слов и документов по их семантической близости. При этом как правило используется скрытое размещение Дирихле, которое хотя и удобно для алгоритмических компьютерных вычислений при проведении тематического моделирования, но не вполне обосновано с лингвистической точки зрения.

Результаты тематического моделирования, проведенного автором настоящей работы, для статей во всех выпусках журнала “Нефтегазовое хозяйство” за период с 2008 по 2016 годы (что упоминалось уже в настоящей главе ранее) показали, что вопреки изначальному предположению о плавной эволюции выявляемых в рамках тематического моделирования тематик от номера к номеру, в различных номерах журнала были зафиксированы принципиально различные тематики. Означает ли это, что указанный журнал в каждом новом выпуске концентрируется на новейших технологических достижениях и не производит отсылку к устаревшим и не использующимся технологическим подходам? И да, и нет. Конъюнктурная составляющая при отборе редакцией журнала публикаций с целью повышения привлекательности издания в широких кругах деятелей

нефтегазовой и смежных областей очевидна. В этом в значительной степени и состоит суть любой, в том числе и узкопрофильной, специализированной журналистики. В то же время о судьбе упомянутых однократно технологий судить по таким однократным публикациям нельзя. Были ли они отвергнуты в практическом применении или были опробованы и показали свою несостоятельность? Или, быть может, прочно вошли за анализируемый период времени длительностью в 8 лет в инструментарий нефтедобывающей отрасли? Такие выводы на основе проанализированной информации сделаны быть не могут. В чем же состоит выход? Он состоит в анализе более широкого круга документов, от патентов до тезисов докладов на отраслевых конференциях. Указанные документы могут быть подвергнуты кластерному анализу с применением различных алгоритмов. При этом если речь идет об использовании классификации с обучающим шаблоном (“с учителем”), то в качестве обучающего шаблона могут быть выбраны экспертные описания основных современных технологических трендов и инновационных тематик в нефтегазовой отрасли. Те же совокупности анализируемых документов, которые не войдут в определенные таким образом классы (кластеры), не должны рассматриваться априори как “шум”, а должны быть подвергнуты дополнительному анализу на предмет того, что они на самом деле представляют собой свидетельства (цифровые артефакты) латентных инновационно-технологических трендов [36].

На базе такой выполненной кластеризации (классификации) документов (цифровых артефактов инновационно-технологического развития нефтегазовой отрасли) могут быть определены многокритериальные интегральные числовые показатели эффективности деятельности конкретных НТЦ, вычисляемые на основе долей распределения цифровых артефактов, произведенных сотрудниками данного НТЦ, по кластерам (классам). Изменение таких распределений во времени могут служить основой для апостериорного прогнозного моделирования эффективности деятельности конкретных НТЦ в будущие периоды времени.

Отдельной темой, хотя и относящейся к технической стороне настоящего исследования, является обеспечение информационного обмена, доступа к документам (цифровым артефактам) и их предобработка, включая приведение в единообразные электронные форматы и стемминг.

Все эти вопросы и будут рассмотрены в последующих главах настоящей работы.

Глава 2. Обзор научной литературы

2.1 Процесс производства знаний

Гуру научного менеджмента Майкл Портер в своей книге «Международная конкуренция: конкурентные преимущества стран» [37] выделяет эффективность научно-исследовательской работы как один из способов конкуренции стран.

Существует множество подходов к трактовке понятия эффективности вообще и в научно-исследовательской работе, в частности. Но важно понимать, эффективность - это не коэффициент [38].

В работе [39] отмечается, что научно-исследовательская работа не может быть однозначно описана и оценена с использованием унифицированных и объективных показателей продуктивности, что свидетельствует о необходимости использования сразу ряда показателей.

В статье Левина [40], автор выступает против исключительно количественной оценки исследователей. Но несмотря ни на что, тенденция формальной оценки научных исследований продолжает распространяться в связи с отсутствием альтернативы отношению к научным текстам, как к основному продукту, производимому исследователями. При этом сам процесс производства научного текста не рассматрива-

ется в деталях в виду его творческой природы и остается “черным ящиком”. Автор данного исследования видит в этом иррациональность и считают своей обязанностью исследовать сам процесс производства научных текстов как основу для дальнейшей оценки продуктивности научной работы и выявления возможностей для повышения продуктивности исследователей.

К основным компонентам процесса научного исследования относят формирования коллаборации исследователей, процесс создания научной статьи и ее публикации. Опубликованная научная статья является одним из воплощений результатов научного исследования. Существует множество методик ведения исследовательской деятельности. Большинство из них использует структурирование научно-исследовательской деятельности на этапы для упрощения ее понимания. Например, в книге [41] выделены следующие семь этапов:

1. Выбор темы научного исследования.
2. Изучение мирового опыта по выбранной теме посредством научных источников.
3. Составление плана научно-исследовательской работы.
4. Накопление материала для проверки обоснованности выдвинутой гипотезы.
5. Обработка данных, построение моделей.
6. Анализ результатов исследования и выводы.

7. Документальное оформление научно-исследовательской работы.

Таким образом, создание научной статьи, как результат научного исследования, может быть представлено в виде формализованного процесса, реализуемого участниками научно-исследовательской группой. Этот процесс принадлежит к категории коллективного социального взаимодействия. И его изучение является нашей задачей в данном исследовании. Поэтому автор поставил задачу рассмотрения именно процесса совместного проведения научно-исследовательской деятельности и написания научной статьи с последующей публикацией [42]. Кроме этого автор данного исследования постарался учесть процессы коллективного мышления и коммуникаций [43], отмеченные в работе [44].

В научной практике исследователи должны делиться результатами своих исследований с коллегами. Публикация статьи в научном журнале является одной из форм коммуникации исследователя с научным сообществом [45; 46]. Помимо публикации статьи, коммуникация может быть осуществлена в виде публикации монографии, тезисов конференций или патентов, а также личных выступлений на конференциях и семинарах. Поэтому научное исследование не может рассматриваться по мнению автора в отрыве от процесса публикации. Таким образом, в коллаборацию коллектива исследователей необходимо включить редакцию научных журналов и комитеты научных конференций.

В самом упрощенном виде редакции и комитеты конференций группируются не по формальным рубрикаторам, типа ГРНТИ, а по определенным ментальным кодам [47], скрытым за описаниями формата и редакционными политиками. Примером такого кода может быть: “Мы принимаем статьи только от членов SPE (Society of Petroleum Engineers)” или “Авторы должны иметь научную степень”. Понятие ментального кода широко применяется при анализе объединения в группы [48; 49]. Ментальный код может состоять из отдельных фрагментов, как молекула ДНК. Важно понимать, что именно на основании совпадения ментального кода производится принятие нового участника в сообщество. Что, в нашем случае, означает принятие редакцией или комитетом конференции научной работы к публикации. Иногда часть ментального кода может быть продекларирована, но это не означает, что существенная его часть, на основании которой и будет приниматься решение не остается внутренним достоянием редакции или программного комитета. В таком случае автор будет испытывать недоумение от того, что ему “немотивированно” отказали, так как существенная часть ментального кода редакции или программного комитета конференции ему не доступна.

Процесс публикации научной статьи так же имеет формальные этапы, в которых, однако, не отражен сетевой процесс работы над результатом:

1. Объявление о дате и теме проводимой конференции;

2. Запрос аннотаций статей (call for papers);
3. Экспертная оценка (peer review);
4. Подготовка текста на двух языках;
5. Создание доклада;
6. Выступление с докладом;
7. Подготовка текста в формате для публикации;

Таким образом, можно говорить о фундаментальном процессе, содержащем логику расширения группы, на основании которого работают как малые группы – соавторы, так и большие группы, включающие в общем случае представителей редакций, организационных комитетов конференций, “гостевых авторов”, переводчиков, экспертов, осуществляющих оценку и т.п. Рассмотрение таких коллабораций необходимо для понимания процесса публикации научной статьи и последующей оценки вклада отдельных участников.

Разделение труда характеризует зрелость производственных процессов. Для рассматриваемого нами процесса написания научных статей это может означать, что создаются специализированные пулы ресурсов для поддержания определенных этапов без персонификации. Например, из прошлого нам известно жаргонное выражение “кооператив по вписыванию формул” применительно к кандидатским диссертациям. Несмотря на маргинальность этого явления, которое публично осуждалось и процветало за счет востребованности в узкой специализации, автор видит в нем ранние предпосылки для разделения труда в процес-

се производства научного исследования и публикации на его основе. В настоящее время в связи с ускорением производства научных исследований появились новые формы разделения труда (и новые требования к результативности научно-исследовательских кадров), которые нуждаются в изучении.

Вопрос коллективного создания знаний и написания научных исследований в частности имеет много аспектов, связанных с этикой исследователя. Должен ли автор, полностью выполнять все этапы работы над исследованием? Если в работе два соавтора, то какое разделение труда не нарушает этических норм исследователя? Какие роли среди соавторов этичны? В общеизвестном “Курсе теоретической физики” Ландау и Лифшица какую роль выполнял Л. Д. Ландау, а какую Е. М. Лифшиц?

После объединения на основе ментального кода происходит развитие отношений в рамках коллабораций в широком (с внешними участниками) и узком (в рамках исследовательской группы) смыслах. Укрепление соавторских отношений в результате написания нескольких работ создает более устойчивые рабочие группы. Существуют примеры продолжающихся на протяжении десятилетий соавторств. С другой стороны, есть примеры, когда, написав одну исследовательскую работу, авторы больше не сотрудничают. В чем причины устойчивых объединений в соавторские коллективы?

Автор считает, что во многих научно-методических источниках основное внимание уделено технологиям написания научной статьи и ее оформлению, но не изучению процесса создания научных статей, поэтому считают данную работу практически полезной для администрирования и планирования НИР в понятиях научного менеджмента по системе Тейлора [50] .

Проблема объективной оценки эффективности НИР находится в центре внимания исследователей уже давно, и это, в первую очередь, связано с вопросами финансирования как бюджетного, так и в рамках грантов. В рамках традиционного подхода выделяются следующие индикаторы оценки эффективности [51]:

- Финансовые
- Кадровые
- Инновационные
- Библиометрические

Собственно в рамках библиометрии учитываются следующие параметры:

- число публикаций в международных журналах характеризует качество статей;
- индикатор цитирования и индекс Хирша показывают степень значимости проводимых исследований и признание научных школ мировым сообществом;
- “публикационная нагрузка” ученых – продуктивность ученых;
- наличие патентов;

– соавторство с зарубежными учеными – показатель международной кооперации".

Как отмечают многие исследователи [52—54], этот набор параметров далек от совершенства, поскольку не даёт полностью объективную картину НИР выбранного учёного или коллектива. Например, индекс Хирша зависит от дисциплины, а также он не спадает, если человек не публиковал новых работ в течение 10 и более лет. Цитатные базы WoS и Scopus, во-первых, неполноценно отражают исследования на русском языке, а во-вторых, разным дисциплинам отводятся в них неравные доли. В данном исследовании проверяется гипотеза, что повышение качества оценки эффективности НИР возможно через учёт дополнительных факторов, о которых будет сказано далее.

2.2 Социальность знаний

Эффективность организации – очень сложный и многогранный концепт [55]. На него оказывают влияние различные факторы. Одним из важных предвестников рыночного успеха научно-исследовательской компании является хорошо развитая коммуникация и кооперация между сотрудниками. Многие теоретические и практические исследования демонстрируют связь между продуктивностью организации и структурой коммуникации её сотрудников,

например, в работах [11; 12]. Исследование социальной структуры организаций и профессиональных сообществ становится одним из главных направлений прикладного анализа социальных сетей. В сфере общественных связей и управления глубоко изучаются модели коммуникаций внутри организаций. Начало этим исследованием положено в 1956 году работе C.H.Cooley “Социальная организация” [13].

Информация о взаимодействиях сотрудников может быть получена различными способами, например, с помощью корпоративных баз данных, общественных опросов и личных отчётов. Однако, данные, полученные такими путями, нужно интерпретировать с некоторыми оговорками, поскольку они не отражают всего механизма профессионального взаимодействия в целостности. Как утверждают Вассерман и Фауст [56], около половины того, что люди сообщают о своих взаимодействиях, по той или иной причине неправильно. Таким образом, людям не очень хорошо удаётся качественно информировать о своих взаимоотношениях, поэтому пути сбора данных должны избегать такой субъективности.

Источником такой информации может быть Google Scholar, arXiv и другие онлайн библиотеки [57]. Рассмотрение открытых научных сообществ так же интересно, как и сужение выборки до одной страны, отрасли и организации.

Одним из более объективных способов анализа человеческих взаимодействий является формальный концептуальный анализ (англ. formal concept analysis, FCA). FCA

представляет собой определенный способ анализа коллекции объектов и их свойств [58]. Идея применить FCA в области анализа социальных сетей уже не нова. В работе [59] он был использован для массового анализа сети. В [60] комбинация формального концептуального анализа и известных методов факторизации была направлена против вычислительной сложности анализа социальных сетей, а также на облегчение визуализации этого анализа. Би-кластеризация и три-кластеризация были применены в [61] для анализа данных, собранных в русской социальной интернет-сети Вконтакте для выделения групп пользователей со схожими интересами, для поиска сообществ пользователей, входящих в состав схожих групп и для выявления интересов пользователей. Формальный концептуальный анализ многократно использовался для анализа социальных сетей, основанного на ссылках [62], для обнаружения криминальных сетей [63]. Другие способы применения FCA можно почерпнуть в работе [64]. Весьма подробный анализ приложений на основе FCA для анализа социальных сетей можно найти в [65—67].

Одним из частных случаев коммуникации является кооперация, которая может в случае научно-исследовательской работы переходить в соавторство при создании научных публикаций.

Публикация научных исследований – главный объект, по которому оценивается эффективность научно-исследовательской работы. Поэтому важно проследить, как проходит

этот процесс, начиная с зарождения исследовательской идеи, проведения эксперимента и заканчивая публикацией работы. Необходимо провести анализ, какие условия способствуют успешной публикации статьи. В рамках данного исследования было изучено соотношение публикаций отдельных учёных и научных коллективов. Было показано, что за последнее десятилетие есть чётко выраженная тенденция ученых объединяться в группы соавторов для публикации статей. Отсюда можно сделать вывод, что одним из факторов, положительно влияющих на опубликование работ, является объединение людей в команды.

В свою очередь, командообразование тоже бывает успешным и неуспешным, оно также поддаётся изучению, в результате которого можно выделить условия успешного командообразования. Задача поиска оптимальных параметров команды соавторов для наиболее продуктивного написания научных статей относится к классу задач оптимизации. Традиционно исследователи обращают внимание на следующие параметры, имеющие значение для продуктивного научного творчества:

- Размер команды
- Ментальные модели сообщества
- Компетенции сотрудников (дополняющие и гомофильные)
- Слабые связи между учёными

В отличие от размера команды, который является явным, а не скрытым признаком, а также легко формали-

зуемым, признак ментальных моделей (англ. mental model) сообщества гораздо труднее поддаётся выявлению и фиксации. Многие исследователи отмечают важность изменения во времени ментальных моделей помимо структуры команды [68; 69]. Понятие ментальной модели является развитием понятий структуры знания [70], схемы знаний [71; 72], и неявной теории [73]. Автор данного исследования трактует понятие ментальной модели как стратегическую согласованность командных компетенций. Например, ментальная модель “Agile geoscience” [74] крупнейшего сообщества ученых-геофизиков [75] основывается на компетенциях “гибкие методики” и “геология”.

Исследователи сходятся в том, что совпадение ментальных моделей участников команды положительно влияет на производительность [76; 77]. Этот факт говорит о связи ментальной модели команды и полного командного кода, которое более подробно раскрыто далее.

Формирование основной системы внутреннего взаимодействия внутри команды согласно исследованию [78] происходит при знакомстве участников по принципу дополняемости. Тем не менее, нельзя полностью отрицать значение гомофильных (совпадающих) компетенций. Во многих работах отмечается динамическая структура гомофилии [79—81], в ходе которой параллельно происходят два процесса. С одной стороны – схожие между собой индивиды формируют социальные связи (социальная селекция). С другой – уже связанные друг с другом люди перенимают

поведение друг друга (социальное влияние). Совокупность этих факторов результирует в гомогенную социальную систему, в которой между индивидами со схожим поведением и характеристиками есть связь, при этом характер связи может быть, как формальным, так и неформальным.

Несмотря на то, что связи между индивидами со схожими характеристиками более вероятны, чем связи между непохожими, уровень схожести также важен. В работе [15] было показано, что социальная схожесть более, чем по одному показателю, приводит к тому, что люди с меньшей вероятностью будут формировать между собой взаимоотношения. Автор объясняет данный эффект тем, что слишком схожие по многим характеристикам люди, как правило, не могут привести что-то новое и конструктивное во взаимные отношения или же в команду.

Для продуктивного сотрудничества необходима не только схожесть интересов, но также и различный профессиональный и жизненный опыт, позволяющий предложить многомерные подходы к ее решению.

2.3 Место текста в научной деятельности

Анализ текста иногда называют *Text Mining*. Суть этого процесса в превращении данных (текста) в высококачественную информацию способную приносить знания.

Важным моментом является то, что при получении знаний человеческие затраты должны быть минимальны.

Полученные из текста знания становятся основой для принятия управленческих решений в организационной среде. Отдельным процессом рассматривается получение текста, иногда называемое созданием корпуса текстов.

Реальный мир находит свое отражение в текстах при помощи авторов, а процесс анализа текста делает обратное: на основе текстов составляет информацию о реальной природе вещей.

Многомодовым подходом к анализу текстов называют процесс учитывания сопутствующей основному тексту информации. Например адрес письма, номер выпуска газеты с новостями или фамилии соавторов научной статьи.

Формально анализ текста производится в следующей последовательности:

1. анализ языка текста
2. анализ содержания текста
3. получение информации об авторе текста
4. вывод определенных переменных, характеризующих природу вещей в тексте

Рассмотрим более подробно методы работы с текстами научных статей.

Обработка текста

Задачи по обработке текста были поставлены в 60-70 годах 20-ого века при обработке натурального языка [82; 83]. Нужно было приводить текст к более удобной для последующего анализа форме. Эту процедуру общепринято называть *нормализацией текста*. Для нормализации текста использовались регулярные выражения (regular expressions), концепцию которых разработал С.К.Клини [84]. Одним из первых, кто использовал регулярные выражения в работе с тестом был К.Томсон [85].

В настоящее время задачи нормализации текста существенно расширились. Нужно не только выделять слова, но и учитывать специальные символы, обозначающие эмоции (Emoji), такие как 8-) [86], выделять хештеги [87], выделять гиперссылки [88] и обрабатывать цитирования [89].

Задача лексического анализа состоит в разделение текста на части - предложения, слова, буквы. Иногда лексический анализ называют токенизацией от английского слова *tokenizing* [90].

Другая задача нормализации текста состоит в определении слов с единой основой и называется лемматизацией. Основа слова не обязательно совпадает с морфологическим корнем слова. Лемматизация для русского языка отличается от лемматизации для английского [91—93]. Поэтому для английского языка используют процедуру лемматизации на

основе частотных алгоритмов [94; 95], так же называемую стемминг от английского слова *stemming*. Но для других языков лемматизация использует еще более сложные алгоритмы. Например, есть стемминг для Древнегреческого языка [96].

Таким образом нормализация текста состоит из трех этапов:

1. Выделения слов из текста
2. Приведения слов к более общим формам
3. Выделении предложений

Для автоматизации задач нормализации текста используют библиотеки на языке программирования Python. Например, библиотеку NLTK [97], содержащую огромное количество различных алгоритмов обработки текста.

Модели текста

Модели, которые присваивают вероятности словам в последовательностях слов называются вероятностными моделями текста [98]. Математически это определение можно записать в виде уравнения. Допустим у нас есть вероятность последовательности из n слов $P(w_1, \ldots, w_n)$, такая, что вероятность третьего слова $P(w_3)$ равна $P(w_3|w_1,w_2)$. Тогда следующее выражение определяет вероятностную модель текста.

$$P(w) = P(w_1, w_2, \ldots, w_n) = \prod_{i^n} P(w_i | w_1, w_2, \ldots, w_{i-1}) \quad (2.1)$$

Так как вычисление $P(w)$ представляет сложность O^n, то современные исследования текста используют представление $P(w)$, как однородной Цепи Маркова и строят приближенные модели [99]:

1. Униграмная модель $P(w_1, w_2, \ldots, w_n) \approx \prod_i P(w_i)$
2. Биграмная модель $P(w_i | w_1, w_2, \ldots, w_{i-1}) \approx \prod_i P(w_i | w_{i-1})$

Можно так же рассматривать n-грамные модели для большего охвата контекста, как в работах [100; 101]. Применительно к задачам распознавания речи это сделано в работах исследователей из ИБМ [102—104].

Tomas Mikolov в работах [105; 106] показывает, что и эти упрощенные модели обладают слишком большой вычислительной сложностью, поэтому его лаборатория разработала векторное представление слов не с помощью априорных распределений как в исследовании [107], а на основе встраиваемых (embedded) векторов.

Из определения (2.1) следует и способ для проверки качества вероятностной модели текста. Для этого пользуются метрикой *Perplexity*:

$$\mathcal{P} \approx \sqrt{\frac{1}{P(w)}}$$

В работах [108; 109] показано, что между метрикой *Perplexity* и относительной энтропией на одно слово $H(W)$ существует следующая зависимость:

$$H(W) = \log_2 Perplexity(W)$$

Клод Шеннон [110] оценил энтропию английского языка как 0.6 - 1.3 бита на букву, прося людей предсказать следующую букву слова. В работе [111] дана оценка нижней границы энтропии английского текста как 1.25, а в работе [112] с использованием триграм модели слов приведена оценка 1.75 бит на слово.

Но есть и другие подходы к оценке качества моделей текста. Например в работах [113; 114] использует метрику основанную не на энтропии, а на попарном сравнении (pairwise ranking approach).

Помимо подхода, развиваемого T.Mikolov, существуют и другие способы векторного представления слов. Следует отметить работу исследователей и Университета Стандфорд, названную GloVe [115]. Векторное представление слов Glove требует существенно меньше вичислений, так как использует только частоты употребления слов, а не вероятности.

Классификация текста

Самым распространенным примером потребности в классификации текстов, наверное, является задача отнесения писем к категории нежелательных почтовых рассылок, называемых спамом. И базовой линией для этой задачи является классификатор на основе алгоритма Наивный Байес (НБ).

Наивная Байесовская текстовая классификация была предложена М.Мароном в работе [116] для присвоения категории принадлежности текста доклада определенному журналу. Его модель представила большинство особенностей, используемых и в настоящее время для задач классификации текстов.

Байесовские методы [117] были также применены к задачам классификации текстов по авторству в пионерской работе Ф.Мостеллера и Д.Уоллеса [118]. Наивный Байес был впервые применен для обнаружения спама в работе [119].

В работах [120—122] было показано что, использование бинарных признаков с мультиномиальным распределением дает лучшие результаты, чем счетчики слов.

Бинарный Байес с Мультиномиальным распределением часто путают с другим вариантом наивных Байесовских алгоритмов, которые также используют двоичное представление того, встречается ли слово в документе: Многовариантный Наивный Байес (МНБ) с использованием

распределения Бернулли. Вариант НБ с рапределением Бернулли оценивает вероятность того, что слово не входит в документ.

В исследовании [123] показано, что МНБ не всегда хорошо обобщается на новые тексты.

Задача определения эмоциональности текста относится к задачам классификации и успешно решается с помощью алгоритмов НБ. Существует ряд хороших обзоров применения анализа эмоциональности текстов среди которых работы [124—126].

Стоит также отметить хороший обзор различных текстовых классификаторов, сделанный К.Маннингом с соавторами [127].

В настоящее время векторное представление частей текста (embedding) приобрело большую популярность. Широко используются методы Word2Vec [105], GloVe [115], StarSparse [128], Fasttext [129], Sent2vec [130]. Поэтому стоить упомянуть и рациональный взгляд на преимущества от использования векторного представления

Основной демонстрацией преимущества от использования векторного представления слов стала формула: king - men = queen. Смысл этой формулы в том, что векторные представления слов (king, men, queen) можно подвергать арифметическим операциям.

Но отнюдь не все слова так складываются. Некоторые очевидные для человека аналогии в векторном представлении не являются близкими векторами [131]. Поиск

смыслов в векторных представлениях предпринят в работе [132—134].

Алгоритм AdaGram предложен в работе [135] для поиска векторных представлений для неоднозначных слов. Исследование уточнения пониманий векторных представлений в зависимости от контекста предпринято в работах [136; 137].

Существенные преимущества в классификации текстов были получены от применения рекуррентных нейронных сетей. Из всей массы работ в этом направлении необходимо отметить многочисленные исследования текстовых моделей на основе нейронных сетей выполненные сотрудниками лаборатории естественного языка при Стенфордском университете [138—142].

Сравнение коллекций научных статей

Согласно информации в отчёте о развитии искусственного интеллекта[1] количество статей по этому научному направлению выросло до 140 тысяч в год (рис. 2.1).

Вряд ли существует учёный, который поддерживает своё понимает развития этой научной области, читая все статьи. Для анализа таких массивов информации используются автоматизированные средства обработки текстов.

[1] https://syncedreview.com/tag/2018-china-ai-development-report/

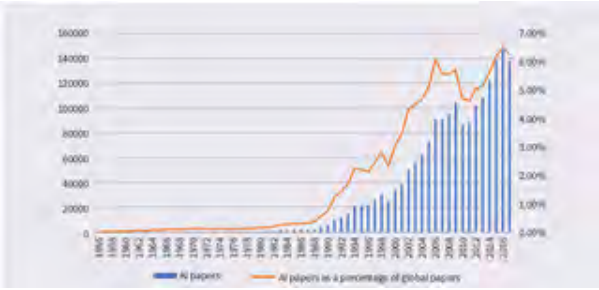


Рисунок 2.1 — Динамика роста количества статей по направлению Искусственный интеллект

Одним из наиболее популярных методов анализа текстов является тематической моделирование.

Сравнительный анализ контента относится к области информационного поиска и ранжирования. Для поиска определённых фраз в корпусе документов производится ранжирование контента для выделения документа с наивысшим рангом.

Задачи сравнения «документ–документ» нашли широкое применение в области поиска плагиата [143] и [144].

В настоящем исследовании авторами ставится задача сравнения корпусов документов: "корпус-корпус". Статистические подходы к решению этой задачи были сделаны в работах [145] и [146]. Современное развитие статистических подходов для этнической стратификации корпусов текстов изучено в работе [147]. В исследовании [148] сделано статистическое профилирование корпусов текстов для их сравнения.

Сравнение корпусов с помощью извлечения тематик и введение метрик было изучено в работах [149] и [150].

Другим возможным подходом может быть построение корреляционной матрицы "документ-документ"и сведение к задаче сравнения документов. В этом случае размерность корреляционной матрицы будет $N_1 * N_2$, где N_i – это количество документов. В корпусы могут входить разное количество документов, поэтому в общем случае $N_1 \neq N_2$.

Двуязычный корпус текстов

Официальным языком научных статей де-факто является английский, но достаточно большой объем научных работ изначально публикуется на родном языке учёного, а только потом переводится на английский в более полном и углублённом виде. Таким образом, можно говорить о существовании двуязычного корпуса документов. В настоящее время в компьютерной лингвистике растёт интерес к двуязычным корпусам текстов для создания моделей машинного перевод. Например, в работе [151] используется Proceedings of the Canadian Parliament на Английском и Французском, а в работах [152; 153] используются субтитры от кинофильмов на нескольких языках. Среди особенностей таких параллельных корпусов в работе [154] выделяют принадлежность языка определённой

области и неполное соответствие смыслов переводов – особенности перевода. Анализу особенностей параллельного перевода посвящено исследование [155] , в котором отмечают несколько уровней параллельности: на уровне слов, фраз, предложений и рассуждений. Перечисленные выше исследования [151—153; 155] фокусируются на выделении пар совпадающих предложений и большое внимание уделяют выстраиванию соответствий между словами. Такой подход является важным этапом в решении задачи статистического машинного перевода (СМП), которая была сформулирована уже более 50 лет назад в работе [156]. Важными вехами в истории создания подходов, основанных на СМП, являются создание моделей I и II центром исследований IBM Watson в 70-х годах прошлого века. В настоящее время существенным выглядит прогресс, достигнутый с помощью архитектуры *Encoder-Decoder* в работах [157; 158].

Концепция архитектуры *Encoder-Decoder* подразумевает, что одновременное обучение производится для двух искусственных нейронных сетей, а в центре находится скрытое представление переводимого текста ν. Схема архитектуры *Encoder-Decoder*, представленная на рисунке 2.2, адаптирована автором по материалам исследования [159].

Авторы данного исследования поставили цель рассмотреть структуру скрытого представления ν, применив в качестве *Encoder* и *Decoder* не искусственные нейронные сети [160], а аппарат тематического моделирования текста с последовательной регуляризацией [161].

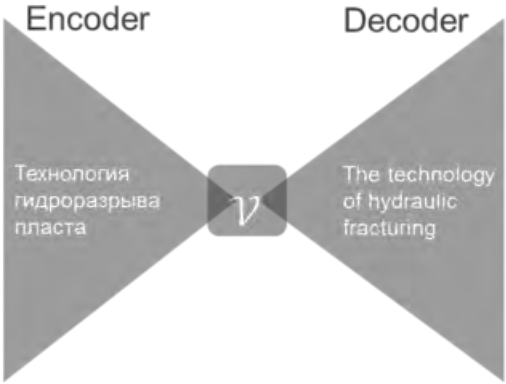


Рисунок 2.2 — Модель Encoder-Decoder для задачи СМП.

Перевод научной статьи с русского на английский может быть выполнен по-разному. Некоторые авторы используют для перевода средства СМП, а некоторые как бы пишут статью заново на английском. Результаты таких подходов отличаются. В случае использования СМП иногда даже говорят, что переводная статья написана на «русском английском» языке. Настолько это заметно для человека. На наш взгляд представляет интерес исследовать методы определения способов перевода статей.

Исследовательская гипотеза автора состоит в следующем:

Для двуязычного корпуса научных статей возможно автоматически выделить пары научных статей, перевод которых сделан с помощью статистического машинного перевода.

2.4 Анализ социальных сетей

В книге [16] отмечается, что базисом для анализа социальных сетей является теория социометрии, основанная J.L.Moreno [17]. Социометрия изучает взаимоположения социальных атомов в группах. Социограммой по Морено является графическое отображение социального выбора членов социальной группы. Социальным выбором может быть выбор лидера, дружба, выполнение совместных задач, и др. Социограма представляет граф, состоящий из вершин и ребер.

В книге [162] M.Tsvetovat с соавторами ставит вопрос: Кто из участников организации, представленной в виде графа, важнее? Таким образом проводя логическую связь между графами и организационной теорией, что отмечено в работе [163].

Хорошим примером измерения инновационности организации с помощью анализа социальных связей служат исследования [164] .

Направление углубленного изучения графов (Social media mining) деятельных сообществ развито в работах [165; 166]. Таким образом, переводя различные метрики графов в свойства для задач классификации составляющих графы вершин и ребер.

Например, в работе [167] метрики графа соавторства использованы для предсказания новых соавторств. А в

работе [168] предсказание соавторств использовано для повышения эффективности научной организации.

Для решения задач предсказания вершин и ребер графов необходимо рассмотреть, как создаются графы. Известны модели Small World [169], Preferential Attachment Model [170], и другие модели случайных графов, состоящих из подграфов.

В работе [171] показано, что выделение подграфов дает возможность выявления социальных подгрупп, объединенных общей темой. Определение таких сообществ не представляется возможным без использования матетамического аппарата теории графов [172].

Одной из методик для выявлении сообществ является построения векторного пространства узлов (node embedding) [173]. Так же как и с векторным пространством частей текста, описанным в разделе 2.3, построение векторного пространства узлов позволяет вводить авторам исследования [174] новые свойства графов на основе близости узлов в векторном пространстве.

Граф соавторства является частным случаем социальной сети. Одним из первых исследований графа соавторства является работа [18], сделанная в 1973 году. С этого времени исследования научной деятельности при помощи графов соавторства не прекращались и обрели статус проверенного инструмента анализа [175]. Например, в недавнем исследовании [19] предпринята попытка предсказания будущих научных исследований на основе графа соавторства, а в ра-

боте [20] построен глобальный граф соавторства на основе Google Scholar, который содержит более 400 тысяч вершин. Оба исследования проведены в 2017 году.

Построение графа соавторства выполняется таким образом, что если два автора сделали совместную научно-исследовательскую работу, то каждый из авторов считается вершиной графа, а факт соавторства ребром графа. Будем называть такой способ создания графа соавторств традиционным. В результате такого, традиционного подхода получают граф, изображенный на рисунке (Рис. 2.3).

Рисунок 2.3 — Пример простого графа соавторства.

Глава 3. Объект и методы исследования

3.1 Цифровые экосистемы

Появление цифровых экосистем является результатом естественного развития научного сотрудничества и информационных технологий. Целью цифровых экосистем является повышение эффективности связей между внутренними и внешними агентами для поддержания бизнеса. В литературе есть два широких определения концепции цифровых экосистем. Первое исходит из структурной и функциональной перспективы, которая видит цифровую экосистему как открытую сетевую среду для эффективного взаимодействия. Второе, напротив, рассматривает цифровую экосистему как открытый кластер слабо связанных компонент, в котором каждый агент является проактивным для собственной выгоды (Рис. 3.1).

Понятие “цифровые артефакты” вошло в обиход вместе с понятием цифровой экосистемы. В широком смысле цифровые артефакты являются синонимами любых информационных результатов действий цифровой экосистемы. По своей информационной природе цифровые артефакты могут сохраняться или разрушаться. Как при сохранении, так и при разрушении происходит видоизменение изначального цифрового артефакта. В исторической перспективе цифро-

Рисунок 3.1 — Экосистема научного инжиниринга

вые артефакты могут быть изучены, как и любые другие продукты деятельности человека.

Цифровые экосистемы можно рассматривать на макроуровне (страна, отрасль) [176] и на микроуровне (корпорация, группа компаний, отдельное предприятие, департамент) [177—179]. Цифровые артефакты также могут существовать на разных уровнях.

Одним из примеров цифровых артефактов на микроуровне являются системы распространения знаний в нефтегазовых компаниях. Система распространения знаний (СРЗ) – это инструмент, помогающий координировать процессы управления и обмена знаниями в области разведки и добычи нефти внутри группы компаний “Газпром нефть” для решения технологических и производственных задач при принятии решений.

Она предназначена для настройки процессов сбора, обработки и распространения знаний с целью извлечения максимальной выгоды от внедряемых в компании практик и технологий. СРЗ реализована в виде информационной системы с несколькими модулями, помогающими получить необходимую информацию по разным аспектам работы на месторождении.

В СРЗ систематизировано представлена информация о лучших практиках, применяемых в “Газпром нефти” в области разведки и добычи. Система позволяет пользователю проводить сравнительный анализ [180] и подбор оптимальных технических решений в соответствии с необходимыми ему критериями. В ней также хранятся данные обо всех проведенных внутри компании испытаниях нового оборудования, что позволяет наиболее эффективно внедрять новое оборудование и технологии на любом месторождении внутри компании.

Наибольший вклад в развитие СРЗ вносят эксперты Научно-Технического Центра “Газпром нефти”. Они формируют крупную структурированную базу знаний по различным областям геологии, геологоразведки и добычи, к которой имеют доступ все сотрудники “Газпром нефти”. СРЗ – это один из инструментов создания инновационного климата внутри компании, необходимого для развития новых, более эффективных технологий разведки и добычи нефти.

3.2 Модели и моделирование социотехнических объектов

Современная парадигма научного исследования заключается в том, что реальные объекты заменяются на их упрощённые представления, абстракции, выбираемые так, чтобы в них отражалась сущность явления, те свойства исходных объектов, которые являются существенными для решения проблемы, которая была поставлена. Объект, который построен вследствие упрощения, называют моделью.

Модели могут классифицироваться по разным признакам: динамические и статические, дискретные и непрерывные, стохастические и детерминированные, имитационные и аналитические.

Статистические модели оперируют характеристиками и объектами, которые не меняются во времени. В динамических моделях изменение параметров модели во времени является существенным. Статистические модели имеют дело с уравнениями балансового типа, установившимися процессами, с предельными характеристиками. Моделирование динамических систем заключается в имитации правил перехода системы из определённого состояния в другое в течение времени.

Модели, у которых состояние изменяется непрерывно во времени, называются непрерывными. Модели, у которых переходы от одного состояния системы в другое происхо-

дят мгновенно, в дискретные моменты времени, называются дискретными.

Стохастические модели, в отличие от детерминированных, учитывают вероятностный характер параметров системы.

При аналитическом моделировании процессы функционирования исследуемой системы отражаются как алгебраические, интегральные, дифференциальные уравнения и логические соотношения, и в определённых случаях анализ таких соотношений может выполняться посредством аналитических преобразований.

При имитационном моделировании структура моделируемой системы – её связи и подсистемы – непосредственным образом представлена структурой модели, а процесс функционирования подсистем в виде уравнений и правил, которые связывают переменные, имитируются на компьютере.

Компьютерные системы предсказательного моделирования (которые также называются системами поддержки принятия инженерных решений) с компьютерными системами проектирования давно применяются в целях автоматизации трудовой деятельности инженера-проектировщика и повышения качества решений, которые принимаются. Но до начала XXI века в предсказательном моделировании применялись исключительно математические модели, основанные на принципах физики, описывающие физические явления и процессы, которые

происходят при функционировании объекта, сложными дифференциальными уравнениями в частных производных с граничными условиями. В содержательных ситуациях для таких уравнений неизвестны ни теоремы о единственности и существовании решения, ни характер зависимости решения от граничных условий и параметров. Численные методы решения этих уравнений обладают значительной вычислительной трудоёмкостью и самих расчётов, и подготовки исходных данных и расчётных сеток. В силу этого существенно сокращаются возможности использования этих моделей в проектировании сложных объектов, в особенности на этапе концептуального или предварительного проектирования, когда рассматривается значительное число разных вариантов решений и особенно высока цена решения, которое выбрано неправильно. Важная часть предсказательного моделирования – это имитационное моделирование, которое используется для исследования сложных информационно-телекоммуникационных систем.

3.2.1 Апостериорный и априорный подходы к исследованию

Рассматривая возможности апостериорного и априорного подхода к исследованию, автор склоняется к тому, чтобы отдать первенство экспериментальному изучению

данного явления, а потом выяснить, какие из теорий смогут составить базу для дальнейшего углубления в изучение феномена соавторства [181].

Современные возможности прямого имитационного моделирования стали настолько удобны для произведения вычислительных экспериментов, что для начального подхода к изучению сложных социальных явлений достаточно быстро могут дать исследователю существенное понимание их природы. Формализм математической модели в данном случае не абстрагирует в мир греческих букв, а приближает к пониманию родовых особенностей исследуемого объекта.

Модель стремиться дать то описание системы, для которого она создается. Но отметим, что создание модели для полного описания социальной системы не является корректной постановкой задачи. Полная модель социальной системы будет настолько же сложна, насколько и сама социальная система. Сформулируем следующее определение модели социальной системы (3.2.1):

Определение 1 Модель $\mathbb{M}_\Omega$ социальной системы Ω может быть использована для получения характеристик $\mathbb{RE}$ с некоторой точностью δ.

Таким образом, целью модели является получение ответов на некоторую совокупность вопросов. Эти вопросы неявно присутствуют (подразумеваются) в процессе анализа, и, следовательно, они руководят созданием модели и направляют его. Это означает, что сама модель должна

будет дать ответы на эти вопросы с заданной степенью точности. Если модель отвечает не на все вопросы или ее ответы недостаточно точны, то говорят, что модель не достигла своей цели.

Агентное моделирование предполагает имитацию поведения системы путем настройки поведения отдельных агентов. На основании результатов поведения отдельных индивидов складывается комплексная картина взаимодействий. Метод агентного моделирования используется в дополнение к методу системной динамики, в рамках которой моделируется поведение всей системы в целом.

Программные алгоритмы агентного моделирования разработаны в нескольких информационных системах, в частности Anylogic и NetLogo. Для решения практических задач эти информационные системы используется в социальных науках, в том числе в экономике и социологии. Важной задачей агентного моделирования является включение информации о взаимодействиях агентов между собой, так как в некоторых социальных системах именно комплексная структура взаимодействий индивидуальных агентов и приводит к более сложным макро-состояниям. Агентное моделирование используется для изучения динамики социальных сетей и взаимного влияния экзогенных и структурных характеристик друг на друга.

Модель для исследования взаимодействия агентов в процессе создания научных статей была реализована автором в программной среде агентного моделирования

AnyLogic, базирующемся на языке Java. В среде AnyLogic для каждого из агентов прописываются определенные правила поведения – эвристики, индивидуальные стратегии. После того, как для каждого из агентов прописываются все правила поведения, запускается серия симуляций. Программные среды для агентного моделирования используются для предсказания коллективного поведения, массовых мероприятий, учебного процесса и многих других социальных процессов.

Для моделирования процессов в данном исследовании использовался метод имитационного моделирования на основе внутренних состояний и действий. Основное преимущество данного подхода состоит в возможности проведения компьютерного эксперимента для понимания поведения системы в целом с помощью настройки графов состояний и действий децентрализованных индивидуальных агентов. Таким образом, в результате была получена база данных поведения агентов для исследования процессов.

В рамках изложенной выше методологии были сформулированы следующие вопросы для исследований:

1. В какой степени научная статья отражает проведенную НИР? Можно ли судить о качестве НИР по опубликованным научным исследованиям?
2. Каковы социальные механизмы объединения исследователей для проведения НИР? Какие виды компетенций и в какой степени влияют на такое объединение?

3. Как зависит время проведения НИР от количества участвующих исследователей? Существуют ли естественные ограничения на количество и состав исследовательских групп и на чем они основаны?
4. Каковы эвристические алгоритмы поведения исследователей по отношению к издательствам и программным комитетам конференций? Существуют ли базовые стратегии поведения? Если возможность идентификации и имитации базовых стратегий?
5. Применимы ли подходы time management (“управление временем”) к НИР? Насколько эффективно рассмотрение научно-исследовательской деятельности как проектной деятельности?
6. Какова модель зрелости научно-исследовательской организации в части проведения НИР? В какой степени возможно определение степени зрелости научно-исследовательской организации на основе анализа публикуемых ею научных статей?
7. Какова структура процессов, составляющих научно-исследовательскую деятельность? Насколько применим процессный подход к изучению научно-исследовательской деятельности? Есть показатели научно-исследовательской деятельности, отражающие характерную структуру составляющих ее процессов?

3.2.2 Теория имитационного моделирования

Имитационное моделирование является методом исследования, при котором изучаемую систему заменяют на модель, которая с достаточной точностью описывает реальную систему, с которой проводятся эксперименты для получения информации об этой системе.

Цель имитационного моделирования заключается в получении приближенных знаний об определенном параметре объекта, без осуществления непосредственного измерения его значений. Такая необходимость возникает, когда измерение невозможно или оно стоит дороже, чем проведение имитации. В то же время для изучения такого параметра есть возможность пользоваться иными известными параметрами объекта и моделью его конструкции. Допуская, что модель конструкции довольно точно описывает объект, автор предполагает, что статистические распределения значений параметра моделирующего объекта, полученные в ходе имитации, будут в определённой степени совпадать с распределением значений параметра реального объекта.

Направления применения имитационного моделирования:

- Агентное моделирование
- Системная динамика
- Дискретно-событийное моделирование

– Динамические системы

Далее рассмотрим более подробно Системную динамику.

3.2.3 Системная динамика

Данный подход разработал и предложил Джей Форрестер в конце 1950х как исследование обратных информационных связей в промышленной деятельности для того, чтобы показать, как организационная структура, усиления (в политиках) и задержки (в действиях и принятии решений) взаимодействуют, оказывая влияние на успешность предприятия.

Приложения системной динамики также включают урбанистические, социальные, экологические системы. Процессы, которые происходят в реальности, представляются в Системной Динамике в терминах накопителей (англ.stocks, к примеру, материальных объектов, людей, знаний, денег), потоков между этими накопителями (flows) и информации, определяющей величину таких потоков. Системная Динамика абстрагирована от определенных событий и объектов и предполагает агрегатный взгляд на процессы. Она концентрируется на политиках, управляющих этими процессами. Моделируя в стиле Системной Динамики, вы представляете структуру и поведение системы в качестве множества взаи-

модействующих отрицательных и положительных обратных связей и задержек.

3.2.4 Принципы построения моделей

Социально-экономическая система может описываться многими системно-динамическими моделями. Выбор факторов, которые подлежат включению в модель, обусловливается вопросами, на которые должен даваться ответ. Но в общем случае база построения модели не может ограничиваться какой-либо узкой научной дисциплиной. Стоит включать в модель экономические, организационные, правовые, технические, трудовые, психологические, исторические и денежные факторы. Все они должны найти своё место при определении взаимодействия элементов системы. Всякий фактор может оказывать решающее влияние на поведение системы.

Обычно в самые важные модели, которые отвечают запросам управления, включаются от 30 до 3000 переменных. Нижний предел близок к минимуму, отражающему основные типы поведения системы, которые интересуют тех, кто принимает решения. Верхний предел ограничен нашими возможностями восприятия системы и всех её взаимосвязей.

Особое внимание стоит уделять таким аспектам исследуемой системы, как:

– временные зависимости,
– прямая и обратная связи,
– искажение информации.

При построении модели её переменные должны соответствовать переменным моделируемой системы и измеряться в тех же единицах. Например, потоки товаров должны измеряться не денежными, а натуральными единицами. Потоки денежных средств рассматриваются отдельно. Денежные и товарные показатели связываются ценами. Товары нельзя представлять, как соответствующие денежные суммы, иначе не будет учитываться значение цен и факт того, что движение денег не является синхронным движению товаров. Заказы на товары не являются товарами, отгруженные товары не являются равнозначными счетам к оплате, а последние не равнозначны денежным средствам.

В модели экономической системы стоит использовать фактические цены, а не индексированные или приведенные. Фактические цены и их колебания ведут к важным психологическим последствиям, к примеру, при установлении величины зарплаты.

Системно-динамическая модель не обязательно должна являться устойчивой. Среди имеющихся социально-экономических систем определенные неустойчивы в математическом понимании. Они не стремятся к равновесному состоянию даже в случае отсутствия внешних возмущений. Социальные системы в высшей степени

нелинейны и большую часть времени противодействуют ограничениям, которые связаны с недостатком рабочей силы, преодолением инфляции, сокращением денежных ресурсов, спадом деловой активности, недостатком средств производства.

3.2.5 Этапы компьютерного имитационного моделирования.

Помимо принципов, существуют и общие этапы компьютерного имитационного моделирования. Как правило, оно включает в себя следующие этапы:

- Понимание системы: понимание того, что происходит в системе, которая подлежит анализу: какой является ее структура, какие процессы протекают в ней.
- Формулировка цели моделирования системы: список задач, которые предполагается решить посредством будущей модели. Список выходных и входных параметров модели, список исходных данных, критерии завершенности будущего исследования.
- Разработка концептуальной структуры модели: структура модели, состав существенных процессов, которые подлежат отображению в модели, зафиксированный уровень абстракции для каждой

подсистемы модели (список допущений), описание управляющей логики для подсистем.

- Реализация модели в среде моделирования: реализованные подсистемы, их поведение, их параметры, реализованная логика связи подсистем.
- Реализация анимационного представления модели: анимационное представление модели, пользовательский интерфейс.
- Проверка корректности реализации модели: убеждение в том, что модель корректно отражает процессы реальной системы, которые требуется анализировать.
- Калибровка модели: фиксация значений параметров, коэффициентов уравнений и распределений случайных величин, которые отражают ситуации, для анализа которых будет использоваться модель.
- Планирование и осуществление компьютерного эксперимента: результаты моделирования – таблицы, графики и т.п., которые отвечают на поставленные вопросы.

Кроме этапов моделирования необходимо рассмотреть принципы сбора данных, необходимых для эксперимента. Об этом будет сказано в следующем подразделе.

3.2.6 Методы сбора данных

Имитационное моделирование является статистическим экспериментом. Его результаты должны базироваться на соответствующих статистических проверках: доверительные интервалы и методы проверки гипотез. Для выполнения данной задачи получаемые наблюдения и имитационный эксперимент должны соответствовать таким требованиям:

1. **Наблюдения имеют стационарные распределения, то есть распределения не меняются при проведении эксперимента.** Результаты наблюдений над моделью находятся в зависимости от длительности периода имитации. Начальный период неустойчивого поведения модели, как правило, называют переходным. Когда результаты имитационного эксперимента стабилизируются, система переходит в установившийся режим. Чем длиннее продолжительность прогона модели, тем выше шанс достижения установившегося состояния.
2. **Наблюдения подчинены нормальному распределению.** Данное требование можно выполнить, если привлечь центральную предельную теорему, которая утверждает, что распределение средней выборки является асимптотически нормальным, вне зависимости от распределения

генеральной совокупности, из которой взята выборка.

3. **Наблюдения независимы.** Природа имитационного эксперимента не гарантирует независимости между последовательными наблюдениями над моделью. Но использование выборочных средних для представления отдельных наблюдений дает возможность смягчить проблему, которая связана с отсутствием независимости.

Существует три самых общих метода сбора информации в ходе имитационного моделирования:

– Метод подынтервалов. Если рассматривается имитация n наблюдений продолжительностью T, в соответствии с этим методом обрезается информация, относящаяся к переходному процессу и остаток результатов имитации делится на n равных частей. Среднее значение искомой величины внутри каждого подынтервала используется как единственное наблюдение. Преимущество этого метода заключается в том, что влияние нестационарных условий снижается. Недостаток состоит в том, что последовательные группы с общей границей являются коррелированными, что ведет к невыполнению предположения о независимости.

– Метод повторения. В этом методе каждое наблюдение представляется независимым прогоном модели, в котором переходный период не учиты-

вается. Вычисление средних величин выборки для каждой группы осуществляется точно таким же образом, как в методе подынтервалов. В этом случае стандартная формула для дисперсии является применимой, поскольку группы между собой не коррелированы. Преимущество этого метода заключается в том, что каждый имитационный прогон модели определяется своей последовательностью случайных чисел из интервала, благодаря чему действительно обеспечивается статистическая независимость получаемых наблюдений. Недостаток заключается в том, что все наблюдения могут быть под сильным влиянием начальных переходных условий.

– Метод циклов. Данный метод может рассматриваться в качестве расширенного варианта метода подынтервалов. В этом методе постарались уменьшить влияние автокорреляции посредством выбора групп так, чтобы обеспечить для каждой из них одинаковые первоначальные условия. В качестве переменной может рассматриваться длина очереди, тогда каждая группа должна начинаться в тот момент, когда длина очереди равна нулю. В отличии от метода подынтервалов, в методе циклов длины интервалов каждой группы могут быть разными. К недостаткам метода можно отнести меньшее, в

сравнении с методом подынтервалов, количество получаемых наблюдений при заданной длине прогона.

Имитационное моделирование является довольно гибким инструментом исследования, который можно эффективно использовать при анализе сложных систем. Его недостаток состоит в том, что любой результат, полученный при имитационном моделировании, подвержен экспериментальным ошибкам и должен проверяться статистическими тестами. Задача получения наблюдений методом имитационного моделирования, которые являются одновременно репрезентативными и независимыми в стационарных условиях, достаточно трудна. Использование специальных методик сбора данных позволяет смягчить эти трудности.

3.2.7 Применение имитационно-прогностических моделей в исторических исследованиях.

Теоретико-методологические проблемы применения имитационно-прогностических моделей пока еще не разработаны. Существуют разные мнения о возможном использовании имитационно-прогностических моделей в истории, но есть большой интерес к их применению. Существующий опыт их практического построения дает

возможность выделения трех типов задач, которые могут быть решены на их основе:

- моделирование альтернативных, то есть субъективно и объективно возможных, но нереализованных на практике исторических ситуаций с тем, чтобы охарактеризовать реальный ход развития более глубоко;
- построение моделей контр фактических (реально не существующих) исторических ситуаций, которые конструируются историком в целях использования данных моделей в качестве эталона оценки реальной исторической действительности;
- имитация исторических явлений и процессов, для обычной характеристики и отражательно-измерительного моделирования которых нет необходимых конкретно-исторических данных.

В последние годы достигнуты существенные успехи в области создания моделей социальной истории. Имеющиеся к настоящему времени модели можно условно разделить на три группы:

- модели-концепции, основанные на выявлении и анализе общих исторических закономерностей, представлении их в виде когнитивных схем, описывающих логические связи между различными факторами, влияющими на исторические процессы (Дж.Голдстайн). Эти модели имеют высокую сте-

пень обобщения, но обладают не математическим, а чисто логическим, концептуальным характером;

– частные математические модели имитационного типа, которые посвящены описанию конкретных исторических событий явлений (Д.Медоуз, Дж.Форрестер). В таких моделях основное внимание уделяется тщательному учету описанию факторов процессов, которые влияют на рассматриваемые явления. Применимость этих моделей обычно ограничена довольно узким пространственно-временным интервалом; они “привязаны” к определенному историческому событию, их нельзя экстраполировать на длительные периоды времени;
– математические модели, которые являются промежуточными между двумя указанными типами. Данными моделями описывается определенный класс социальных процессов без претензии на детальное описание особенностей для каждого конкретно-исторического случая. Их задача заключается в выявлении базовых закономерностей, которые характеризуют протекание процессов рассматриваемого вида. В соответствии с этим данные математические модели называются базовыми.

3.3 Модель процесса публикаций научно-практических статей

Все исследователи сталкивались с тем, что опубликовать результаты исследования почти так же сложно, как и выполнить само исследование. Рассмотрим процесс публикации результатов исследований детально и проанализируем возможности его ускорения и упрощения для авторов. Отправной точкой для нашего анализа будем считать готовый текст, описывающий с точки зрения исследователей результат их научно-исследовательской работы. Традиционно этот текст называют рукописью.

В современном мире скорость публикации рукописей является критическим фактором для роста научного вклада страны в международную науку. Публикация статей требует от исследователей широкого спектра навыков по администрированию и коммуникациям, которые не всегда являются характерными для ученых. Необходимость приобретения этих навыков отдельными учеными-авторами создает риски потери фокусировки на исследовательских вопросах и отнимает у ученых время, которое можно с пользой потратить на науку. С другой стороны, привлекая в соавторы людей, например, для перевода статьи на английский или лоббирования командировки на конференцию, авторы размывают исследовательский профиль организации и создают так называемых “гостевых” соавторов.

Исторически задача учёного состоит в том, чтобы сделать результат исследования доступным для наиболее широкого круга заинтересованных лиц; в этом состоит суть процесса публикации результатов исследования. Основная цель данной диссертации – это исследовать процесс публикации рукописи, понять узкие места, определить возможности по их устранению и предложить усовершенствования. Ниже изображен логический каркас (research framework) исследования в виде схемы (Рис. 3.3):

Рисунок 3.2 — Логический каркас исследования.

Как видно из рисунка, логический каркас исследования включает в себя организационную среду (соавторы и их рукопись), процесс публикации, издателей, показатели продуктивности и результаты публикации. В следующем подразделе каждый из компонентов логического каркаса будет рассмотрен подробнее.

3.3.1 Рукопись

Как упоминалось ранее, рукопись по форме - это текст. Методологически рукописи разделяют на следующие основные виды:

– Монография
– Научная статья
– Тезисы доклада

Научная статья – это произведение небольшого объема, обычно от 5 до 20 страниц. По содержанию научные статьи разделяются на три типа:

– Научно–теоретические статьи
– Научно–практические статьи
– Научно–методические статьи

Научно–практические статьи посвящены научным экспериментам и реальному опыту. Далее будет рассмотрен именно этот тип рукописей.

3.3.2 Соавторы

Большинство исследований выполняют научно-исследовательские коллективы, а не авторы-одиночки. Как следствие, рукописи тоже пишутся в результате коллективного труда. Согласно исследованию [181], в нефтегазовой отрасли

распределение количества соавторов имеет вид, представленный на рисунке (Рис. 3.3.2).

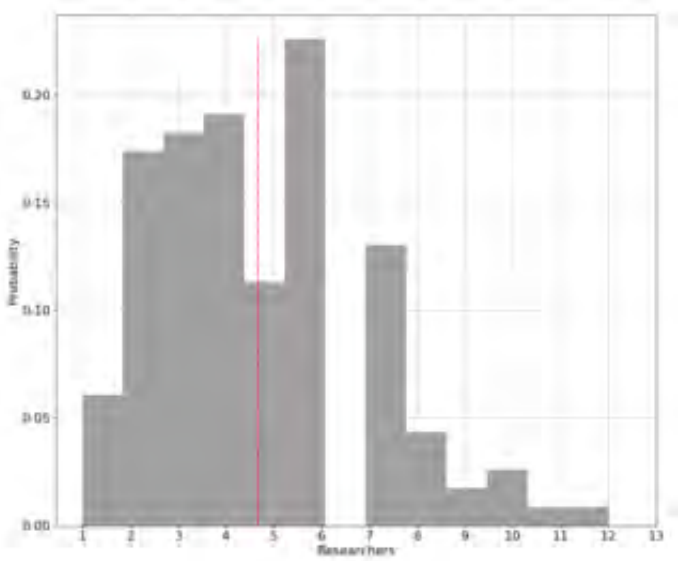


Рисунок 3.3 — Распределение количества соавторов научных статей в нефтегазовой отрасли. Красной линей обозначено среднее значение: 4.67. Стандартное отклонение распределения равно 2.28.

3.3.3 Организационная среда

Научные исследования проводятся сотрудниками научно-исследовательских подразделений. В нефтегазовой отрасли такие подразделений могут принадлежать

профильным институтам, научно-техническим центрам, сервисным организациям и другим участникам экосистемы [182]. Таким образом, соавторы работают в организационной среде. Организационная среда во многом определяет коммуникации между соавторами, что важно на нашего исследования.

3.3.4 Процесс публикации

Процесс публикации состоит из двух типов действий:

– Взаимодействие соавторов с издателем
– Взаимодействие соавторов между собой

Объектом обоих действий является рукопись и сопутствующие дополнительные материалы: анкеты, презентации, письма, рецензии. Основная задача взаимодействий с издателем состоит в удовлетворении условий для публикации статьи в данном издании. Обычно требования к авторам обозначены на веб сайтах издателей и могут отличаться. Взаимодействия соавторов между собой в процессе публикации включают следующие действия:

– Формирование списка возможных издателей
– Изучение специфики требуемых издателями тематик
– Определение временных ограничений на подачу рукописи

- Составление плана доработок рукописи под требования издателей
- Сбор сопутствующих документов по требованиям издателей
- Подготовка презентации для доклада (обязательно для публикации в материалах конференций)
- Выступление с докладом (командировка)
- Подтверждение авторства в сообществах ученых и индексах.

3.3.5 Издатели

Наиболее значимыми считаются издатели, рекомендованные к публикации Высшей аттестационной комиссией (ВАК). В “Перечне рецензируемых научных изданий” ВАК по состоянию на 20.09.2017 содержится 2172 издания. Выберем издателей по одной, наиболее близкой для нефтегазовой отрасли специальности — 25.00 “Науки о земле”. Таких изданий оказалось 147 штук. Далее автор разработал следующий алгоритм для сбора данных:

1. Осуществляем поиск веб сайта по названию издания
2. На сайте издания ищем раздел “Для авторов”
3. Собираем перечень требований к рукописи и заносим в таблицу.

4. Переходим к следующему издателю.

Воспользуемся результатами исследования [183], чтобы выбрать издателей с наибольшей публикационной активностью и импакт-фактором по международным реферативным базам. Список содержит 16 журналов. Все 16 журналов имеют общие правила для авторов, разработанные издательством МАИК “Наука/Интерпериодика”. Каждое издание имеет свой допустимый объём публикаций, обусловленный количеством статей в одном выпуске и количеством выпусков в год. Чем больше рукописей поступает к издателю, тем выше конкуренция за право быть опубликованным.

3.3.6 Результаты публикации

Результатом публикации является определенный вклад в науку. Задача максимизации доступности результатов исследования в эпоху Интернет может быть решена с помощью использования интернет ресурсов. Приведем лишь некоторые способы для увеличения аудитории:

- Международные реферативные базы (Scopus, WoS),
- Электронные библиотеки (например, eLibrary.ru),
- Присвоение научной статье идентификатора цифрового объекта (DOI),

- Привязка научной статьи к автору в онлайн сообществах ученых (например, ResearchGate),
- Публикация статьи в открытых библиотеках (например, arxiv.org),
- Привязка статьи к идентификационному номеру ученого (например, ORCID, SPIN),
- Индексы цитирования (например, РИНЦ).

Индекс цитирования, например, Российский индекс научного цитирования (РИНЦ), является одним из распространенных наукометрических показателей и применяется для формальной оценки в научных кругах. Альтернативами индексу цитирования являются экспертная оценка и оценка по импакт-фактору научных журналов. Углубленные методы библиометрического анализа дают возможность рассмотреть вклад автора с различных точек зрения. Большое внимание в частности уделяется анализу публикаций с помощью графов соавторства, рассмотренных далее в разделе 3.9. Пример графа соавторств приведен на рисунке (Рис. 3.3.6).

Графы соавторств позволяют визуально выделить наиболее значимых ученых по данной тематике. Например, на рисунке (Рис. 3.3.6) мы можем видеть такой кластер.

Принцип построения графов соавторств заключается в отнесении количества публикаций по выбранному ключевому слову к вершинам – авторам, а фактов соавторства к ребрам графа. Такой принцип построения графа позволяет

Рисунок 3.4 — Граф соавторств для ключевого слова *Нефтяные оторочки*

Рисунок 3.5 — Фрагмент графа соавторства по ключевому слову Нефтяные оторочки. Изображены только узлы, относящийся к профессору Rahim Masoudi.

анализировать его с помощью методов Анализа социальных сетей (Social network analysis, SNA).

3.3.7 Показатели продуктивности публикаций

Показатели продуктивности процесса публикаций должны давать интегральную характеристику процессу и позволять проводить сравнения различных реализаций процесса. Наиболее полезными представляется показатели, приведенные в таблице (Таб. 1):

Таблица 1 — Показатели продуктивности процесса публикаций

Название показателя продуктивности	**Описание**
Эффективность публикаций	Отношение количества опубликованных рукописей к общему количеству написанных рукописей
Доля опубликованных рукописей на одного автора	Отношение количества опубликованных рукописей к числу авторов
Доля отвергнутых издателями рукописей на одного автора	Отношение количества отвергнутых рукописей к числу авторов

Предполагается, что процесс более продуктивен, когда *Эффективность публикаций* стремится к единице, *Доля опубликованных рукописей на одного автора повышается*, а *Доля отвергнутых издателями рукописей на одного автора* стремится к нулю. Стратегии управления процессом публикации через показатели продуктивности приведены в таблице (Таб. 2).

Таблица 2 — Стратегии управления продуктивностью процесса публикаций через показатели продуктивности.

Название показателя продуктивности	**Максимальная продуктивность**	**Минимальная продуктивность**
Эффективность публикаций	Стремится к единице	Стремится к нулю
Доля опубликованных рукописей на одного автора	Увеличивается	Уменьшается
Доля отвергнутых издателями рукописей на одного автора	Стремится к нулю	Увеличивается

Отметим, что приведенные показатели продуктивности никак не характеризуют качество самой научной статьи. В данном исследовании автор не ставит задачу оценки качества научной работы.

На основании вышеизложенных методических принципов может быть построена модель процесса с помощью системной динамики.

3.4 Экосистема научного издательства

Каждое научное издание озабочено своей контентной аутентичностью. Но субъективное понимание тематической ниши главным редактором издания может существенно отличаться от реального положения дел.

Журнал может дублировать другое издание по научным направлениям, а может не соответствовать собственным, декларированным направлениям. Журналы, издаваемые десятилетиями, проходят естественные фазы жизненного цикла: старение и деградацию. В результате приоритеты научного издания могут мигрировать до диаметрально противоположных. Например, вместо академических статей по математической логике большинство статей окажется посвящено популяризации олимпиад по математике для студентов.

Помимо тематического наполнения журнала, важно и сообщество авторов. С течением времени критерии попадания в номер для авторов начинают отличаться. Возникает определённое неравенство. Например, зрелому автору пуб-

ликоваться первый раз в данном журнале сложнее, чем аспиранту известного профессора.

Здоровая экосистема журнала обладает характерным поведением и находится в согласии с динамикой научного контента. Проблема своевременного обнаружения вышеперечисленных симптомов дисгармонизации жизненно важна для редакционных коллегий журнала. Гомеостаз читаемости и цитируемости, позволяет установить стабильные метрики развития журнала. Но такая статистика не всегда бывает доступна в отличии от текстов статей.

Авторы разработали методический подход для сравнительного анализа контента журналов с целью выявления степени соответствия контента декларируемым научным направлениям. Одной из составляющей этой методики, названной автором T4C, является возможность обнаружения аномального поведения экосистемы журнала, выявления коммерциализации научного контента, гостевых соавторов, само цитирования, "выкупленных выпусков"и других отклонений.

Предложенный автором подход основан на методологии тематического моделирования [24] и теории графов.

Научная новизна этой методики состоит в новых метриках парного модального тематического моделирования с последовательной регуляризацией [184].

В результате проведённого в настоящей работе эксперимента было показано, что разработанная автором ме-

тодика T4C, продуктивно работает на коллекциях слабо структурированных научных текстов на английском языке.

3.5 Теория суррогатного моделирования

Суррогатная модель лежит в основе нового направления моделирования в инженерии. Она является математическим методом составления модели, базирующейся на результатах испытаний и/или вычислительных экспериментов, проведенных с разнообразными объектами одного рассматриваемого класса. В некоторых случаях суррогатное моделирование является единственным способом решения инженерно-технической задачи.

Задачей суррогатного моделирования является оптимизация исходной сложной функции [185] таким образом, чтобы максимально уменьшить область расчета и свести его к минимуму. Для упрощения многих инженерных задач строится суррогатная модель целевой функции, которая впоследствии заменяет саму целевую функцию.

Концепция создания суррогатных моделей состоит из следующих этапов:

– Характеристика объекта Z, определяющая свойства объекта в некоторых условиях, может быть описана в виде функциональной зависимости $Z = \Phi(X,Y)$,

где переменная X описывает сам объект, а переменная Y задает условия функционирования.

– Функция Ф является неизвестной, и для ее вычисления проводятся вычислительные эксперименты.
– Имеется некоторое количество измерений $\Xi = \{X_i,Y_i,Z_i = \Phi_i(X_i,Y_i), i \in \mathbb{R}\}$, где значение $Z_i = \Phi_i(X_i,Y_i)$ характеристики Z получено методом M_i для объекта, имеющего описания X_i, в условиях функционирования Y_i.
– По известному множеству Ξ с помощью тех или иных математических методов анализа и обработки данных строится функция $\Phi^s(X,Y)$, значение которой принимаются в качестве приближенного значения характеристики Z для объекта с описанием X в условиях функционирования Y.

  Если все значения в множестве Ξ получены при помощи одной и той же модели M и $\Phi^s(X,Y) = \Phi^m(X,Y)$, то построенная функция Φ^s может рассматриваться как "заменитель" (суррогат) функции Φ^m.

Суррогатное моделирование успешно применяется в таких областях как электротехника, нефтяное дело, водное хозяйство, военное дело, машиностроение и химическая отрасль.

Незаменимо применение суррогатных моделей и в строительстве для оптимизации аэродинамической формы для выявления оптимальной формы уникальных граждан-

ских сооружений, таких как высотные здания и большепролетные мосты, которые окружены турбулентным потоком.

Следующие задачи нефтегазовой отрасли так же могут быть решены с использованием суррогатных моделей:

- Построение суррогатных модели резервуара (surrogate reservoir model);
- Оптимизации расположения скважин;
- Анализ неопределенности прогноза добычи нефти;
- Автоадаптация модели резервуара по данным;
- Задача оптимизации при суперэлементном моделировании разработки нефтяных месторождений.

В ряде вычислительных экспериментов для решения задач нефтегазовой отрасли используется вышеописанный мета-алгоритм. Например, сначала в гидродинамическом симуляторе производятся расчеты значения функции для определённых узловых значений параметров X_i на основании физических законов движения жидкостей в пористой среде M_i, а потом, заданная таким числовым образом функция Φ используется для получения значений функции Y_i либо на более детализированном множестве значений параметров, либо для значений параметров, выходящих за рамки узловых значений X_i.

Одна из основных причин возникновения описанного выше мета-алгоритма построение суррогатной модели заключаются в ограничениях на скорость гидродинамической моделирования. В будущем, когда в любое время любой специалист организации сможет варьировать значения па-

раметров в широком диапазоне и в режиме близком к реальному времени получать искомые значения функции потребность в суррогатных моделях скорее всего отпадет. А пока моделирование производится на дорогостоящих высокопроизводительных кластерах, специалистами за времена, измеряемые в часах, а иногда и днях для одного набора параметров существует потребность в прозорливой подготовке данных которые могут понадобиться в дальнейшем. Так как, потребность в изменении параметров может возникать по несколько раз в день и у самых разных специалистов различных подразделений организации, то применение суррогатного моделирования является насущной необходимостью. Получаемая суррогатная модель Φ^s, иногда ее называют прокси-моделью [186; 187], превосходит изначальную модель Φ^m по вычислительной силе во много раз, то есть, не требует большого объема вычислительных ресурсов и работает в режиме близком к реальному времени.

3.6 Непараметрические модели

Для того, чтобы понять, что такое непараметрические модели рассмотрим параметрические модели. Параметрической модель p для значений y, зависящих от переменных X и параметров θ будет иметь вид $p\left(y|,\theta\right)$. Нахождение пара-

метров θ с помощью методов максимизации апостериорной вероятности $p(\theta|y,X) \longrightarrow \max_\theta$.

Для поиска оптимальных параметров математической модели используются методы оптимизации.

Численные методы оптимизации:

- Градиентные и безградиентные,
- Одно и много критериальные,
- Робастные (для задач оптимизации в условиях неопределенности),
- Основанные на суррогатных моделях (surrogate – based),

Рассмотрим методы Байесовской оптимизации, которые наиболее часто применяются в суррогатном и имитационном моделировании. При этом данные и модель являются "черным ящиком".

Пусть дана функция $f(x)$ и нам нужно найти x при котором она достигает максимума $f(x) \longrightarrow \max_x$. Добавим условие при котором расчет каждого значения $f(x)$ это ресурсоемкая задача. Такое условие встречается в следующих случаях:

- x - это географические координаты скважины, а $f(x)$ – это количество нефти, которое можно добыть, пробурив скважину с координатами x. В таком случае одно значение $f(x)$ стоит миллионы рублей;
- x - это гиперпараметры искусственной нейронной сети глубокого обучения, $f(x)$ – это целевая метрика

точности предсказания. В этом случае одно значение $f(x)$ будет занимать месяцы работы;
- x - это лекарство, а $f(x)$ - эффективность лекарства против болезни. В таком случае одно значение $f(x)$ будет стоить жизни одного подопытного животного.

Таким образом, постановка задачи состоит в том, чтобы оптимизировать целевую функцию за минимальное количество попыток. При этом использование суррогатных моделей целевой функции позволяет сделать каждый шаг оптимизации менее ресурсоемким. Введем функцию ценности обнаружения $\mu(x)$, которая характеризует выгоду полученную от оптимизации $f(x)$ при использовании суррогатной модели $\hat{f}$. Функция ценности обнаружение является количественной оценочной функцией для минимализации количества попыток. Расмотрим следующие $\mu(x)$:

- Maximum probability of improvement (MPI): $\mu(x) = P(\hat{f}(x) \geqslant f^* + \varepsilon = \Phi\left(\frac{\mathbb{E}\hat{f}(x) - f^* - \varepsilon}{Var[\hat{f}(x)]}\right)$, где f^* - текущее лучшее значение.
- Upper confidence bound (UCB): $\mu(x) = \mathbb{E}\hat{f}(x) + \eta Var[\hat{f}(x)]$
- Expected improvement (EI): $\mu(x) = \mathbb{E}\max(f(x) - f^*,0) = Var[\hat(f(x)] \cdot [z\Phi(z) + \varphi(z)]$, где $z = \frac{\mathbb{E}\hat{f}(x) - m(x)}{Var[\hat{f}(x)]}$
-

3.7 Байесовские методы для определения параметров НТЦ

Рассмотрим результаты деятельности НТЦ как наблюдения x, тогда в самом общем смысле в качестве задачи поставим найти распределение случайной величины θ, приводящей к имеющимся наблюдениям.

Согласно теореме Байеса имеем выражение 3.1.

$$p(\theta|x) = \frac{p(x|\theta)\ p(\theta)}{\sum_i p(x|\theta_i)\ p(\theta_i)} \tag{3.1}$$

Для вычисления апостериорного распределения $p(\theta|x)$ на основании функции правдоподобия $p(x|\theta)$, априорного распределения с плотностью вероятности $p(\theta_i)$ и полной вероятностью $p(x) = \sum_i p(x|\theta_i)\, p(\theta_i)$.

Вычисление полной вероятности $p(x)$ является сложной задачай, поэтому воспользуемся принципом максимизации апостериорной вероятности $p(\theta|x)$. Найдем такие параметры θ_{MAP} при которых выражение $p(\theta|x)$ максимально. Принцип максимализации апостериорной вероятности (Maximum A Posteriori, MAP) можно записать в виде:

$$\begin{aligned} \theta_{MAP} &= \arg\max_{\theta} p(\theta|x) \\ &= \arg\max_{\theta} \frac{p(x|\theta)\ p(\theta)}{p(x)} \end{aligned}$$

Так как полная вероятность $p(x)$ не зависит от θ, то можем убрать знаменатель и получить формулировку для

оптимизационной проблемы в виде:

$$\theta_{MAP} = \arg\max_{\theta} p\left(x|\theta\right) p\left(\theta\right) \tag{3.2}$$

Уравнение 3.2 не содержит $p\left(x\right)$ и может быть решено с помощью численных методов. Но данный подход страдает от следующих проблем:

- Нет инвариантности относительно параметров распределения θ_{MAP};
- θ_{MAP} не применима в качестве априорного распределения;
- Нет возможности оценить байесовский достоверный интервал (credible interval).

Рассмотрим частный случай θ_{MAP} когда вероятности всех θ одинаковы - юниформное распределение. Тогда задача поиска θ сводится к поиску максимального значения для $p\left(x|\theta\right)$. Такой подход называют методом оценки максимального правдоподобия (maximum likelihood estimation, MLE). Запишем выражение оптимизационной задачи для метода оценки максимального правдоподобия.

$$\theta_{MLE} = \arg\max_{\theta} p\left(x|\theta\right) = \arg\max_{\theta} \prod_{i} p\left(x_i|\theta\right) \tag{3.3}$$

Без потери общности можем максимизировать логарифм от правой части выражения 3.3.

$$\theta_{MLE} = \arg\max_{\theta} \log p\left(x|\theta\right) \tag{3.4}$$

$$\begin{aligned}\theta_{MLE} &= \arg\max_{\theta} \log \prod_i p\left(x_i|\theta\right) \\ &= \arg\max_{\theta} \sum_i \log p\left(x_i|\theta\right)\end{aligned}$$

Покажем подробнее как MAP преобразуется в MLE для случая юниформного распределения θ:

$$\begin{aligned}\theta_{MAP} &= \arg\max_{\theta} \sum_i \log p\left(x_i|\theta\right) p\left(\theta\right) \\ &= \arg\max_{\theta} \sum_i \log p\left(x_i|\theta\right) \; const \\ &= \arg\max_{\theta} \sum_i \log p\left(x_i|\theta\right) \\ &= \theta_{MLE}\end{aligned}$$

Еще одним подходом к оценке θ является метод сопряжённых априорных распределений. В теореме Байеса 3.1 изменяемым является только член $p\left(\theta\right)$, так как функция правдоподобия $p\left(x|\theta\right)$ определяется моделью, а $p\left(x\right)$ данными.

Распределение априорной вероятности называется сопряженным с распределением постериорной вероятности, если они относятся к одному семейству распределений.

Поясним вышеизложенное на примере. Пусть $p\left(x|\theta\right)$ и $p\left(\theta\right)$ являются нормальными распределениями. Для распределения $p\left(\theta\right) = N(x|\mu_0,\sigma_0^2)$ с математическим ожиданием μ_0 и дисперсией σ_0^2 выражение 3.1 может быть записанно в виде 3.5.

$$p(\theta|x) = \frac{\mathbb{N}(x|\theta)\,\mathbb{N}(\theta|\mu_0, \sigma_0^2)}{p(x)} \tag{3.5}$$

Произведение двух нормальных распределений будет так же нормальным распределением, и параметры постериорного распределения можно будет вычислить по следующим формулам.

$$\mu = \frac{\left(\frac{\mu_0}{\sigma_0^2} + \frac{\sum_{i=1}^{n} x_i}{\sigma^2}\right)}{\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}} \tag{3.6}$$

$$\sigma = \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right)^{-1} \tag{3.7}$$

Таким образом использование сопряжённых семейств распределений позволяет избежать сложных вычислений полной вероятности.

3.7.1 Скрытые параметры модели

Говоря о характеристиках НТЦ, как объекта исследования мы не можем измерить такие параметры как интеллектуальный капитал (ИК). Хотя ИК влияет на результаты деятельности НТЦ, которые мы можем измерить, например, на количество публикаций и количество авторов. Будем называть такие параметры как ИК скрытыми параметрами [188].

Для определения ИК можно рассмотреть подход на основании машинного обучения. Например, на основе искусственной нейронной сети. Тогда нам для обучения искусственной нейронной сети будет необходим набор данных, содержащий значения ИК для различных компаний с разными параметрами: количеством публикаций, количеством сотрудников, и.т.п. Из литературы известно, что для обучения искусственных нейронных сетей необходимы наборы данных с сотнями тысяч образцов и сотнями параметров. В действительности такого набора данных для НТЦ не существует. Но даже есть представить, что такой набор данных есть, то он будет содержать много пропущенных значений, противоречивых данных и других проблемных данных.

С другой стороны, Байесовская статистика может работать с небольшими наборами данных. Что приводит нас к рассмотрению вероятностного подхода к оценке скрытых параметров. Первым шагом для построений вероятностной модели является построение зависимости наблюдаемых параметров. И на первый взгляд все параметры будут зависит друг от друга. Например, чем больше авторов, тем больше публикаций, чем больше сотрудников с учеными степенями, тем больше публикаций в журналах из списка ВАК и т.п.

Но такая полносвязанная структура модели не позволяет нам построить структурированную вероятностную модель, так как нам необходимо будет оценить вероятность каждой возможной комбинации параметров. А количество

таких комбинаций будет экспоненциально расти с количеством рассматриваемых параметров.

Одним и решений может быть введение скрытых параметров, таких как ИК, которые уменьшают количество связей. Предположим, что НТЦ обладает ИК от которого зависит число публикаций и количество авторов. Таким образом, количество комбинаций для вероятностной оценки существенно сокращается.

3.7.2 EM-алгоритм

Рассмотрим вероятностную формулировку неравенства Йенсена. Пусть $(\Omega,\mathcal{F},\mathbb{P})$ — вероятностное пространство, и $X\colon \Omega \to \mathbb{R}$ — определённая на нём случайная величина. Пусть также $\varphi\colon \mathbb{R} \to \mathbb{R}$ — выпуклая (вниз) функция. Тогда, если $X, \varphi(X) \in L^1(\Omega,\mathcal{F},\mathbb{P})$, то $\varphi(\mathbb{E}[X]) \leqslant \mathbb{E}[\varphi(X)]$, где $\mathbb{E}[\cdot]$ означает математическое ожидание. Или другими словами для выпуклой функции f и распределения вероятностей t получаем следующее выражение 3.8:

$$f(\mathbb{E}_{p(t)}\, t) \geqslant \mathbb{E}_{p(t)} f(t) \tag{3.8}$$

Для дальнейшего рассмотрения приведем следующее определение расстояния Кульбака — Лейблера 3.9.

$$\mathcal{KL}(q\|p) = \int_x q(x) \log \frac{q(x)}{p(x)} \mathrm{d}x \tag{3.9}$$

Отметим, что точнее будет называть расстояния Кульбака — Лейблера мерой различия двух распределений $q(x)$ и $p(x)$, так как по определению данная мера не обладаем симметрией: $\mathcal{KL}(q\|p) \neq \mathcal{KL}(p\|q)$.

Покажем еще одно полезное свойство расстояния Кульбака — Лейблера: $\mathcal{KL}(q\|p) \geqslant 0$. Для этого произведем следующие вкладки:

$$\begin{aligned}
\mathcal{KL}(q\|p) &= E_q\left(-\log\frac{q}{p}\right) && (3.10)\\
&= E_q\left(-\log\frac{p}{q}\right) \leqslant \log\left(E_q\frac{q}{p}\right) && (3.11)\\
&= \log\int_x q(x)\,\frac{q(x)}{p(x)}\,\mathrm{d}x && (3.12)\\
&= 0 && (3.13)
\end{aligned}$$

Рассмотрим применение EM-алгоритма для нахождения скрытых параметров НТЦ. Предположим, что у на есть модель НТЦ со скрытыми параметрами. Обозначим скрытые параметры как t_i, а наблюдаемые параметры как x_i. Тогда функцию правдоподобия можно выразить как 3.14.

$$p(x_i|\theta) = \sum_c p(x_i|t_i = c)\, p(t_i = c|\theta) \tag{3.14}$$

Где $p(t_i = c|\theta)$ - это априорная вероятность того, что t принимает значение . Задача состоит в том, чтобы максимизировать вероятность функции правдоподобия по θ. Так

как логарифм является выпуклой непрерывно возрастающей функцией, то будем искать максимум логарифма от $p\left(x_i|\theta\right)$. Предположим так же, что все N измерений x_i были сделаны независимо. Тогда вероятность $X = \prod_i^N p\left(x_|\theta\right)$

$$\log p\left(X|\theta\right) = \sum_i^N \log p\left(x_i|\theta\right) = \sum_i^N \log \sum_c p\left(x_i|t_i = c|\theta\right) \tag{3.15}$$

Стоит отметить, что мы можем искать максимум выражения 3.15 с помощью градиентных методов. Например, с помощью метода стохастического градиентного спуска. Но автор сознательно применил другой алгоритм и покажет его преимущества далее.

Применим неравенства Йенсена 3.8 к выражению 3.15 и получим $\log p\left(X|\theta\right) \geqslant \mathfrak{L}(\theta, q)$. Далее выберем функцию $\mathfrak{L}(\theta, q)$ так, чтобы ее легко было максимизировать.

$$\mathfrak{L}(\theta, q) = \sum_i^N \sum_c q(t_i = c) \log \frac{p\left(x_i, t_i = c|\theta\right)}{q(t_i = c)} \tag{3.16}$$

И в итоге для параметра θ и распределения q получим неравенство 3.17:

$$\log p\left(X|\theta\right) \geqslant \mathfrak{L}(\theta, q) \tag{3.17}$$

Теперь для поиска максимума $mathfrakL(\theta, q)$ применим следующий итерационный алгоритм из двух шагов для каждой итерации k:

– Фиксируем θ^k и выбираем q^k так чтобы $\mathfrak{L}(\theta^k, q^k)$ была максимальной
– Получаем $q^{k+1} = \arg\max_q \mathfrak{L}(\theta^k, q)$

Первый шаг принято называть E-шаг, а второй M-шаг. Вместе они представляют EM-алгоритм результатом, которого является θ для скрытой переменной t.

3.7.3 E-шаг

Рассмотрим более подробно E-шаг. Максимизация функции нижней границы $\mathfrak{L}(\theta^k, q^k)$ означает, что расстояние между $\mathfrak{L}(\theta^k, q^k)$ и функцией максимального правдоподобия $\log p\left(X|\theta^k\right)$. Запишем это уравнение для k-й итерации и покажем, что это расстояние можно выразить через рассто-

яния Кульбака — Лейблера.

$$
\begin{aligned}
DIST &= \log p\left(X|\theta\right) - mathfrakL(\theta, q) \\
&= \sum_{i}^{N} \log p\left(x_i,\theta\right) - \sum_{i}^{N}\sum_{c} q(t_i = c) \log \frac{p\left(x_i,t_i = c|\theta\right)}{q(t_i = c)} \\
&= \sum_{i}^{N} \Big\{ \log p\left(x_i|\theta\right) \sum_{c} q(t_i = c) - \sum_{c} q(t_i = c) \log \frac{p\left(x_i,t_i = c|\theta\right)}{q(t_i = c)} \Big\} \\
&= \sum_{i}^{N}\sum_{c} q(t_i = c) \Big\{ \log p\left(x_i|\theta\right) - \log \frac{p\left(x_i,t_i = c|\theta\right)}{q(t_i = c)} \Big\} \\
&= \sum_{i}^{N}\sum_{c} q(t_i = c) \Big\{ \log p\left(x_i|\theta\right) - \log \frac{p\left(x_i,t_i = c|\theta\right)}{q(t_i = c)} \Big\} \\
&= \sum_{i}^{N}\sum_{c} q(t_i = c) \log \frac{p\left(x_i|\theta\right) q(t_i = c)}{p\left(x_i,t_i = c|\theta\right)} \\
&= \sum_{i}^{N}\sum_{c} q(t_i = c) \log \frac{p\left(x_i|\theta\right) q(t_i = c)}{p\left(t_i|x_i,\theta\right) p\left(x_i|\theta\right)} \\
&= \sum_{i}^{N}\sum_{c} q(t_i = c) \log \frac{q(t_i = c)}{p\left(t_i|x_i,\theta\right)} \\
&= \sum_{i}^{N} \mathcal{KL}(q(t_i) \| p\left(t_i|x_i,\theta\right))
\end{aligned}
$$

Таким образом, максимизация функции нижней границы $\mathfrak{L}(\theta^k, q^k)$ эквивалентна минимизации суммы расстояний Кульбака — Лейблера для $q(t)$ и $p\left(t|x,\theta\right)$. Так как расстояния Кульбака — Лейблера неотрицательны по определению, то мы можем приравнять их нулю для поиска глобального

минимума.

$$0 = \sum_{i}^{N} \mathcal{KL}(q(t_i) \| p\left(t_i | x_i, \theta\right)) \tag{3.18}$$

Также из определения расстояния Кульбака — Лейблера известно, оно равно нулю только в случае если оба распределения совпадают.

$$q(t_i) = p\left(t_i | x_i, \theta\right) \tag{3.19}$$

Уравнение 3.19 означает, что для нахождения оптимального распределения $q(t)$ мы должны выбрать его равным постериорному распределению $p\left(t|x,\theta\right)$.

3.7.4 M-шаг

На M-шаге производится максимизация функции правдоподобия 3.16 при фиксированном $q(t)$ по θ.

$$\begin{aligned}
\mathfrak{L}(\theta, q) &= \sum_{i}^{N} \sum_{c} q(t_i = c) \log \frac{p\left(x_i, t_i = c | \theta\right)}{q(t_i = c)} \\
&= \sum_{i}^{N} \sum_{c} q(t_i = c) \log p\left(x_i, t_i = c | \theta\right) - \sum_{i}^{N} \sum_{c} q(t_i = c) \log
\end{aligned}$$

Отметим, что так как выражение $\sum_i^N \sum_c q(t_i = c) \log q(t_i = c)$ не зависит от θ, то при дифференцировании оно обнулится. Таким образом, выражение 3.20 можно

преобразовать следующим образом.

$$\mathfrak{L}(\theta, q) = \mathbb{E}_q \log p\left(X,T|\theta\right) + const \tag{3.20}$$

Напомним, что в выражении 3.20 X - это все данные, а T - это все значения скрытых переменных. $\mathbb{E}_q$ обозначает математическое ожидание распределения q. Так как мы выбираем распределения для X и T, то в наших силах обеспечить чтобы $p\left(X,T|\theta\right)$ была гладкой и непрерывной. Такой выбор значительно упростит нахождения экстремума по θ.

3.7.5 Сходимость EM-алгоритма

EM-алгоритм предназначен для нахождения локальных экстремумов функции максимального правдоподобия. Для этого используется функция нижней границы $\mathfrak{L}(\theta^k, q^k)$, которая в процессе оптимизации не убывает 3.21.

$$\log p(X|\theta^{k+1}) \geqslant \log p(X|\theta^k) \tag{3.21}$$

3.7.6 Использование EM-алгоритма для выявления скрытых тематик в тексте

Научный текст является одним из проявлений деятельности НТЦ. Выявление тематик текста может быть сделано

с использованием распределения Дирихле. Байесовская модель для постериорного распределения скрытых тематик в тексте может быть записана в следующем виде.

$$p\left(W,Z,\Theta\right) = \prod_{d=1}^{D} p\left(\theta_d\right) \prod_{n=1}^{N_d} p\left(z_{dn}|\theta_d\right) p\left(w_{dn}|z_{dn}\right)$$

$$p\left(\theta_d\right) \sim Dir(\alpha)$$

$$p\left(z_{dn}|\theta_d\right) = \theta_{dz_{dn}}$$

$$p\left(w_{dn}|z_{dn}\right) = \Phi_{z_{dn}w_{dn}}$$

$$\sum_{w} \Phi_{tw} = 1$$

$$\Phi_{tw} \geqslant 0$$

Таким образом, что W - это текстовые данные (научные статьи, документы), Φ - распределение слов в каждой тематике, Z - распределение тематик для каждого слова, Θ - распределение тематик в документе. Оптимизационная задача для поиска тематик выглядит следующим образом:

$$P(W|\Phi) \to max_{\Phi} \tag{3.22}$$

Для использования EM-алгоритма выпишем явно уравнения для E-шага и M-шага:

E-шаг:

$$\mathcal{KL}(q(\Theta)\, q(Z) \| p\left(\Theta,Z|W\right)) \to \underset{q(\Theta)\, q(Z)}{\text{minimize}} \tag{3.23}$$

M-шаг:

$$\mathbb{E}_{q(\Theta)\, q(Z)} \log p\left(\Theta,Z,W\right) \to \underset{\Phi}{\text{maximize}} \tag{3.24}$$

Полученные выражения для Е-шага (3.23) и М-шага (3.24) позволяют получить скрытые тематики их текста.

3.8 Моделирование самоорганизующихся команд в научной среде

Самоорганизация рабочих групп представляет большой интерес для научно-технических организаций, ищущих новые формы эффективной организации труда сотрудников. Рассмотрение феномена самоорганизации [189], как альтернативы формированию рабочих групп привело к построению модели процесса самоорганизации. Рассмотрение жизненного цикла рабочей группы по отношению к поставленной цели позволило ввести формальные критерии оценки эффективности работы группы и прогнозировать ее продуктивность.

Важным фактором, влияющим на продуктивности рабочей группы, является характер решаемой задачи. Автор предложил собственную классификацию творческих задач нефтегазовой отрасли на основании компетенций, необходимых для их решения.

В рамках разработанной методики жизненного цикла рабочей группы и классификации задач была построена математическая модель самоорганизации рабочих групп для решения творческих задач. Калибровка математической мо-

дели появления рабочих групп была проведена на данных Газпромнефть НТЦ.

В созданной системе показателей эффективности был проведен цифровой имитационный эксперимент [190] для выявления основных характеристик самоорганизации рабочих групп.

В результате была сделана кластеризация творческих задач [191] по эффективности решения различными рабочими группами, составлены рекомендации по созданию организационных мероприятий, способствующих повышению вероятности самоорганизации рабочих групп, выделены критерии отнесения творческих задач к различным типам рабочих групп, определены основные критерии для формирования эффективных рабочих групп.

Моделирование групповых действий индивидуумов зависит от потенциальных участников группы и целей. Например, фанаты футбольного клуба легко объединяются в группу, но не имеют конкретной цели. С другой стороны, ученые могут объединяться в исследовательскую группу для конкретной цели, например, написания научной статьи. Одинаковы ли принципы объединения в этих случаях?

Разделяют два процесса появления групп:

1. Самоорганизация – процесс упорядочения сотрудников за счёт внутренних факторов, без внешнего специфического воздействия,
2. Формирование – назначение участников группы из вне.

Пример формирования группы в организационной среде:

Начальник управления принимает решение сформировать рабочую группу для создания системы разработки месторождения из двух разработчиков месторождений.

Пример самоорганизации группы в организационной среде:

Несколько разработчиков месторождений решили, что им вместе нужно усовершенствовать методы обработки данных о технических режимах работы скважин с помощью методов машинного обучения и самозабвенно работают в вместе по выходным над этой задачей.

Решение о создании группы внутри ведет к самоорганизации, решение о создании группы из вне формирует группу. Отметим, что на практике процесс появления рабочих групп представляет суперпозицию самоорганизации и формирования. Но для исследовательских целей в данной работе авторы намерено рассматривают формирование и самоорганизацию раздельно для выявления родовых признаков этих явлений.

Группа создается на определенное время и для решения конкретной задачи. В этом смысле на лицо признаки проектной деятельности – уникальность результата и ограниченность ресурсов для достижения поставленной цели. Таким образом, обоснованным представляется применять проектную методику оценки эффективности группы, как проектной команды.

Структура группы определяется характером решаемых задач. Для задач массового обслуживания, например, в рабочие группы в центрах обработки вызовов объединяют специалистов с определенным, одинаковым профилем компетенций. Разделение обязанностей в такой группе почти нет: типовые задачи по обслуживанию требуют стандартизованных действий членов группы. Нагрузка равномерно распределяются по членам группы.

Группа, созданная для решения творческой задачи менее однородна. Для решения творческой задачи необходимы специалисты с разными компетенциями. Образно можно представить, как задача декомпозируется на компетенции участников. И это не равномерное распределение, участникам достаются разные объемы работы в рамках их компетенций. Для приведенного выше примера (Пример 3.8), задача требует компетенций в технических режимах работы скважин и методах машинного обучения. Что будет, если компетенций нужных для достижения цели нужно больше, чем есть у группы? Каждый из участников группы выполнит работы в рамках своих компетенций, но цель не будет достигнута, так как останутся невыполненные работы. Это распространенная ситуация, как при неправильном планировании групп (формировании), так и при самоорганизации групп. Результат работы в такой ситуации оказывается отрицательным, но отношение к этому результату разное в случае самоорганизации и формирования.

Основными критериями эффективности рабочих групп является результат их деятельности и сроки достижения этого результата – это общепринятые организационные показатели эффективности. Исследование феномена самоорганизации рабочих групп в динамике является сложнейшей организационной задачей. Поэтому автор использовал в данной работе математическую модель феномена самоорганизации рабочих групп. Математическая модель дает возможность изучить наиболее характерные аспекты феномена самоорганизации, но обладает определённой степенью приближения, неточности. Вычислительный эксперимент на основании созданной модели самоорганизации рабочих групп, который представлен далее состоит в том, чтобы оценить результаты работы различных групп над различными задачами. В связи с такой постановкой эксперимента возникают следующие исследовательские вопросы:

1. Как происходит самоорганизация групп? Какие сотрудники могут самоорганизоваться, а какие нет? Как на самоорганизацию влияют компетенции, опыт, социальные факторы?
2. Какие организационные условия необходимы для самоорганизации групп в научно-технической среде? Конкурсы? Обучения? Мероприятия?
3. С каким типом задач самоорганизованные группы справляются эффективнее, чем сформированные?

4. Каковы принципы формирования групп для наиболее эффективного решения творческих задач?

Оценке эффективности научно-исследовательских и опытно-конструкторских проектов, а также исследованию факторов, влияющих на результативность научной деятельности посвящены многочисленные публикации, см., например, [192—195]. Как правило, в этих работах научный коллектив рассматривается как “чёрный ящик”, производящий научные результаты, и оценка его эффективности производится только на основе выпускаемых результатов, внутренняя структура исследовательской группы обычно не берется в учёт. Самоорганизующиеся команды подробно изучены в работе [196]. Отдельно исследуются мотивирующие факторы [195] и факторы, влияющие на результативность [194].

При этом тема моделирования и анализа командной работы также является хорошо проработанной и активно исследуется с середины XX века, см. [197—199]. Формальное описание профиля компетенций — это тема многочисленных исследований и публикаций, см., например, [199; 200].

Первым приближением может быть модель ограниченная существованием фиксированного набора определенных навыков. При этом профиль компетенции каждого сотрудника можно описать в виде вектора значений, в котором каждая координата описывает уровень его владения соответствующим навыком.

Вектор, описывающий профиль компетенций команды, получается в результате простого сложения профилей компетенций участников. Такая модель естественно возникает, если измерять уровень компетенции производительностью при выполнении соответствующего типа задач. Тогда естественно предполагать, что при совместной работе в команде производительность участников складывается.

Аналогичным вектором можно также описать и профиль задачи. Для подготовки и проведения научного исследования с учетом ограничения по времени требуется определенный уровень производительности для каждого типа задач.

В данной работе рассматриваются самоорганизующиеся малые команды, в которых инициатива создания исходит от сотрудников. Это допущение соответствует реальной ситуации в большинстве научных коллективов, где администрация может различными способами мотивировать сотрудников подать заявку на участие в той или иной научной конференции или рекомендовать подготовить статью для определенного журнала, но итоговое решение, как правило, остается за научным сотрудником.

В данной работе предполагается, что список компетенций и уровень опыта являются критериями, на основе которого сотрудник принимает решение о присоединении к команде.

В качестве входных данных модели рассматривается набор тематик, которые соответствуют последовательности

поступающих приглашений от конференций и журналов, в которые открыт прием заявок. Для каждого мероприятия или издания известна одна или несколько тем. Подготовка статьи по заданной тематике требует определенного набора компетенций.

Компетенции определяются пространством научной деятельности. В нефтесервисной индустрии набор компетенций отличается от набора компетенций в деревообрабатывающей индустрии.

Опыт описывает вектор определенной длинны и направления в пространстве компетенций. Проекции вектора опыта на оси компетенций показывают опыт в соответствующей компетенции.

Задача, например тема научной статьи, так же представляет вектор в пространстве компетенций. Темы могут требовать компетенций, которыми не обладают авторы по отдельности. Каждый соавтор закрывает только часть требуемых для решений задачи (написания статьи) компетенций.

3.8.1 Старт процесса командообразования

Процесс образования команды начинается с принятия первым участником решения о создании команды для подготовки заявки на конференцию или статьи в сборник.

Происходит это следующим образом. Незанятый сотрудник просматривает список приглашений и производит оценку своих компетенций с точки зрения объявленных тематик. Если хотя бы одна из его компетенций соответствует или превосходит требования цели, он принимает решение о создании команды и становится первым ее участником. В начальный момент профиль компетенций команды совпадает с профилем первого участника. Следующие участники будут присоединяться к этой команде учитывая требования, соответсвующие выбранной тематике, а так же профили компетенций других членов команды.

3.8.2 Присоединение новых участников к команде

Второй (последующий) участник узнает от одного из членов команды о цели и оценке текущих компетенций команды. Эта информация распространяется между сотрудниками, которые достаточно хорошо знакомы друг с другом. В модели это представлено с помощью коммуникационного графа. Каждый последующий участник производит оценку своих компетенций с точки зрения потребностей команды для достижения цели и принимает решение о присоединении к команде. Решение положительно, если хотя бы одна

из компетенций этого участника при добавлении к профилю команды приближает ее к поставленной цели.

3.8.3 Финализация состава команды

В виду ограниченности времени на решение поставленной задачи, время на формирование команд тоже не может быть безграничным. Если в течение отведенного отрезка времени команду с требуемым набором компетенций сформировать не удалось, процесс останавливается, участники освобождаются от принятых обязательств и переключаются на поиск другой задачи. Если же команда успешно сформирована, то считаем, что входящие в нее сотрудники заняты некоторое время и итогом этой работы является публикация.

3.8.4 Формальная модель компетенций

Пусть N обозначает количество ключевых навыков, которые необходимы для работы в данной предметной области, W - множество сотрудников организации. Тогда профилем компетенций для сотрудника называется вектор $\vec{\kappa}(w)$ (3.25).

$$\vec{\kappa}(w) = (\kappa_1, \ldots, \kappa_{\mathbf{N}}), \text{ где } w \in W, \kappa_i \in \mathbf{R}^+. \quad (3.25)$$

Профиль компетенций команды T состоящей из M человек - это вектор той же размерности N, который определяется как сумма по всем участникам команды:

$$\vec{\kappa}(T) = \sum_{i=1}^{\mathbf{M}} \vec{\kappa}(w_i), \text{ где } T = \{w_1, \ldots, w_{\mathbf{M}} : w_i \in W\}. \quad (3.26)$$

Неформально i-я компонента вектора соответствует производительности человека и команды при выполнении задач определенного типа. Профиль тематики p имеет тот же тип, а именно является N-мерным вектором:

$$\vec{\kappa}(p) = (\kappa_1, \ldots, \kappa_N). \quad (3.27)$$

Тут i-я компонента вектора соответствует минимальной производительности команды, при которой все задачи соответствующего типа будут выполнены гарантировано в срок и с надлежащим качеством.

3.8.5 Модель принятия ключевых решений

Для реализации процесса командообразования ключевыми являются функции, моделирующие логику принятия решения на разных этапах формирования команды, а именно:

- $\alpha(w, p)$ описывает выбор цели первым участником команды, а именно $\alpha(w, p) = 1$ если сотрудник w рассматривая цель p принимает положительное решение о создании команды и $\alpha(w, p) = 0$ в противном случае;
- $\beta(w, T, p)$ формализует принятие решения о присоединении к команде вторым и последующими участниками;
- $\gamma(T, p, t)$ моделируют решения о самороспуске в момент времени t на основании сопоставления профиля созданной команды и профиля задачи.

В данном исследовании предполагается, что α, β и γ являются детерминированными булевозначными функциями, которые зависят только от профиля компетенций человека, команды и задачи соответственно.

$$\alpha(w, p) = \alpha'(\vec{\kappa}(w), \vec{\kappa}(p)), \tag{3.28}$$

$$\beta(w, T, p) = \beta'(\vec{\kappa}(w), \vec{\kappa}(T), \vec{\kappa}(p)), \tag{3.29}$$

$$\gamma(T, p, t) = \gamma'(\vec{\kappa}(T), \vec{\kappa}(p), t). \tag{3.30}$$

Пусть далее $\mathbf{K}$ обозначает всё пространство возможных значений вектора компетенций. Тогда тот факт, что в нашей модели алгоритм командообразования зависит только от профилей компетенций участника, команды и цели, задает тип функций α', β' и γ':

$$\alpha' : \mathbf{K}^2 \to \{0, 1\}, \quad \beta' : \mathbf{K}^3 \to \{0, 1\}, \quad \gamma' : \mathbf{K}^2 \to \{0, 1\} \tag{3.31}$$

Эти функции можно описать в виде следующих логических формул:

$$\alpha'(x,y) = 1 \iff \exists i(x_i \geqslant y_i) \tag{3.32}$$

$$\beta'(x,y,z) = 1 \iff \exists i[(x_i > y_i) \wedge (y_i < z_i)] \tag{3.33}$$

$$\gamma'(x,y,t) = 1 \iff \exists i(x_i < y_i) \wedge (t > \tau_{\max}) \tag{3.34}$$

3.8.6 Процесс формирования команды

На момент командообразования фиксирован список открытых задач P и для каждой конкретной задачи $p \in P$ задан её профиль $\kappa(p)$. Также фиксировано множество сотрудников W и для каждого сотрудника $w \in W$ известен профиль его компетенций $\kappa(w)$. Кроме этого задан граф коммуникаций между сотрудниками $G \subseteq W \times W$. Ещё одним параметром является время $\tau_{\max}$ в течение которого команда должна сформироваться.

На каждом шаге последовательно происходит следующее.

1. Каждый сотрудник w_0, который не включен ни в одну из команд и не получил приглашение о вступлении в команду рассматривает список целей P. В случае если находится p_0 для которой

$\alpha(w_0,p_0) = 1$, сотрудник принимает решение о создании новой команды T_0 и отправляет приглашения присоединиться к команде всем соседям в коммуникационном графе G.

2. Если сотудник w_1 не был включен в команду и получил приглашение войти в команду T_1, созданную для решения задачи p_1, он принимает приглашение, если $\beta(w_1,T_1,p_1) = 1$ и отправляет приглашения всем своим соседям в графе G. В противном случае приглашение отклоняется.
3. Если для какой-то команды T_2, созданной для решения задачи p_2, выполняется условие $\gamma(T_2,p_2) = 1$, команда приступает к работе и все приглашения аннулируются.
4. Если для какой-то команды T_3, созданной для решения задачи p_3, спустя заданное время $\tau_{\max}$ выполняется условие $\gamma(T_3,p_3) = 0$, эта команда расформировывается и все приглашения аннулируются.

Несмотря на то, что что α, β и γ являются детерминированными, алгоритм допускает большую степень неопределенности, которая связана с недетерминированным характером взаимодействия объектов внутри системы. В частности, на результат существенно влияют следующие параметры, которые реализуются вероятностно:

– очередность рассмотрения списка задач свободным сотрудником;

– очередность рассмотрения сотрудником полученных приглашений;
– очередность, в которой выбираются сотрудники для применения очередного шага алгоритма.

Построенная модель является основой для дальнейших исследований процесса образования и функционирования проектных команд в научной среде. В частности, на её основе планируется разработать методику оценки эффективности научно-исследовательской деятельности. Также интересным направлением работы является уточнение и расширение модели, в частности:

– Модели компетенций могут быть уточнены с привлечением аппарата нечёткой логики.
– При моделировании долгосрочных периодов появляется необходимость учитывать профессиональное и карьерное развитие сотрудников и сопряженные с этим изменения в их профилях компетенций.
– Функции α, β и γ, описывающие процесс принятия ключевых решений, могут быть уточнены путем учёта других индивидуальных и командных характеристик, а также специфики задач.
– Алгоритм командообразования может иметь более сложную итеративную логику, учитывающую различные подходы к гибкому управлению проектами.
– Отдельной проработки заслуживает ситуация с неуспешным завершением проекта. В терминах научной деятельности это означает, что написанная

публикация не была принята к печати, но полученные результаты являются хорошим заделом для дальнейшей работы. В текущей работе автор сделал допущение, что сотрудники не пишут *в стол*, а каждое соавторство ведет к публикации.

3.9 Методика графа соавторства

Сложившаяся практика построения графов соавторства подразумевает использование математического аппарата теории графов. Традиционно для построения графов соавторства используют неориентированные графы. Граф соавторства представляет наглядную визуализацию выбранного научного сообщества и позволяет производить анализ с помощью таких распространенных метрик графов, как: Betweenness centrality [201—203] и Closeness centrality [204—206]. Данные метрики, как и метрика Degree, предназначены для формального выделения важных вершин графа.

3.9.1 Двудольные графы

Существенным аспектом для построения графа соавторств является выборка данных для анализа. Обычно исследователи используют публичную библиографическую информацию, содержащую список соавторов. Источником такой информации может быть Google Scholar, ArXiv и другие онлайн библиотеки. Рассмотрение открытых научных сообществ так же интересно, как и сужение выборки до одной страны, отрасли [207] и даже организации [181]. Добавление в граф полей, связанных с аффиляцией автора, позволяет сделать исследования отношения организаций. Как пример, в работе [207] авторы анализируют связи между исследовательскими институтами и индустриальными научными центрами в нефтегазовой индустрии. Такой подход к выборке позволяет проанализировать топологию связей между организациями, на основании принадлежности авторов к организации.

Отметим, что все приведенные выше исследования не принимают в расчет содержание исследовательских статей. Эта особенность будет важна в дальнейшем. Среднее количество соавторов может изменяться в зависимости от индустрии, но в целом количество соавторов растет. Отметим этот факт, как структурную особенность исследуемой области.

В приведенных выше исследования граф соавторства строится на неориентированный граф. Авторы равнозначны в соавторстве, хотя на деле это не так. В работе автора [42] проанализирована структура команды соавторов и сформулированы возможные роли в процессе исследования.

Кроме того, в традиционном построении графа соавторства информация о всех совместных исследовательских работа содержится в ребрах графа. Часто ребра рисуют различной толщины или цвета в зависимости от количества совместных работ, но данная характеристика ребер не рассматривается в контексте метрик графа, так как не отражает коммуникационный смысл повторного соавторства. С учетом этих ограничений сформулируем следующий исследовательские вопросы:

Исследовательский вопрос 1 Существуют ли другие способы построения графа соавторств?

Исследовательский вопрос 2 Какими преимуществами и недостатками обладают различные способы построения графа соавторств?

Исследовательский вопрос 3 Каковы количественные, сравнимые характеристики графов соавторств?

В приведенных выше исследованиях граф соавторства строится как неориентированный граф: статьи становятся равнозначными ребрами, соединяющими авторов. Автор

данного исследования считает, что более информативным будет построение графа соавторств как двудольного графа. Такой подход позволяет включить в граф соавторств информацию о научных статьях. На Рисунке 3.9.1 приведен основной принцип построения графа соавторств на основе направленного двудольного графа.

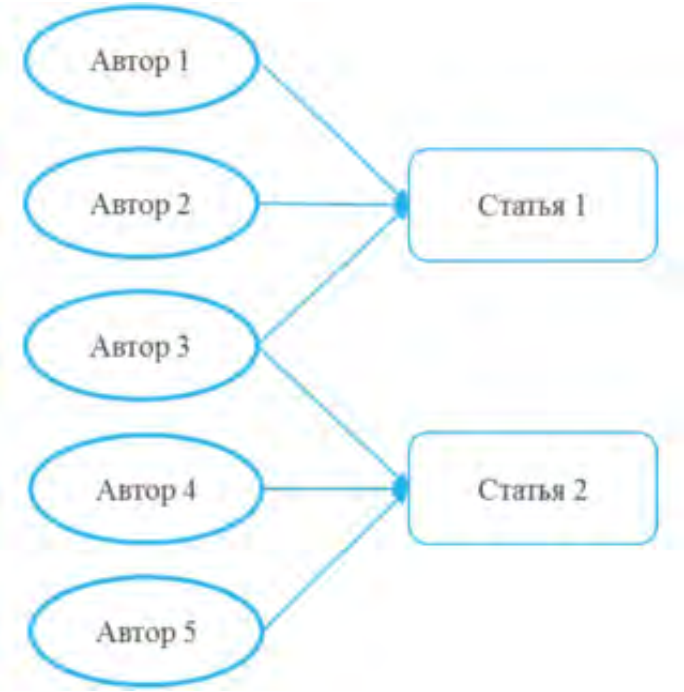


Рисунок 3.6 — Двудольный граф соавторств.

Преимущества такого подхода состоят в том, что в графе соавторств становится возможным сохранить для дальнейшего анализа библиографическую информацию о статье:

1. Название статьи
2. Год издания
3. Издатель
4. Ключевые слова

Отметим, что традиционное представление графа соавторств в виде неориентированного графа является проекцией двудольного графа на множество вершин авторов. Поясним это более подробно. Ориентированный граф $G = (V,E)$ называется двудольным, если множество его вершин можно разбить на две части $A \cup P = V$, так, что

- ни одна вершина в A не соединена с вершинами в P
- ни одна вершина в P не соединена с вершинами в A.

В рассматриваемом случае A - это множество авторов, P - это множество статей. A и P - являются долями графа G. Отметим, что граф G может быть как полным так и неполным в зависимости от того имеют ли авторы соединения со всеми статьями. Приведенный на Рисунке 3.9.1 двудольный граф является неполным. Обозначим G_A проекцию графа G на множество вершин A. Граф G_A является традиционным представлением графа соавторств и отображен на рисунке 3.9.1.

Из рисунка 3.9.1 видно, что при построении проекции атрибутами ребер графа G_A могут стать только интегральные характеристики соавторства, например, количество соавторств двух авторов.

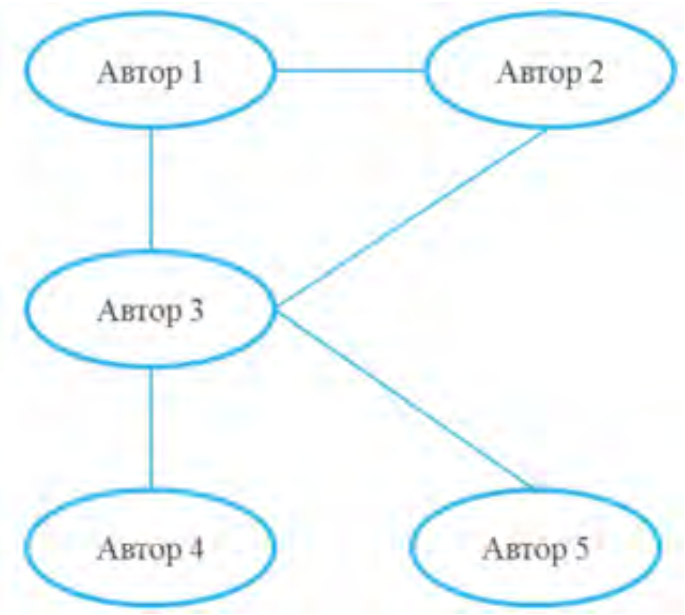


Рисунок 3.7 — Неориентированный граф соавторств.

3.9.2 Моделирование графов соавторства

Для моделирования графов соавторства как социальной сети широко применяется следующие стохастические подходы:

1. Случайные графы.
2. Модель “Маленький мир” (small-world).
3. Модели, основанные на концепции преимущественного присоединения (preferential attachment).

Одним из существенных ограничений стохастических моделей является фиксированное количество рассматрива-

емых вершин, или их постоянный рост. На практике в организации изменяется количество потенциальных соавторов. Так же важно понимать, что стохастические модели ставят целью моделировать граф с определенными параметрами. Такими как, кластеризация, плотность и др.

С другой стороны, формирование малых групп, к которым относится и группа соавторов научной статьи, моделируется на основании принципа дополнительных компетенций, относящегося к классу детерминированных методов создания графов соавторства.

Для предсказания новых вершин графа соавторства применяется комбинированный подход на основе машинного обучения с предварительным отбором признаков авторов, статистических показателей активности за несколько последних временных промежутков, а также структурные индексы влияния и локальные метрики в сети соавторства. Полученные в данной работе результаты позволили сделать вывод о применимости методов предсказания связей для анализа коллаборативных шаблонов поведения в крупной организации с динамической структурой коллектива, а также меняющимися внешними и внутренними факторами, влияющими на индивидуальную и коллективную публикационную активность.

Основой для прогнозирования изменений графа соавторств для научно-технического центра являются следующие компоненты:

– Текущая структура графа соавторств.

- Внешние по отношению к научно-техническому центру воздействия.
- Внутренние изменения научно-технического центра.

Рассмотрим каждую из компонент подробнее. Текущая структура графа соавторств представляет набор метрик, описывающих данный граф соавторств. К таким метрикам принято относить следующие:

- Для ребер
    - Common Neighbours (CN)
    - Salton Index (SI)
    - Jaccard Index (JI)
    - Hub Promoted Index (HPI)
    - Hub Depressed Index (HDI)
    - Leicht-Holme-Newman Index (LHN1)
    - Preferential Attachment Index (PA)
    - Adamic-Adar Index (AA)
    - Resource Allocation Index (RA)
- Для вершин
    - Degree centrality
    - Betweenness centrality
    - Closeness centrality
    - Harmonic centrality
    - Clustering

Каждая из этих метрик представляет характерный набор признаков (features) графа соавторств, влияющих на прогноз его изменений. Внешние по отношению к научно-техническому центру воздействия заключаются в

публикационной политике редакций, публикующих научные статьи. В простейшем случае отсутствие возможности опубликовать статью из-за ограничений по объему выпуска журнала приводит несостоявшемуся соавторству. Основные зависимости публикационной активности научно-технического центра от редакций рассмотрены в работе [42]. Внутренние изменения научно-технического центра вызваны изменениями в составе персонала. В организацию приходят новые сотрудники, некоторые сотрудники увольняются. В процессе наставничества и обучения сотрудники приобретают новые компетенции. В результате проведения НИР рождаются новые исследования и научные заделы. Часто, изменения внутренних требований к качеству публикаций также могут стать причиной структурных изменений, подтверждая принцип “publish or perish”, и влияя как на структуру коллектива, так и на параметры активности отдельных сотрудников и исследовательских коллективов. Рассмотрим подробнее в чем состоит прогноз развития графа соавторств для научно-технического центра. Под развитием будем понимать возникновение новых вершин и ребер. Граф соавторств может рассматриваться как накопленным итогом за период, так и инкрементальными изменениями по годам. Далее будем рассматривать факт авторства, как признак вершины графа соавторств. Другими словами, сотрудник, представляемый вершиной графа соавторства, может как написать, так и не написать статью в следующем временном периоде. Процесс

прогнозирования в данном случае будет решать задачу бинарной классификации. Для каждого сотрудника будет определяться вероятность создания статьи по определенной тематике. Статья является коллективным усилием работы соавторов, обладающих определенным набором компетенций, нашедших свое применение в цели исследования. В этом состоит основная идея принципа дополнительности компетенций. Авторы с одинаковыми компетенциями не имеют рационального обоснования для объединения с целью проведения научного исследования. Будем считать, компетенции атрибутами вершин графа. Для выявления компетенций необходимых для написания статьи будем использовать ключевые слова, а при их отсутствии – метод тематического моделирования текста работы.

3.10 Современные процессы организации труда на основе гибких методик

Гибкие (Agile) методики разработки программного обеспечения широко применяются в различных индустриях. Написание программного кода по своей сути является процессом создания логически структурированного текста так же, как и написание научной статьи. Коллективная работа над написанием научных статей требует разделения обязанностей для повышения продуктивности так же, как

и написание программного кода требует выделения специалистов для тестирования и документирования.

Использование ролевой модели гибких методик представляется перспективным кросс-индустриальным опытом для применения, но нуждается в теоретической проверке. Одним из вариантов проверки гипотез, показавшим себя в условиях, когда постановка реального эксперимента представляется высоко затратной, является метод имитационного моделирования. Автор видит дополнительные преимущества от институциализации процесса написания научных статей и применения проверенных индустриальных показателей эффективности для его оценки.

Провозглашение основных принципов гибких методик в виде манифеста [208] обозначило насущную необходимость перехода к более эффективным методам разработки программного обеспечения. Решительность этого шага многократно себя оправдала на практике и в последствии нашла теоретические обоснования bonner2016empirical. Суть гибких методик может быть изложена по-разному, но для данного исследования нами выбрана следующая формулировка:

1. Приоритет командных взаимодействий
2. Приоритет работающего программного кода
3. Приоритет реакций над планом

Современные методики написания научных статей остаются на позициях последовательного, “водопадного” подхода. Такой подход был уместен во времена Ньютона,

когда один уникальный ум работал над трудом всей своей жизни. В условиях современной скорости обмена научной информацией одиночки остаются не у дел. Им на смену приходят научные коллективы. Интуитивно понятно, что от согласованной работы научного коллектива соавторов зависит их продуктивность: оптимальное соотношения качества и скорости публикации результатов научных исследований в виде научных статей, доступных наиболее широкому кругу заинтересованных лиц. В гибких методиках разработки программного обеспечения образование команды основано на принципах самоорганизации [196; 209]. Самоорганизующиеся команды в работе [210] разделены на три типа:

1. “Команда пилотов самолёта”: Управление воздушным судном.
2. “Компьютерные команды”: Создание новых программных продуктов.
3. “Команда КВН”: Решение сложных проблем.

Для целей дальнейшего исследования нас больше будут интересовать тип “Компьютерные команды”.

3.10.1 Размеры команд

Размеры команд играют важную роль. В гибких методиках разработки программного обеспечения рассмат-

риваются малые (5-7), большие (10-50) и сверхбольшие команды (100-200) [211].

Важно отметить, что приведенные оценки сходятся с полученными в работе [181] для команд соавторов: современные творческие команды соавторов в среднем состоят из 3 участников. В дальнейшем изложении будем подразумевать, что число участников команды состоит в среднем из 3 соавторов.

3.10.2 Образование команд

Гибкие методики [208] подразумевают под самоорганизацией только возможность работы команды с ограниченным управлением из вне. На взгляд автора данного исследования, целесообразно рассмотреть, как образуются команды в деталях.

Для рассмотрения механизма необходимо понимать, что команда образуется с определённой целью. Рассматривая образование команд авторы исследования [212] предлагают эмпирический вероятностный алгоритм присоединения нового участника к уже сформированной группе (Рис.3.10.2):

Таким образом, авторы работы [212] оценивают влияние внутренней структуры команды на её расширение.

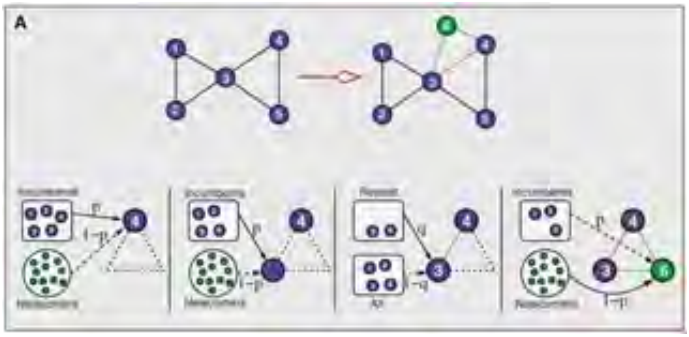


Рисунок 3.8 — Алгоритм образования команд на основе вероятностей (p - для новых участников, q - для участников группы) [212]

В работе [210] отмечено, что основным фактором для самоорганизации команд являются индивидуальные компетенции. При этом компетенции каждого участника оцениваются с точки зрения полезности для достижения цели. В исследованиях [213; 214] утверждается, что такая оценка приводит к появлению системы статусов участников команды, которая выражается в иерархичности коммуникаций. Для настоящего исследования нам достаточно того, что:

1. Цель является образующим базисом для команды;
2. Цель декларирует потребности в компетенциях участников команды;
3. Участники производят оценку компетенций друг друга для достижения цели.

Базовый алгоритм образования команды для двух участников может быть представлен в виде следующей временной последовательности (3.10.2):

Рисунок 3.9 — Базовый алгоритм образования команды.

Приведенная на Рис. 3.10.2 последовательность описывает основное образующее команду действие - *парное объединение*. Можно сказать, что после присоединения участников у команды появляется собственный профиль компетенций для достижения заданной цели. Компетенции команды получаются в результате суперпозиции компетенций участников. Некоторые из компетенций необходимых для достижения цели "закрыты опытом" участников, а некоторые нет. Следующий за первым участник присоединяется уже с учётом профиля компетенций команды. Для удобства дальнейшего изложения сформулируем следующие утверждения:

Утверждение 1 Сотрудники объединяются в команду для достижения цели.

Утверждение 2 Компетенции команды являются функцией от компетенций участников.

Утверждение 3 Объединение первого участника с командой для достижения цели происходит по таким же принципам, что и объединение команды из n участником с n+1 участником.

3.10.3 Парное объединение

Рассмотрим подробнее *парное объединение*. Организационная среда задает размерность N_{comp} пространства компетенций. Каждый участник организационной среды a обладает вектором компетенций c_i таких, что $i \in N_{comp}$. Каждая компетенция участника c_i характеризуется опытом e_i . Опыт участника – это натуральное число, $e_i \in \mathbb{N}$. В результате можно сказать, что участник обладает вектором опыта в пространстве компетенций. Отметим, что пространство компетенций организационной среды обладает существенно большей размерностью, чем вектор компетенций участника. Изначально команда t_0 не содержит участников и не обладает собственными компетенциями (Рис. 3.10.3).

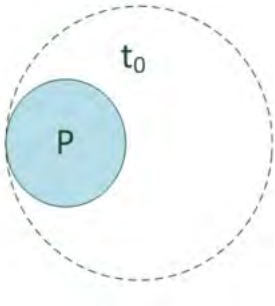


Рисунок 3.10 — Схема команды без участников.

Пусть P обозначает цель для объединения команды, c_j – вектор компетенций, а e_j – опыт по каждой компетенции необходимый для достижения P. В результате успешного объединения для достижения цели будет образована команда t_1(Рис. 3.10.3).

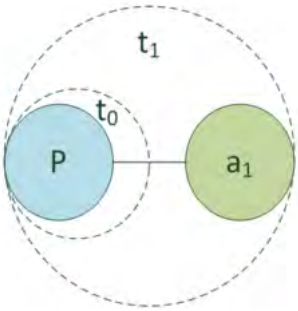


Рисунок 3.11 — Схема команды из одного участника.

Команда t_1 обладает новым вектором компетенций. Так как в t_1 только один участник a_1, то вектор компетенций t_1 будет совпадать с вектором компетенций a_1. При присоединении к t_1 участника a_2 будет образована команда t_2 (Рис. 3.10.3).

Так как участник присоединяется ко всем элементам команды , то можно привести схему (Рис. 3.10.3) к виду графа команды (Рис. 3.10.3).

Цель является P атрибутом ребра, связывающего a_1 и a_2, поэтому можем преобразовать граф команды с двумя участниками к виду (Рис. 3.10.3).

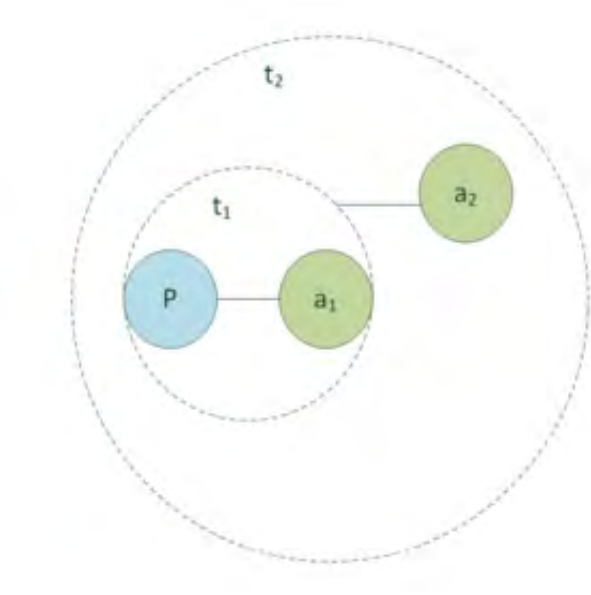


Рисунок 3.12 — Схема команды из двух участников.

Для случая написания научных статей граф команды t_2, изображенный на Рис.3.10.3 обозначают $g(t_2)$ и называют графом соавторства, где под P подразумевают научную статью. Информация о истории создания команды в такой нотации не приводится. На Рис.3.10.3 приведен пример фрагмента графа соавторства. Вершинами графа являются исследователями, а ребрами – совместная научная публикация. Граф соавторства является ненаправленной сетью.

Отметим, что для наглядности на Рис.3.10.3 размер вершин отражает количество научных статей, написанных участником.

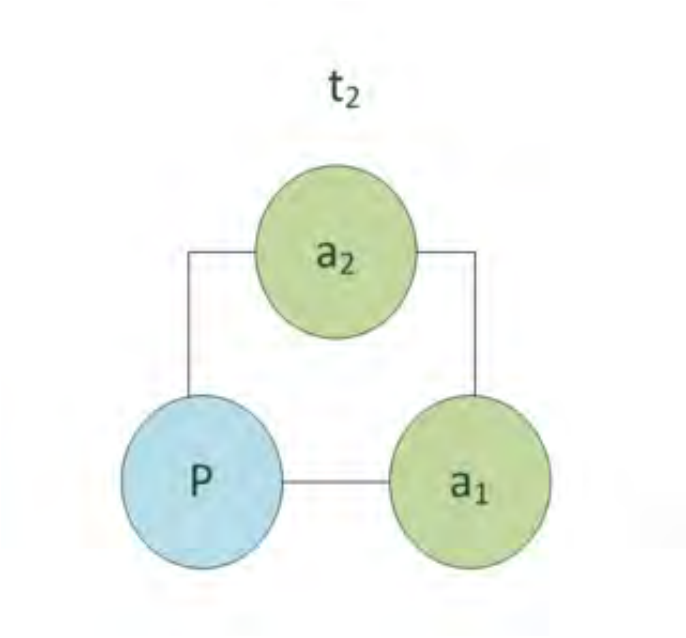


Рисунок 3.13 — Граф команды из двух участников с избыточными связями.

Командный код

Введем понятия *полного командного кода (ПКК)* и *остаточного командного кода (ОКК)*. Эти понятия играют ключевую роль в образовании команды. Составляющими командного кода являются компетенции. По своему типу *полный* и *остаточный командный код* – это вектора в пространстве N_{comp} .

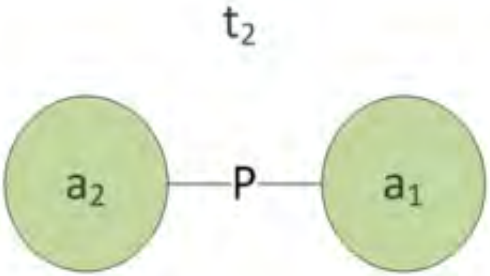


Рисунок 3.14 — Граф команды из двух участников.

Рисунок 3.15 — Фрагмента графа соавторства для нескольких команд.

Рассмотрим, как участником производится оценка своих компетенций с точки зрения потребностей в компетенциях для достижения цели.

Характеристики цели P являются основанием для образования команды. То есть, первый участник команды a и цель P должны быть объединены на основании представления о компетенциях. Другими словами, необходимыми условиями для достижения цели должно быть обладание a определенным набором компетенций и опыта. С точки зрения множеств компетенции сотрудника a и цели P должны находиться в одном пространстве и иметь пересечения. Наличие пересечений будет приводить к объединению в команду.

Введем функцию оценки в виде $\Phi(P,t,a), \Phi \in [0,1]$. Результатом Φ будет вероятность возможности объединения участника a в команду t для достижений цели P. Тогда функция Φ для n-ого участника может быть записана в виде $\Phi_n(P,t_n,a_n)$.

По мере присоединения участника вектор компетенций команды будет изменяться. В него будут входить компетенции новых участников, а опыт по одинаковым компетенциям будет складываться.

$$ut_{n-1} = \prod_{a_j}^{a_{n-1}} \sum_{i}^{N_{comp}} \left\{ c_j * e_i \right\} \tag{3.35}$$

Величину ut_{n-1} будем называть *полным командным кодом (ПКК)*. *ПКК* характеризует потенциал команды для достижения целей.

В соответствии с вышеописанным алгоритмом (Рис. 3.10.2) функцию Φ можно представить в виде:

$$\Phi = P \cdot \prod_{a_j}^{a_{n-1}} \sum_{i}^{N_{comp}} \left\{ c_j * e_i \right\} \cdot a_n \tag{3.36}$$

Важную смысловую часть в выражении (3.36) несет компонент rt_n^P, который автор называет *остаточным командным кодом – ОКК.*

$$rt_n^P = P \cdot \prod_{a_j}^{a_{n-1}} \sum_{i}^{N_{comp}} \left\{ c_j * e_i \right\} \tag{3.37}$$

ОКК rt_n^P характеризует незакрытые t_n командой компетенции цели P. Нулевой вектор в качестве *ОКК* характеризует полную укомплектованность компетенциями команды для достижения цели.

С учетом *ОКК* можно преобразовать выражение (3.36) следующим образом:

$$\Phi = rt_n^P \cdot a_n \tag{3.38}$$

Выражение (3.38) имеет интуитивно понятный смысл: *для оценки возможности присоединения к команде новый участник должен выяснить обладает ли он необходимым опытом в требуемых для выполнения цели компетенциях с учетом того, что существующая команда уже закрыла часть из необходимых компетенций своим опытом.* В работах [215; 216] такой принцип образования команд называют комплементарным.

Гомогенность команд

Мы рассмотрели образование команд на основе дополнительности (комплементарности) компетенций. Второй движущей силой для образования команд является гомогенность.

Гомогенность групп в социальных сетях, или склонность людей со схожими характеристиками формировать связи между собой, также называемая гомофилией, является важным фактором формирования и эволюции социальных сетей [79]. Во многих работах отмечается динамическая структура гомофилии [80; 81], в ходе которой параллельно происходят два процесса. С одной стороны – схожие между собой индивиды формируют социальные связи (социальная селекция). С другой – уже связанные друг с другом люди перенимают поведение друг друга (социальное влияние). Совокупность этих факторов результирует в гомогенную социальную систему, в которой между индивидами со схожим поведением и характеристиками есть связь, при этом характер связи может быть, как формальным, так и неформальным.

Несмотря на то, что связи между индивидами со схожими характеристиками более вероятны, чем связи между непохожими, уровень схожести также важен. В работе [15] было показано, что социальная схожесть более, чем по одному показателю, приводит к тому, что люди с меньшей

вероятностью будут формировать между собой взаимоотношения. Автор объясняет данный эффект тем, что слишком схожие по многим характеристикам люди, как правило, не могут привести что-то новое и конструктивное во взаимные отношения или же в команду. Для продуктивного сотрудничества необходима не только схожесть интересов, но также и различный профессиональный и жизненный опыт, позволяющий предложить многомерные подходы к ее решению.

Основным объединяющим фактором в команде являются компетенции участников, влияющие на достижения цели. Основываясь на понятии *ОКК*, введенного ранее, можно рассмотреть остаточные компетенции участника, то есть компетенции, не востребованные для объединения команду для достижения цели. Влияние этой части компетенций на команду может как усиливать ее, так и ослаблять во время работы.

Работа команд

Начало работ по достижению цели определяется участниками команды и не зависит от процесса образования команды. Показатели производительности могут быть только у работающей команды. Например, важное для научной сферы деятельности понятие *научный задел* означает ни что

иное как работы, выполненные командой имеющей не пустой *остаточный командный код.*

Формирование основной системы внутреннего взаимодействия внутри команды согласно исследованию [78] происходит при знакомстве участников. Таким образом, для данного исследования будем пренебрегать временем установления устойчивой работы каналов коммуникаций.

В гибких методиках разработки программного обеспечения наибольшее внимание уделяется именно коммуникациям внутри команды [217] и с внешними агентами [218], которые по сути тоже являются командой, но в более широком смысле.

Сформулируем следующее утверждение:

Утверждение 4 Характеристики работы каналов коммуникаций соответствуют характеру работы команды.

Таким образом, измеряя работу коммуникационных каналов можно сделать заключения о характере работы команды. Отметим, важное следствие: такой тип измерения производительности команды не создаёт дополнительной нагрузки на сотрудников в отличии от методик оценки основанных опросах.

Вопрос измерения вклада отдельных участников или результата команды рассмотрен в ряде работ [219; 220] и все исследователи склоняются к тому, что измерять нужно и командную производительность (Team Performance), и индивидуальная продуктивность (Individual Performance). В

исследовании [221] приведена следующая схема измерений (Рис. 3.10.3).

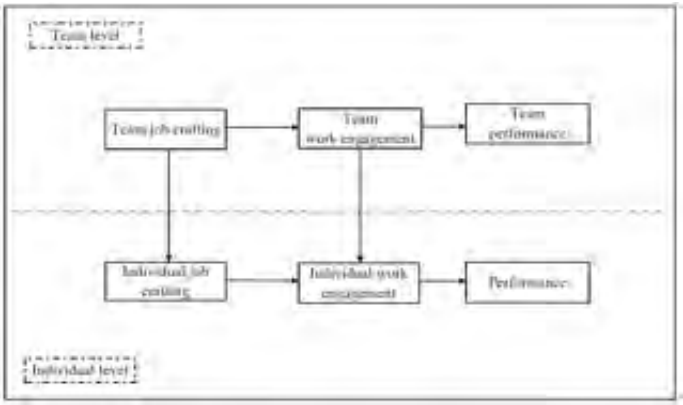


Рисунок 3.16 — Уровни измерения производительности команды и участника [221].

Например, измерение Individual Performance с помощью опросов исследуется в работе [222] путём введения Creative Solution Diagnosis Scale (CSDS) – шкалы креативности. Измерить Individual Performance сотрудника по такой шкале авторы [222] предлагают с помощью Consensual Assessment Technique (CAT), которая требует дополнительных усилий от сотрудников. Сильные и слабые стороны метода опросов для измерения Individual Performance изложены в фундаментальной работе [223].

Вопрос метода измерения Individual Performance находит интересную постановку в современной концепции “sensible organization” [224]. Авторы исследования [224] помимо измерения традиционных цифровых коммуникационных

каналов надевают на сотрудников браслеты, отслеживающие перемещения и другие параметры организма.

Вопросы зависимости производительности команд от структуры команд рассмотрены в исследовании [181].

Методика Scrum

Одной из распространённых гибких методик командной работы является методика Scrum [225]. Scrum предназначен для получения наилучших из возможных результатов для командной разработки сложных интеллектуальных продуктов. В классическом Scrum существует 3 базовых роли:

- Product owner – отвечает за соответствия целям
- Scrum master – отвечает за эффективное взаимодействие в команде
- Команда разработки (Development team)

Рекомендуемый размер Scrum команды — 5-7 человек соответствует принятым в данном исследовании ограничениям. Согласно идеологам Scrum [225], команды большего размера требуют значительных ресурсов на коммуникации, в то время как команды меньшего размера уменьшают размер работы, который команда может выполнить в единицу времени.

Основой Scrum является Sprint, в течении которого выполняется работа над продуктом. Sprint имеет одинаковую продолжительность на протяжении всего процессы создания продукта, рекомендуется одна неделя. Задача Sprint состоит в том, чтобы материализовать продукт в текущем виде. Продуктом в данном исследовании является научная статья.

Методика Scrum декларирует необходимость в определенных видах деятельности, не связанных с исследованиями и написанием текста, которые приводят к лучшей результативности. Кроме этого Scrum задает определенный ритм для этих дополнительных деятельностей.

Введем показатели, на которые влияет применение Scrum к процессу написания научных статей:

1. Ускорения обмена сообщениями в каналах коммуникаций;
2. Потеря работ из-за дублирования при отсутствии своевременных коммуникаций о прогрессе проведения исследований;
3. Потеря работ из-за несоответствия написанной статьи правилам публикации.

С точки зрения формализма графа соавторства применение Scrum приведет к выделению вершин графа, обеспечивающих функции *Product owner (PO)* и *Scrum master (SM)* (Рис.3.17)

С графом $g(t_2^{Scrum})$, отображенным на Рис.3.17 можно произвести преобразование аналогичное сделанному выше с

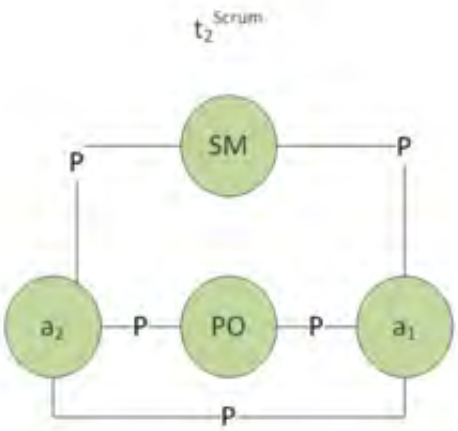


Рисунок 3.17 — Соавторство с ролями Scrum.

графом $g(t_2)$. Как мы видим, Scrum роли PO и SM соединяют вершины a_1и a_2. Из чего следует, что PO и SM являются характеристиками ребер графа, соединяющего a_1 и a_2. Преобразованный граф соавторства с применением Scrum ролей отображен на Рис. 3.18.

Роли Scrum согласно [208] не должны вмешиваться в содержательную часть работы команды, а лишь ускорять информационный обмен и устранять информационные барьеры. Сформулируем это в виде гипотезы, формальное доказательство которой отложим для дальнейших исследований:

Гипотеза 1 Введение ролей Scrum в процесс соавторства не изменяет вид графа соавторства.

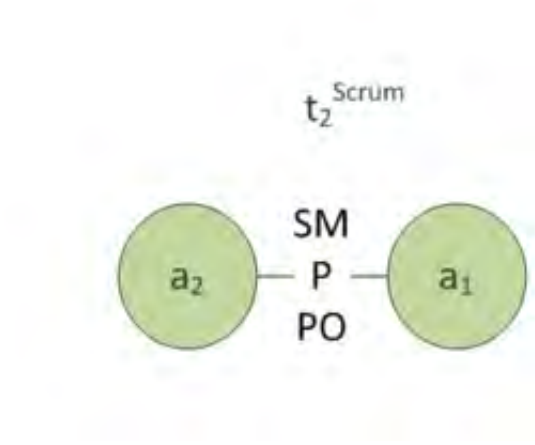


Рисунок 3.18 — Граф соавторства с атрибутами Scrum.

Теперь рассмотрим показатели производительности работы команд.

Показатели производительности команд

В современной работе [226] рассмотрены вопросы разработки показателей, измеряющих скорость перехода продукта из фазы исследований (research) в фазу разработки (development). Авторами [226] предложена интегральная модель для таких показателей. Объектом измерения авторы считают знания, а показатели основывают на процессе Knowledge Management. Никаких конкретных KPI авторы

не предлагают, но описывают пространственные оси своей модели – процессы, инструменты и люди.

Связь между способностью команды собраться и ее продуктивностью исследовали в работе [227]. Стоит отметить, что в работе [227] образование команды подразумевает формирование, а не самоорганизацию.

Состав команды во времени не постоянен и говорить от том, что образование команды в тот или иной момент времени завершено не корректно. Участники могут покидать команду и участвовать в нескольких командах одновременно. Важная веха в работе команды определяется нулевым OKK, когда все компетенции необходимые для достижения цели представлены участниками команды. Сформулируем это в виде утверждения:

Определение 2 Команда считается укомплектованной, тогда и только тогда, когда ее ОКК равен нулевому вектору в пространстве N_{comp}. Минимальное время, в котором ОКК стал равен нулевому вектору называется Временем комплектации (T_c).

Отметим, что T_c может быть больше времени, отведенного издательством или программным комитетом научной конференции на подготовку. Таким образом, статья не будет обладать требуемыми качествами в срок и не будет принята к публикации.

Показатели, наиболее точно отражающие динамику выполнения работы, будут основываться на изменении в

динамике всех параметров команды. Введем функцию применения командой опыта в определенных целью компетенциях: $E(P,t)$. Факторами, негативно влияющими на E будут сложность коммуникаций внутри команды $\Xi(g)$ и необходимость заниматься деятельностью, не направленной на создания научных статей $\Gamma(t)$.

И $\Xi(g)$, и $\Gamma(t)$ будут увеличивать время, требуемое на написание научной статьи. Таким образом, команда может не достигнуть цели в определенные сроки.

Сформулируем две рассмотренные причины не достижения командой цели:

Определение3 Несостоявшейся научной статьей (ННС) будем считать статью, не уложившуюся во временные рамки публикационного процесса с требуемым качеством.

Отношение количества несостоявшихся статей $(Frac_{notpub})$ к количеству опубликованных статей является показателем производительности процесса написания научных статей.

Другим более очевидным показателем производительности является время, затраченное на публикацию научной статьи (T_{pub}).

3.11 Анализ текста

3.11.1 Анализ текста на основании тематик

В последние годы бурно развиваются методики тематического моделирования. Недавние исследования привели к развитию нескольких основных направлений: вероятностного [228], на основе SVD [229] и генеративного [230]. Тематическое моделирование определяет каждую тему как распределение некоторого количества слов с определенными вероятностями. Большинство современных тематических моделей строятся на основе распределения Дирихле (LDA, Latent Dirichlet Allocation) [231]. Трудно представить, что настолько универсальное распределение как LDA будет одинаково хорошо работать для любых текстов. Необходимы тонкие настройки алгоритма на конкретный проблемный домен. Поэтому автор сосредоточился на основном мировом источнике для научно-практических статей нефтегазовой отрасли – библиотеке OnePetro. Важно отметить, что OnePetro охватывает широкий спектр инженерных дисциплин и содержит тексты на английском посвященные именно практическим аспектам применения новых технологий в нефтегазовой отрасли. Авторами этих

статей являются сотрудники нефтяных компаний со всего мира.

Формальная постановка задачи тематического моделирования следующая. Пусть зафиксирован словарь терминов W, из элементов которого складываются документы, и дана коллекция D документов $d \in D$. Для каждого документа d известна его длина n_d и количество $n_d w$ использований каждого термина w. Пусть $\Phi = (\varphi_{wt})$ - матрица распределений терминов (w) в темах (t), а $\Theta = (\theta_{td})$ - матрица распределений тем (t) в документах (d). Тогда задача тематического моделирования состоит в том, чтобы найти такие матрицы Φ и Θ, чтобы выполнялось равенство (3.39).

$$p(w|d) = \sum_{t \in T} \varphi_{wt} \theta_{td}, \tag{3.39}$$

где φ_{wt} - вероятности терминов w в каждой теме t, θ_{td} – вероятности тем t в каждом документе d, а $p(w|d)$ – вероятность появления термина w в документе d.

Уравнение (3.39) можно представить в матричном виде $\Phi \cdot \Theta$. При этом легко показать, что данная задача имеет много решений (3.40).

$$\Phi \cdot \Theta = \Phi \cdot \Lambda \cdot \Lambda^{-1} \cdot \Theta = \hat{\Phi} \cdot \hat{\Theta}, \tag{3.40}$$

где $\hat{\Phi} = \Phi \cdot \Lambda$, а $\hat{\Theta} = \Lambda^{-1} \cdot \Theta$.

Из уравнения (3.40) следует, что матрицы $\hat{\Phi}$ и $\hat{\Theta}$ так же будут являться решениями уравнения (3.39). Но не все матрицы Φ и Θ будут содержать хорошо интерпретируемые тематики. Таким образом, в задачу (3.39) необходимо

ввести условия способствующие получению адекватных и интересных тематик. Образно можно сказать, что необходимо оцифровать специфику предметной области текста для встраивания в алгоритм поиска оптимальных матриц Φ и Θ. Отметим, что при использовании LDA для создания тематической модели такой настройки на предметную область не производится. Для решения подзадачи настройки тематической модели на предметную область автором использован механизм регуляризаторов.

Пусть $p(t)$ — распределение тем в коллекции документов:

$$p(t) = \sum_{d} p(d)\,\theta_{td} \tag{3.41}$$

Тогда полезным представляется регуляризатор на основе дивергенции Кульбака-Лейблера:

$$\mathcal{KL}(\Theta) = -\tau \sum_{t \in T} \ln \left(\sum_{d \in D} p(d)\,\theta_{td} \right) \to max \tag{3.42}$$

Где τ – параметр регуляризации, который нужно подобрать в зависимости от предметной области коллекции документов. Требование максимизации $\mathcal{KL}(\Theta)$ будет означать обнуление вероятностей появления документов и приведет к большей разрежённости матрицы Θ. Вторым механизмом для регуляризации может быть обратное действие – увеличение вероятностей для тематик, которые присутствуют во многих документах. Такие тематики называют шумовыми. Для получения уплотнений строк матрицы Θ с шумовыми тематиками можно применить регуляризатор

(3.42) с обратным знаком. Таким образом, матрица Θ после регуляризации будет представлять чередование зон разрежённости для основных тематик и уплотнений для шумовых тематик.

Полученную тематическую модель необходимо формально проверить на качество. Для этого в процесс обучения необходимо встроить метрики качества модели. А после достижения формальных критериев сходимости на основании метрик провести визуализацию модели для общего контроля качества. Основной метрикой для выявления факта сходимости модели тем является метрика Perplexity вычисляемая по формуле (3.43).

$$\mathcal{P}(D, \Phi, \Theta) = \exp\left(\frac{-1}{n_d} \sum_{d \in D} \sum_{w \in d} n_{dw} \ln\left(\sum_{t \in T} \varphi_{wt} \theta_{td}\right)\right) \quad (3.43)$$

Метрика Perplexity не нормирована и поэтому не может быть использована для сравнения сходимости разных моделей. Общая логика состоит в том, что чем меньше Perplexity, тем лучше модель. Поэтому для принятия решения о достаточной сходимости модели руководствуются тем, что Perplexity перестает значительно уменьшаться с ростом количества итераций обучения.

Результирующая модель тематик может быть рассмотрена как мягкая кластеризация. В таком случае к полученным тематикам могут быть применены инструменты визуализации, используемые для кластеров. Например,

могут быть применены методы обучения на основе многообразий (Manifold Learning): t-distributed Stochastic Neighbor Embedding (TSNE) и Multidimensional scaling. Результаты работы алгоритма TSNE зависят от выбранной метрики расстояния между векторами. При размерности векторного пространства в несколько сотен применяют следующие метрики:

– Косинусная мера (Cosine): $\frac{v_1 \cdot v_2}{\|v_1\|_2 * \|v_2\|_2}$
– Евклидово расстояние (Euclidean):$\|v_1 - v_2\|_2$

Для эффективного использования визуализации тематической модели с помощью методов обучения на основе многообразий необходимо представить слова, составляющие тематики, в векторном пространстве (Vector Space Model). Такая процедура называется word embedding. Для нее часто используют метод GloVe описанный в исследовании [115]. Альтернативным методом word embedding является FastText [232], поэтому автор данного исследования решил провести качественное сравнение обоих методов word embedding на выбранной коллекции. Оба метода обучают векторные представления слов на основании того, как часто слова употребляются вместе. Отличие между ними состоит в том, что FastText условно можно назвать “предиктивной”, а GloVe основывается только на частотах слов. В этом свете GloVe гораздо проще, а автор данного исследования верит, что простота в бизнесе — это залог эффективности.

Библиотека BigARTM [233] позволяет выстраивать последовательно несколько регуляризаторов и управлять

группами тематик. Такой инструмент является уникальным на момент написания данного исследования. Широко используемые на западе методы построения topic models на основе LDA не дают таких возможностей.

3.11.2 Анализ эмоциональной окраски текстов

Анализ тональности текста предназначен для выявления в текстах эмоционально окрашенной лексики. Иногда исследователи выделяют термин Opinion mining, подчеркивая тем самым задачу поиска в текстах оценочных суждений. Кроме академического изучения тональности текста как одного из разделов компьютерной лингвистики существует ряд прикладных исследований, направленных на улучшение процесса принятия управленческих решений.

Применение рекуррентных и сверточных нейронных сетей для анализа тональностей позволило достичь значительно большей точности по сравнению с моделями основанными логистической регрессии.

Автор сфокусировался на методике выбора оптимальной архитектуры и гиперпараметров нейронной сети, позволяющие обучить классификационную модель на публичном наборе данных, содержащем оценочные суждения, и затем предсказать фрагменты текста из научно-практических ста-

тей, содержащие оценочные суждения с заданной степенью точности.

Примененные автором методические подходы могут быть представлены в следующем методическом каркасе исследования (Рис. 3.11.2). Рассмотрим более подробно каждый из элементов методического каркаса.

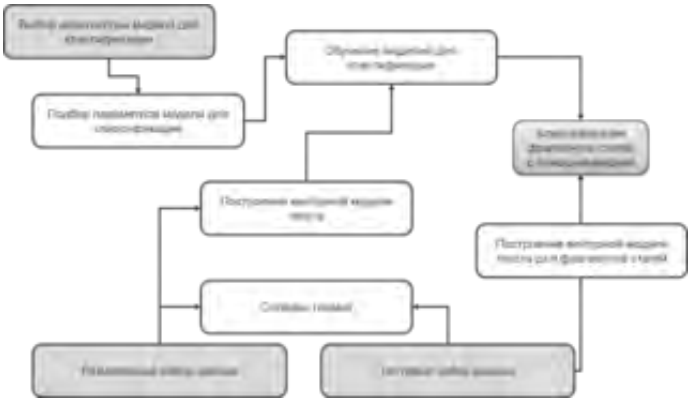


Рисунок 3.19 — Методический каркас исследования эмоциональной окраски текстов.

3.11.3 Методика сравнения корпусов текстов

Методический каркас исследования (методика Т4С) представлен на рисунке 3.20.

Рисунок 3.20 — Методика T4C - методический каркас исследования.

Мультимодальная тематическая модель строит распределения тем на терминах $p(w|t)$, авторах $p(a|t)$, метках времени $p(y|t)$, связанных документах $p(\hat{d}|t)$, рекламных баннерах $p(b|t)$, пользователях $p(u|t)$, и объединяет все эти модальности в одну тематическую модель. Модальный подход к обучению тематической модели позволяет проводить единое обучение, а потом рассматривать матрицу θ для различных значений одной из модальностей. Например, $\theta(M_1|V_1)$ будет содержать только распределение тематик по документам, относящимся к модальности M_1 со значением V_1. Важно отметить, что при этом пространство тематик, матрица φ, является общим для всех документов.

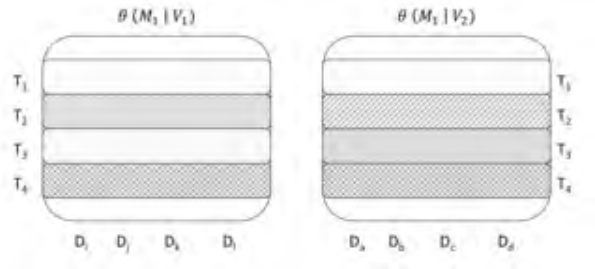


Рисунок 3.21 — Пример матриц θ для разных значений модальности M_1.

На рисунке 4.55 приведён пример матриц θ для разных значений модальности, показывающий как могут выглядеть плотности для тематик. Мы видим, что для тематик T_1 и T_4 плотности для разных значений модальности M_1 могут визуально похожи, а для тематик T_2 и T_3 плотности визуально могут отличаться. Представленная на рисунке 4.55 схема позволяет визуально оценить степень похожести двух коллекций документов с помощью модальной тематической модели. На количественном уровне определить степень похожести можно проведя классификацию значений модальности по матрице θ. Вектора со значениями вероятностей тематик для каждого документа могут быть разбиты на обучающую и тестовую выборку для обучения классификатора определять метки значений модальности.

Возвращаясь к задаче обучения тематической модели, важно отметить, оно сводится к максимизации логарифма правдоподобия $\mathcal{L}(\Phi,\Theta)$ с дополнительными аддитивными ре-

гуляризаторами $\mathcal{R}(\Phi,\Theta)$:

$$\mathcal{L}(\Phi,\Theta) = \sum_{d\in D}\sum_{w\in d} n_{dw} ln \sum_{t\in T} \varphi_{wt}\theta_{td} \tag{3.44}$$

$$\mathcal{L}(\Phi,\Theta) + \mathcal{S}(\mu,\Phi,\Theta) \to \max_{\Phi,\Theta} \tag{3.45}$$

$$\mathcal{S}(\mu,\Phi,\Theta) = \sum_i \mu_i R_i(\Phi,\Theta) \tag{3.46}$$

Слагаемое $\mathcal{S}(\mu,\Phi,\Theta)$ в уравнении (3.47) представляет суперпозицию применяемых регуляризаторов R_i с весами μ_i. n_{dw} - это количество повторений слова w в документе d.

Оптимизационная задача (3.45) не может быть решена градиентными методами, так как функционал $\mathcal{L}(\Phi,\Theta)$ не является дифференцируемым. Поэтому для нахождения Φ и Θ используется двухступенчатый подход к оптимизации, называемый EM-алгоритмом.

3.11.4 Изучение когерентности двуязычного корпуса текстов

Для формальной оценки качества нескольких машинных переводов используют метрику BLEU [234]. В нашем случае данная метрика не подходит, так как мы имеем только один вариант перевода. Можно так же рассмотреть гипотетический способ для проверки выдвинутой исследовательской гипотезы, в котором носителем английского языка

выполняется перевод каждой статьи и потом производится сравнение переводов. Но такой способ требует значительного количества ресурсов. И мы не сможем оценить варианты перевода, в которых авторы расширили свою научную статью при переводе. С точки зрения метрики BLEU такие варианты не являются точными.

С другой стороны, можно наоборот выполнить перевод русскоязычных научных статей на английский с помощью таких средств СМП, как Moses [235] или Phrasal [236]. Таким образом, сделав baseline-оценку, отклонения от которой можно будет измерять с помощью метрики BLEU. Но для того чтобы обучить СМП нужен большой двуязычный корпус научных статей по определённой тематике, которого у авторов нет.

Для небольшого набора данных автор выбрал методику выделения тематик из текста с помощью последовательной регуляризации, предложенную в работе [161]. Обучение модели для выделения тематик производится на каждом одноязычном корпусе документов отдельно. Затем для каждой пары документов сравниваются выделенные тематики.

Суть тематического моделирования состоит в разложении векторного представления текста с помощью двух матриц (3.47):

$$p(w|d) = \sum_{t \in T} \varphi_{wt} \theta_{td} \tag{3.47}$$

Где φ_{wt} матрица вероятностей слов w в каждой тематике $t \in T$, а θ_{td} распределение вероятностей тематик t в документе d, а $p(w|d)$ – условная вероятность слова w в документе d . Для обучения данной модели использовался механизм минимизации кросс энтропии с последовательным добавлением регуляризирующих слагаемых. Без регуляризации матрицы φ_{wt} и θ_{td} не представляют практического интереса.

В множество тематик T целесообразно включить тематики двух типов – основные (sbj_i) и шумовые (nz_j). К шумовым тематикам можно отнести введение и обзор научных источников. Например, большинство статей по СМП будут цитировать во введении одни и те же фундаментальные для этой области знаний научные работы. Хотя основная стема статей будет отличаться. Для шумовых тематик авторами проведена регуляризация со сглаживанием, для основных тематик регуляризация с разряжением. Таким образом, снижен уровень шума в основных тематиках.

Подбор регуляризационных коэффициентов τ был выполнен авторами по методике описанной в исследовании [237]. После обучения с регуляризацией матрицы φ_{wt} и θ_{td} стали разряженными на 80%.

Содержание матриц φ_{wt}^{rus} и φ_{wt}^{eng} представляют распределение тематик для каждого документа. Таким образом, для анализа скрытого представления переводимого текста ν становится достаточным проанализировать соответствие тематик для разных языков. При этом возникает задача

перевода одного названия темы с русского на английский. Ввиду высокой степени разрежённости матрицы φ объем этих работ будет не большим и может быть выполнен с помощью электронного словаря по нефтегазовой тематике.

При сопоставлении φ_{wt}^{rus} и φ_{wt}^{eng} возникает задача выравнивания тематик для ν. Другими словами, одна тематика на русском может соответствовать нескольким тематикам на английском, одной тематике или оставаться без соответствия. Сложность такого выравнивания в матричном представлении описывается квадратной матрицей с размерностью n_T^2.

На основании полученного матричного представления ν можно провести классификацию связей между статьями на разных языках.

3.11.5 Оптимизация тематической модели текста

Тематические модели успешно используются для кластеризации текстов уже на протяжении многих лет. Один из наиболее распространённых подходов к тематическому моделированию на основе LDA [231] моделирует выбранное в качестве параметра фиксированное количество тем на основе распределения Дирихле, для слов и документов. В результате получается плоская, мягкая вероятностная

кластеризация терминов по темам и документов по темам. Все полученные тематики равноправны, они сами по себе не создают каких-либо характерных признаков, которые могли бы помочь исследователю определить наиболее полезные темы, то есть выбрать подмножество тем, которые лучше всего подходят для интерпретации человеком. Проблема нахождения метрики, характеризующей такую интерпретируемость, является предметом изучения многих исследователей [238—241].

Тематическая модель не умеет читать мысли исследователя и поэтому должна иметь параметры настройки на задачу, которую собирается решать исследователь. Тематические модели на основе LDA обладают следующими параметрами согласно исследованиям [242; 243] :

- alpha: параметр априорного распределения Дирихле для документов-тем,
- beta: параметр априорного распределения Дирихле для тем-слов,
- tn: Количество тем,
- b: Количество отбрасываемых начальных итераций при семплировании по Гиббсу,
- n: Количество присемплов,
- si: Интервал семплирования.

В исследовании [243], опубликованном в 2018 году, предпринята попытка нахождения оптимальных значений вышеприведённых параметров с помощью алгоритма Differential Evolution [244]. В качестве кост-функции (мет-

рики) была выбрана модифицированная метрика Jaccard Similarity. В результате был создан новый алгоритм LDADE, в котором появились свободные параметры от алгоритма Differential Evolution, которые тоже нужно будет оптимизировать.

Существует разница между оценкой полного набора тем и оценкой отдельных тем для фильтрации нежелательной информации (шума). Для оценки полного набора тем исследователи обычно смотрят на метрику перплексия [245] для корпуса документов. Такой подход не очень хорошо работает по результатам исследований [246; 247] потому что метрика перплексия не имеет явного минимума, а с ростом итераций выходит на асимптоту [248].

Наиболее распространённое использование метрики перплексия состоит в том, чтобы обнаружить «эффект локтя», то есть когда характер роста упорядоченности модели принципиально изменяется. Перплексия зависит от мощности словаря и распределения частот слов в коллекции, отсюда получаем её недостатки:

- невозможно оценивать качество удаления стоп–слов и нетематических слов
- нельзя сравнивать методы разреживания словаря
- нельзя сравнивать униграммные и n-граммные модели.

Сами авторы LDA сделали исследование качества тематик с помощью Байесовского подхода в работе [249]. Следует отметить, что вопрос оптимального количества те-

матик решён с помощью иерархического процесса Дирихле (HDP) [250], не для документов, а для коллекции в целом. Поясним разницу между Latent Dirichlet Allocation (LDA), иерархическими процессами Дирихле (HDP) и иерархическими распределениями Дирихле (hLDA) [251; 252], так как это разные модели. LDA создаёт плоскую, мягкую вероятностную кластеризацию терминов по темам и документам по темам. В модели HDP вместо фиксированного количества тем для документа количество тем генерируется процессом Дирихле, что приводит к тому, что количество тем также является случайной величиной. «Иерархическая» часть имени относится к другому уровню, добавляемому процесс Дирихле, создающий количество тем, а самим темы по-прежнему являются плоскими кластерами. Модель hLDA является адаптацией LDA, которая моделирует темы как распределение нового, заранее определённого количества тем, взятых из распределения Дирихле. Модель hLDA по-прежнему рассматривает количество тем как гиперпараметр, то есть независимо от данных. Разница в том, что кластеризация теперь иерархическая: модель hLDA изучает кластеризацию первого набора тем, предоставляя более общие абстрактные отношения между темами (а, следовательно, словами и документами). Отметим, что все три описанные модели (LDA, HDP, hLDA) добавляют новые свободные параметры, которые требуют оптимизации, как отмечено в исследовании [253]. Одним из основных требований к тематическим моделям является интерпрети-

руемость человеком [254]. Другими словами, содержат ли темы слова, которые, согласно субъективным суждениям человека, являются репрезентативными для единой когерентной концепции. В работе [255] Ньюмен показал, что человеческая оценка интерпретируемости хорошо коррелирует с автоматизированной мерой качества, называемой когерентностью. В исследовании [256] 2018 года предлагается минимизировать энтропии Реньи и Цаллиса для нахождения оптимального количества тем в тематическом моделировании. В этом исследовании тематические модели, полученные из больших коллекций текстов, рассматриваются как неравновесные сложные системы, где количество тем рассматривается как эквивалент температуры. Это позволяет вычислять свободную энергию таких систем - значение, через которое легко выражаются энтропии Реньи и Цаллиса. Полученные на основе энтропий метрики позволяют найти минимум в зависимости от количества тем для больших коллекций, но на практике небольшие коллекции документов встречаются так же достаточно часто. В исследовании [257], опубликованном в 2018 году, года предложен матричный подход к повышению точности определения тематик без использования оптимизации. Но с другой стороны в исследовании [258] отмечено, что повышение точности модели противоречит с интерпретируемостью человеком. В частности, в исследовании [259], завершённом в 2018 году, создан фреймворк VisArgue, предназначенный для визуа-

лизации процесса обучения модели с целью определения наиболее интерпретируемых тематик.

Использование статистической меры TF-IDF в качестве метрики количественной оценки качества тематик изучено в работе [260]. Так же есть ряд исследований совмещения преимуществ тематических моделей и плотных представлений векторов слов [261—264] .

Мотивацией проведённого автором исследования стал тот факт, что изучение стабильной метрики для качества тематик продолжаются. И использование кластерного анализа является одним из инструментов для анализа стабильности тематик [265] и оптимального количества тем [266], но при этом не рассматриваются преимущества от возможностей специальной подготовки тематической модели с последовательной регуляризацией и плотного представления векторов слов.

Для валидации качества кластеров разработано достаточно много метрик. Например, метрики Partition Coefficient [267], Dunn Index [268], а также DPI [269] и её модификации [270; 271], Silhouette [272], которые задействованы в алгоритмах кластеризации. Но в случае тематической модели мы уже получаем кластеры тематик и не нуждаемся в алгоритме кластеризации, а только в оценке полученных кластеров. Для валидации кластеров необходимо их рассмотрение в пространстве обладающим понятиями близости и удалённости. Для слов такими пространством является векторное представление слов. Значительные ре-

зультаты в этом направлении получены в исследованиях [273—275]. Слова, представленные в виде плотных векторов, отражают смысловое представление и обладают свойствами близости и удалённости. Таким образом, представив тематики в виде плотных векторов автор настоящего исследования создал новую вариацию метрики DPI для тематик, которую автор назвал cDPI.

Рассмотрим способы построения тематической модели для конкретной коллекции документов. Будем называть коллекцию однородной, если она содержит документы одного типа. Например, коллекция научных статей одной конференции, созданных по единому шаблону, является однородной. В случае однородной коллекции научных статей, каждый документ обладает схожей структурой, постулируемой шаблоном конференции. Все научные статьи состоят из введения, представления результатов исследования и заключения. Таким образом, с точки зрения гипотетической тематической модели можно представить документ в виде распределения основной темы и вспомогательных тем: введения и заключения. Конечно, основные темы в разных документах могут быть разные. Но мы можем ограничить коллекцию научных статей выбором определённых рубрик из тематического рубрикатора конференции. Тогда число тем нам будет известно. На рисунке 3.22 представлена матрица распределения тем по документам.

Как мы видим на левой части рисунка 3.22 при построении тематической модели выделены такие темы,

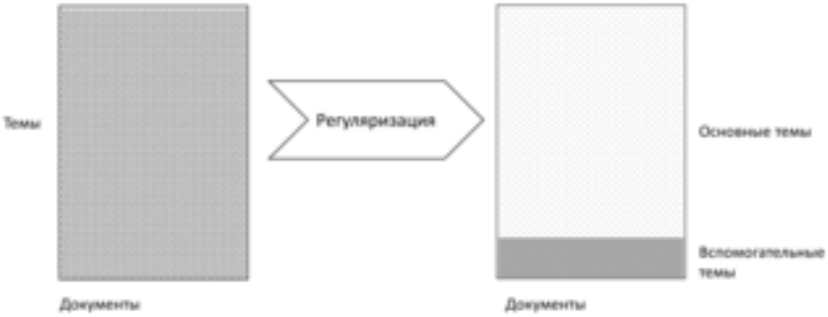


Рисунок 3.22 — Схема матрицы “тема-документ”.

которые распределены по документам достаточно однородно. Такая картина вероятностей матрицы “тема-документ” может быть получена с помощь, например, модели на основе алгоритма LDA [231]. А на правой части рисунка 3.22 показан результат работы модели с последовательной регуляризацией ARTM [276]. Основные и вспомогательные темы выделены с помощью управления процессом обучения модели. Принцип отнесения темы к вспомогательным может быть сформулирован, как наличие такой темы в подавляющем количестве документов. То есть, вероятности вспомогательной темы будут распределены по документам однородно и плотно. А основная тема будет представлена в виде разряженного вектора для каждого документа, так как каждый документ характеризуется одной основной темой. Покажем, что существующие внутренние метрики тематической модели не подходят для определения оптимального количества тем. Для этого рассмотрим внутренние автома-

тизированные метрики качества тем. Введём понятие ядра тематик:

$$W_t = \{w \in W \,|\, p\,(t|w) \geqslant threshold\} \tag{3.48}$$

На основе ядра тем могут быть рассчитаны следующие метрики качества тематической модели:

- Чистота тем: $Purity = \sum_{w \in W_t} p(w|t)$
- Размер ядра тематик : $|W_t|$
- Контраст тематик : $\frac{1}{|W_t|} \sum_{w \in W_t} p(t|w)$
- Когерентность тематик : $Coh_t = \frac{2}{k(k-1)} \sum\limits_{i=1}^{k-1} \sum\limits_{j=1}^{k} PMI(w_i, w_j)$, где k is the interval — окно в котором вычисляются совместные употребления слов, поточечная взаимная информация $PMI(w_i, w_j) = \log \frac{N \cdot N_{w_i w_j}}{N_{w_i} \cdot N_{w_j}}$, $N_{w_i w_j}$ – число документов, в которых слова w_i и w_j хотя бы один раз встречаются в окне k. N_{w_j} – число документов, в которых слово w_i встретилось хотя бы один раз, а N – это количество слов в словаре.

Как видно из формул для внутренних метрик тематической модели каждая из этих метрик может быть измерена для разного количества тем (tn). Рассмотрим поведение метрики *Размер ядра* в зависимости от количества тем. При увеличении количества тем размер ядра будет уменьшаться, так как при построении матриц “тема–слова” и “документ–тема” должны выполняться нормирующие условия: сумма вероятностей должна быть равна единице. Для метрик *Чистота тем* и *Контраст тем* характер изменений

при росте количества тем также будет монотонно убывающим, так как сумма вероятностей тематик, входящих в ядро будет уменьшаться. С другой стороны, для метрики *Когерентность тем* поведение с ростом количества тем будет монотонно возрастающим, так как будет расти вклад от *PMI*. Конкретный характер изменения рассмотренных метрик может отличаться, поэтому целесообразно с помощью численных методов попробовать найти экстремум, если он есть. Рассмотрение качества тематик коротких сообщениях с точки зрения кластеров было проведено в работе [277] с помощью NMF (Non-negative Matrix Factorization) и метрик отражающих энтропию кластеров. Матричный подход (LSI + SVD) к выделению кластеров тематик из программного кода был исследован в работе [278] с модифицированной метрикой близости векторов. В исследовании качества тематической модели [266] использована метрика Silhouette Coefficient [272] с Евклидовым расстоянием для разряженных векторов тематик. Таким образом, в этих работах не исследованным остаются кластеры в пространстве плотных векторов слов, составляющих тематики, и не Евклидовые расстояния в метриках. В работах [248; 279; 280] обнаружена и исследована нестабильность тематик относительно порядка обрабатываемых документов. Поэтому для вычисления метрики качества тематик необходимо провести расчёты для корпуса документов со случайным порядком, чтобы исключить наличие зависимости от порядка документов. В работе [281] показана возможность стабилизации

тематической модели с помощью регуляризации. На основе проведённого анализа авторами был сформулирован методический каркас, изображённый в виде схемы на рисунке 3.23 .

На рисунке 3.23 изображена последовательность действий, повторяемая для одного корпуса документов значительное количество раз, по порядку сравнимому с количеством документов в корпусе. Справа отображены действия, которые выполнены однократно: формирование словаря, настройка параметров регуляризации тематической модели и преобразование разряженного пространства представления тематик в плотное представление. На основании данного методического каркаса были разработаны и проведены цифровые эксперименты, описанные в следующем разделе.

Рисунок 3.23 — Методический каркас исследования.

Глава 4. Апробация и результаты

4.1 Постановка эксперимента для прямой и обратной задач.

В исследовании есть два крупных направления, которые тесно между собой переплетаются и дают комплексный, углубленный взгляд на изучаемый объект. Перечислим эти направления:

- *Изучение деятельности НТЦ по внешним проявлениям.* К внешним проявлениям относятся цифровые артефакты деятельности организации - это опубликованные научные статьи, материалы конференций, информационные сайты в сети Интернет и новости о компании. Изучение цифровых артефактов производится с помощью подходов, основанных на анализе текстов и соавторов.
- *Изучение НТЦ изнутри.* К исследованиям в этом направлении относятся моделирование научной деятельности, эффективность производственных процессов, самоорганизации малых творческих коллективов и модели персонала научной организации.

Допустим, что мы рассматриваем конкретную организацию с определенным количеством сотрудников, бюджетом и планом работ.

Нас интересуют вопросы эффективности данной организации. И с этой точки зрения для нас представляются важным следующие вопросы:

- Каков интеллектуальный потенциал организации? Какие научные исследования организация способна выполнить самостоятельно, а какие необходимо выполнять совместно с другими научными организациями. Очевидно, что при выполнении совместных исследований возникают коммуникационные издержки и исследование нуждается в дополнительной координации. Но для эффективности важна не только принципиальная граница "может-не может но и распределение по времени и затрачиваемым усилиям.
- Какова загруженность персонала? Известно, что в условиях высокой загрузки производительность деградирует. Но для вопросов эффективности данный эффект необходимо рассматривать в динамике, так как возвращение из деградированного к нормальному состоянию занимает определенное время. Кроме того, важна сегментация загрузки по типам сотрудников. Новички могут быть как перегружены, так и недогружены работой. От этого зависит текучесть персонала. Но загрузка экспертов значительно более существенно влияет на эффективность. Эффекты интеллектуальной усталости экспертов драматически влияют на эффективность.

– Научный задел организации истощен? Какова динамика создания научного задела? Есть ли прорывные направления в научных исследованиях, ведущихся внутри организации? Кто участвует в создании научного задела?

Перечисленные параметры организации невозможно измерить. Но от них принципиально зависит эффективность НТЦ. Методы оценки данных параметров разработанные автором, дают методологические подходы к прояснению поставленных выше вопросов.

Прямой метод измерения в настоящем исследовании сводится к моделированию динамики организационной среды для получения цифровых артефактов. Для этого автором созданы модели персонала, модели командообразования [282] и модели продуктивности НТЦ. Результатом многопрогонного эксперимента с этими моделями являются синтетические цифровые артефакты деятельности научной организации: соавторства, тематики, направления развития и др.

Обратный метод постановки эксперимента в свою очередь анализирует реальные цифровые артефакты деятельности НТЦ. А именно научные статьи, материалы конференций и т.п. И на основании цифровых артефактов автор строит модель соавторства [283], модели научных тематик, модели научных направлений и научных школ в организации.

4.2 Модель процесса публикаций научных статей

В настоящем исследовании была проанализирована публикационная активность научно-технического центра ПАО “Газпромнефть” в электронной библиотеке OnePetro международного сообщества нефтегазовых инженеров. Полученная зависимость изображена на рисунке (Рис. 4.1).

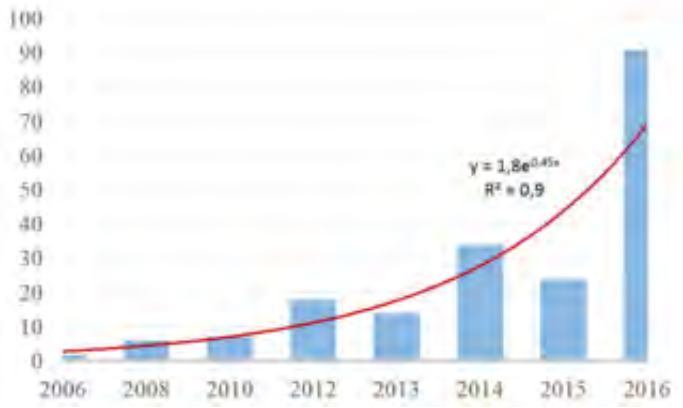


Рисунок 4.1 — Количество публикаций сотрудников Газпромнефть НТЦ в электронной библиотеке OnePetro и линия тренда.

Экспоненциальный рост публикаций в одном издании не может продолжаться бесконечно. Каждое издание имеет свой предельный объем публикаций, рукописи, поступающие сверх допустимого изданием объема публикаций, повышают конкуренцию за право быть опубликованным.

Но в результате отбора некоторые качественные рукописи отвергаются издателями. Для изучения процесса публикации была разработана имитационная модель. Когнитивная карта модели процесса публикаций приведена на рисунке (Рис. 4.2).

Рисунок 4.2 — Когнитивная карта модели процесса публикаций.

Созданная модель процесса публикаций содержит два накопителя:

– Researchers — исследователи,

– PublishersQueue — очередь рукописей.

Модель управляется посредством следующих свободных параметров (Таб. 3):

На основании когнитивной карты модели процесса публикаций был проведен цифровой эксперимент. На рисунке (Рис. 4.3) представлена зависимость эффективности публикаций от времени при различном количестве издателей.

Таблица 3 — Свободные параметры модели процесса публикаций.

Название параметра	**Описание**
Publishers	Количество издателей
PapersPerIssue	Количество статей в выпуске
IssuesPerPublisher	Количество выпусков на одного издателя в год
NewPapersRate	Скорость создания рукописей
InitialResearchers	Начальное количество исследователей
NewResearchersRate	Скорость появления новых исследователей

Падение эффективности публикаций как мы видим имеет резкий, лавинообразный характер. Такой характер поведения эффективности публикаций требует особого внимания, чтобы не пропустить начало стагнации и принять организационные меры для расширения количества издателей, участвующих в процессе публикации.

Принцип разделения труда ведет к повышению эффективности процессов. Гипотетически расширение ролевой модели может повысить эффективность процесса публикаций. Независимо от количества соавторов процесс публикации определяет следующие роли:

- Продюсер — носитель основной идеи исследования
- Редактор — изменяет текст рукописи
- Рецензент — диалектическая противоположность Продюсера, оппонирует, отвечает за выводы и результаты исследования
- Переводчик — если статья не на родном языке авторов, то требуется технический перевод и вычитка

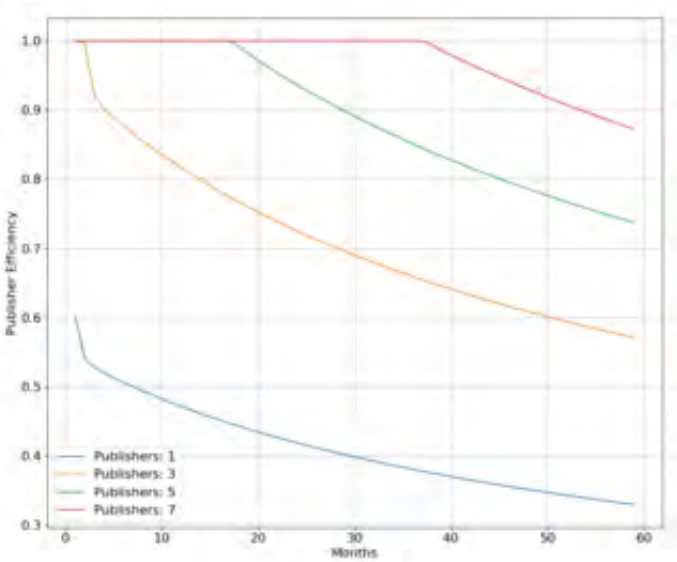


Рисунок 4.3 — Кривая зависимости эффективности публикаций от времени при различном количестве издателей (1,3,5,7).

– Специалист по работе с издателями — отвечает за поиск издателей и внешние коммуникации
– Дизайнер — картинки, презентация для доклада
– Докладчик — представляет результат в устном виде на конференции (если нужно и столько сколько нужно раз)
– Исследователь данных — проведение компьютерных расчетов.

С учетом перечисленных ролей научно-исследовательская команда не изменяется. Авторами исследования остаются именно те ученые, которые его провели. Повы-

шается качество рукописи, а коммуникации становятся более профессиональными. Отметим, что функции внешних и внутренних корпоративных коммуникаций обычно присутствуют в организационной среде, но не имеют фокусировки на работу с индивидуальными потребностями исследователей.

4.3 Измерение интеллектуального капитала НТЦ

Интеллектуальный капитал (ИК) по своей природе является составным показателем продуктивности научно-исследовательской организации основным продуктом которой являются знания [284]. В структуру ИК входят:

– Человеческий капитал

– Организационный капитал

Человеческий капитал (ЧК) - включает в себя знания и навыки. Организационный капитал (ОК) включает в себя технологии и процессы. Другими словами, ЧК характеризует опыт сотрудников, а ОК характеризует то как сотрудники применяют свой опыт к поставленным задачам в данной организации.

Помимо создания интеллектуального капитала можно рассмотреть и его разрушение – сотрудники, которые вели исследования увольняются и уносят с собой знания.

Вклад сотрудников в интеллектуальный капитал не равнозначен. Далее авторы определяют роли, которые относят к “ядру команды”. Потеря сотрудников, состоящих в ядре команды драматически сказывается на производительности. К ядру относят сотрудников с высоким уровнем опыта и наиболее востребованными в организации навыками.

Существует достаточно много подходов, описывающих жизненный цикл сотрудника внутри организации или должности , однако большинство исследований соглашается в выделении 4 основных этапов относительно уровня продуктивности:

1. начальный этап,
2. накопление опыта,
3. продуктивный этап,
4. спад продуктивности.

При этом этап адаптации (начальный этап и накопление опыта) может отличаться в зависимости от вида деятельности и уровня позиции, но в среднем для специалистов и руководителей среднего звена занимает до полугода, около года для руководителей высшего звена.

Наибольший процент увольнений среди новичков, поэтому больше внимания необходимо уделять социальной адаптации новых сотрудников, встраиванию новичков в процессы и наставничеству.

При распаде творческой команды утечка мозгов бывает разная и не всегда наносит вред производительности.

Другими словами, иногда уход опытного, но имеющего отличную от большинства ментальную модель сотрудника, уменьшает сдерживающие факторы роста ИК.

Существуют понятия текучести кадров “по собственному желанию” и “по инициативе организации”. С точки зрения ИК обе составляющие имеют негативное влияние. В Российской практике сложилось устойчивое понятие “текучести кадров”: показатель, фиксирующий уровень изменения состава вследствие увольнения и перехода на другую работу по личным мотивам. В понятие текучести обычно не включают переход сотрудника к другому работодателю через перевод, что сильно искажает российские результаты по сравнению с иностранными. В разных индустриях и сферах промышленности, а также на разных уровнях управления “нормой” считают различные значения текучести персонала (от 2-5 до 80%), что обусловлено особенностями бизнеса и категориями сотрудников. Так, например, для розничной торговли и массовой сферы обслуживания характерны самые высокие показатели, тогда как для тяжелой промышленности в целом нормальные достаточно низкие значения (5-10%). В целом, можно отметить, что уровень текучести повышается по мере выхода на работу более молодых поколений X, Y.

Важно так же отметить связь выгорания, усталости и текучести персонала, что имеет негативное воздействие на продуктивность организации. Положительной обратной связью обладают текучесть кадров и уменьшение произ-

водительности труда. Организации с высокой текучестью кадров обычно испытывают больше проблем с производительностью труда и с накоплением ИК.

Наиболее значимой составляющей ИК является продуктивность организации, отражающая отношение эффективного персонала к общему числу сотрудников.

Автор данного исследования построил модель ИК на основе продуктивности организации. Для этого была создана модель численности персонала. Модель численности обычно решает задачи прогнозирования численности персонала в зависимости от определенных драйверов численности, как правило, внешних (количество проектов, задач, клиентов, объектов обслуживания) на основе, текущей или заданной производительности труда. Основной проблемой моделей численности, разрабатываемых организациями, является линейные зависимости численности от драйверов и отсутствие учета фактора адаптации персонала (то есть перехода от новичков к опытным сотрудникам), а также применимость только в конкретной организации с ее драйверами/процессами.

Задачей данного эксперимента является рассмотрение поведения ИК в условиях нагрузки на персонал. Для оценки изменений ИК в условиях нагрузки была создана модель выполнения заданий. Обе модели в отдельности и общая модель ИК, построенная ни их взаимодействии описаны далее.

На рисунке (Рис.4.4) приведена когнитивная карта модели численности персонала, разработанная автором данного исследования по рекомендациям из [285] .

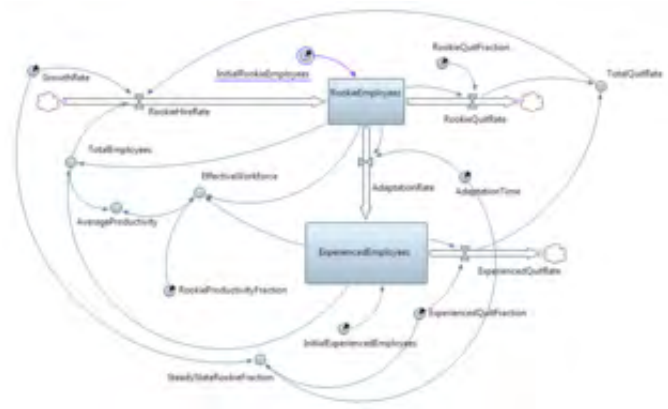


Рисунок 4.4 — Когнитивная карта модели численности персонала.

Модель численности персонала состоит из двух накопителей – Rookie Employees (Новички) и Experienced Employees (Опытные сотрудники) и четырех потоков:

– Набор новичков (RookieHireRate),
– Увольнение новичков (RookieQuitRate),
– Адаптация новичков в опытных сотрудников (AdaptationRate),
– Увольнение опытных сотрудников (ExperiencedQuitRate).

Свободные параметры модели численности персонала приведены в таблице 4:

Динамические переменные модели численности персонала перечислены в Таблице Таб. 5:

Таблица 4 — Свободные параметры модели численности персонала.

Название параметра	**Обозначение параметра**	**Значение параметра**
Скорость набора персонала	Growth Rate	0.01 в неделю (от Total Employees)
Начальное количество новичков в компании	Initial Rookie Employees	40 человек
Начальное количество опытных сотрудников в компании	Initial Experienced Employees	60 человек
Время адаптации новичка в опытного сотрудника	Adaptation Time	50-100 недель
Доля вклада новичка в продуктивность персонала	Rookie Productivity Fraction	30-80%
Доля увольнений новичков	Rookie Quit Rate	0.01
Доля увольнений опытных сотрудников	Experienced Quit Fraction	0.004

Потоки перечисленные выше вычисляются по формулам, приведенным в Таблице 6:

Для моделирования нагрузки модель численности персонала может быть дополнена процессами выполнения заданий. На Рис. 4.5 приведена когнитивная карта модели

выполнения заданий, разработанная авторами данного исследования.

Рисунок 4.5 — Когнитивная карта модели выполнения заданий.

Модель выполнения заданий состоит из двух накопителей: ServiceBacklog (очередь заданий) и StandardTimePerTask (Стандартное время на выполнение задания). Модель выполнения заданий управляется свободными параметрами, приведенными в Таблице 7:

Динамические переменные модели выполнения заданий приведены в Таблице 8:

Очередь заданий (ServiceBacklog) управляется экзогенной динамическим потоком поступления новых заданий (TaskArrivalRate) и потоком выполненных заданий

(TaskCompletionRate). Точка равновесия для модели выполнения заданий определена следующим уравнением (4.1):

$$EffectiveWorkforce = DesiredWorkforce \tag{4.1}$$

Модель численности персонала представляет основу для модели выполнения заданий. Вместе эти модели представляют динамику навыков и процессов, что характеризует интеллектуальный капитал организации, как мы уже отмечали ранее. Модели соединены динамическими переменными, приведенными в Таблице 9:

Для выявления поведения ИК исследуем кривые производительности. Кривая производительности является графическим представление изменения скорости обучения определённому виду деятельности. На рисунке 4.6 приведены кривые производительности для модели численности персонала при различных значениях времени адаптации новичков. В модели численности персонала динамической переменной, характеризующей производительности, является Средняя продуктивность персонала (Average Productivity).

Как мы видим из рисунка 6 для Доли вклада новичка равной 80% при временах адаптации более 50 недель кривая обучаемость стремиться к нижней асимптоте, а при временах более 60 недель стремиться к верхней асимптоте Средней продуктивности. Таким образом, демонстрируя разный характер поведения. На практике это означает, что

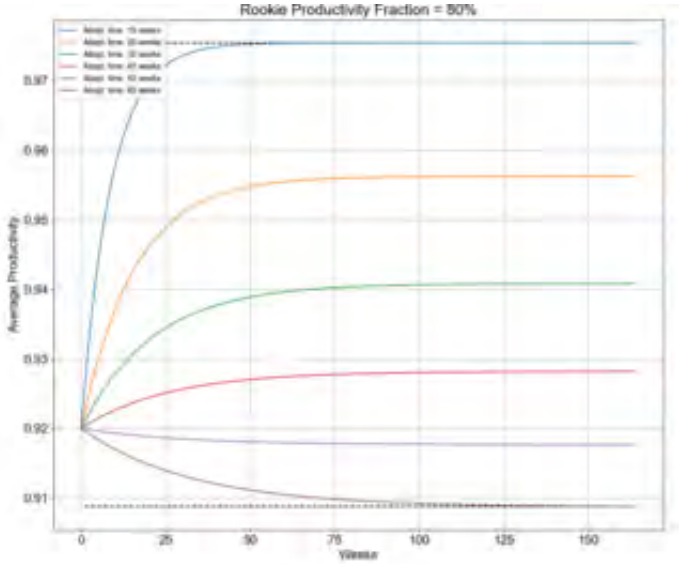


Рисунок 4.6 — Кривые производительности для различных значений Времени адаптации новичков.

при значительном времени адаптации новичков организационная производительность падает, так как количество опытных сотрудников в коллективе уменьшается по отношению к новичкам, а вклад в продуктивность от новичков меньше чем от опытных сотрудников.

С другой стороны, из кривых на рисунке 4.7 видно, что для Времени адаптации равному 20 неделям кривые продуктивности имеют единый характер и отличаются скоростью выхода на предельное значение – асимптоту.

Небольшие Доли вклада новичков означают, что сложность заданий не подразумевает участия в них неподготовленных сотрудников. С другой стороны, большие Доли

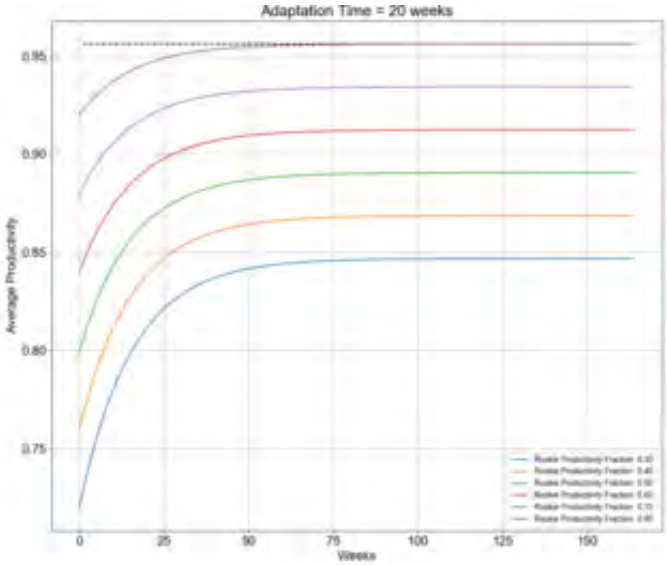


Рисунок 4.7 — Кривые производительности для различных значений Доли вклада новичков.

вклада новичков означают, что задания позволяют даже неопытному сотруднику работать с высокой отдачей, приближающейся к отдачи опытных сотрудников.

Для моделирований ИК под нагрузкой мы будем использовать экзогенную функцию для потока заданий (Рис. 4.8), для создания разных нагрузок.

Из рисунка 4.8 мы видим, что в пиковых нагрузках производительность падает, но за счет адаптации новичков организация, восстанавливает производительность, когда нагрузка спадает. Для различных времен адаптации в модели ИК кривые производительности будут иметь вид, представленный на рисунке (Рис. 4.9).

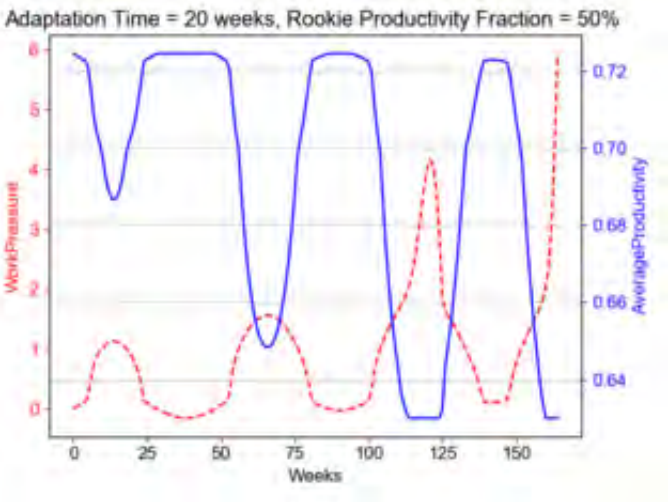


Рисунок 4.8 — Кривые производительности и нагрузки для модели ИК.

На рисунке (Рис.4.10) представлены кривые производительности для Времени адаптации равному 20 неделям с учетом нагрузки. Мы можем наблюдать различное поведение производительности до выхода на асимптоты при различных долях участия новичков, что отражает тот факт, что возможное включении новичков в решение заданий (до адаптации) характеризует эти задания, как достаточно простые и типовые.

Отметим, что при небольшом времени адаптации новичков (20 недель) и высокой доли участия новичков в продуктивности (80%) относительное падение продуктивности ниже, чем при невысокой высокой доли участия новичков в продуктивности (30%). Это наблюдение подтвер-

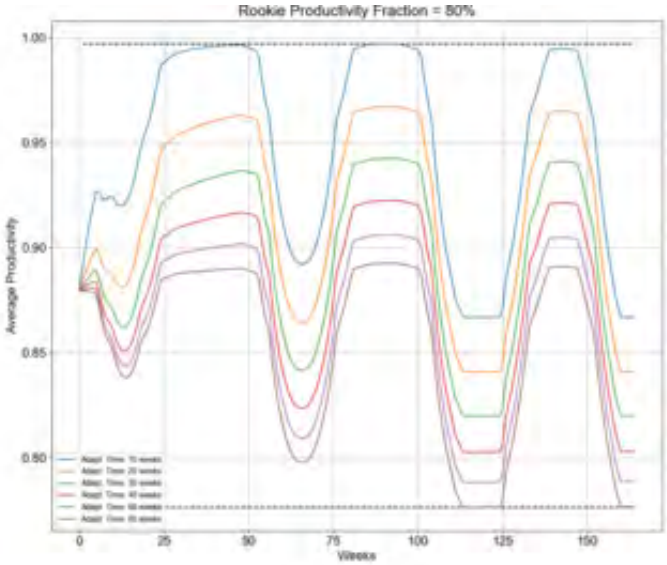


Рисунок 4.9 — Кривые производительности для различных значений Времени адаптации новичков с учетом нагрузки.

ждает тот факт, что при возрастании нагрузки коротких и простых заданий для новичков их продуктивность падает меньше, чем на сложных заданиях.

Для моделирования эффекта “выгорания” и “усталости” сотрудников в условиях длительной работы в режиме удлинённой недели в модель ИК введены следующие зависимости:

1. Эффект “выгорания” состоит в увеличении скорости увольнения опытных сотрудников в зависимости от времени работы в условиях удлинённой недели.

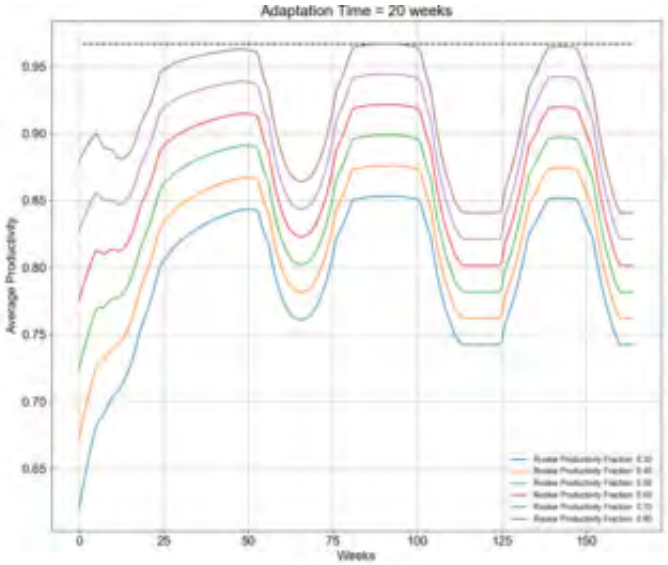


Рисунок 4.10 — Кривые производительности для различных значений Доли вклада новичков с учетом нагрузки.

2. Эффект “усталости” сотрудников состоит в уменьшении производительности сотрудников в зависимости от времени работы в условиях удлинённой недели.

На рисунке (Рис.4.11) изображен результат симуляции модели ИК для 500 недель. Такой длительный срок выбран с целью показать эффекты “выгорания” и “усталости” персонала и как следствие падение продуктивности вызванной работой в условиях удлинённой рабочей недели.

Падение производительности, вызванное длительной высокой нагрузкой, драматически влияет на ИК. В заключе-

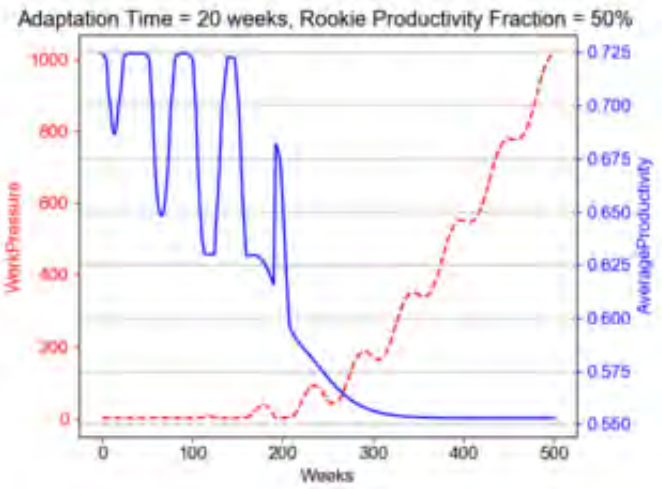


Рисунок 4.11 — Кривые производительности и нагрузки для модели ИК в условиях удлинненой рабочей недели.

ние на рисунке (Рис.4.12) представлены кривые изменения человеческого капитала – опытных сотрудников, новичков и общего числа сотрудников. Отдельно приведена кривая требуемого количества сотрудников для выполнения поступающих заданий. Мы видим, что количество новичков растет быстрее, чем количество опытных сотрудников.

Приведенные результаты эксперимента подтверждает теоретические работы по изучению процессов управления интеллектуальным капиталом. Новизна данного исследования состоит в выработке количественных оценок, помогающих уточнить стратегию управления интеллектуальным капиталом научно-исследовательской организации.

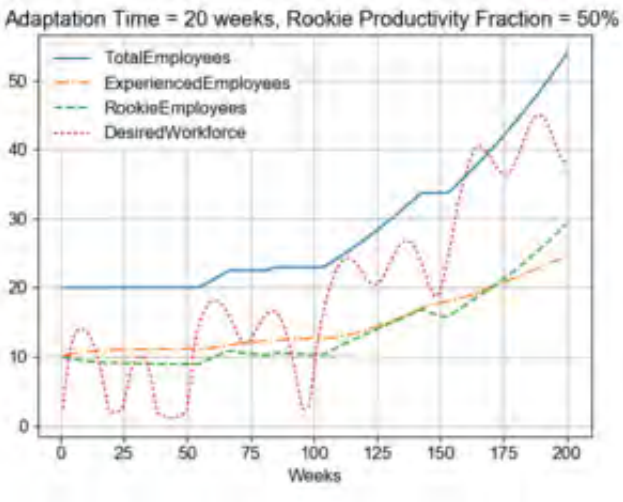


Рисунок 4.12 — Кривые изменения человеческого капитала по модели ИК.

Рассмотренная автором ситуация работы в условиях высокой нагрузки является типичной для российской экономики в современных условиях и особенно актуально в нефтегазовой отрасли.

4.4 Результаты моделирования командообразования в научной деятельности

Допустим, что в отраслевой научно-исследовательской организации Ω работают лаборатории λ_i , где $i \in (1 \dots N_\lambda)$

. Обозначим множество лабораторий $\Lambda = \{\lambda_i, \ldots, \lambda_{N_\Lambda}\}$. В лабораториях работают научные сотрудники $A = \{a_i, \ldots, a_{N_A}\}$.

Обозначим множество тематик t_i, где $i \in (1, \ldots, N_T)$, по которым организация Ω ведет НИР как $T = \{t_1, \ldots, t_{N_T}\}$. Тогда деятельность организации Ω по выполнению НИР может быть описана следующими компонентами (4.2):

$$\mathbb{M}_\Omega = \left\{S, \Xi, \Psi, E\right\}, \text{где} S = \{\Lambda, A, T, P, X\} \tag{4.2}$$

Помимо вышеопределенных компонент в уравнении 4.2 присутствуют:

- $\Xi = \{\xi_1, \ldots, \xi_{N_\Xi}\}$ – множество связей между субъектами (знакомство, соавторство и др.),
- $\Psi = \{\psi_1, \ldots, \psi_{N_\Psi}\}$ – множество действий субъектов (“поиск темы”, “отправка тезисов” и др.),
- $P = \{\rho_1, \ldots, \rho_{N_P}\}$ – множество научных работ,
- $X = \{\chi_1, \ldots, \chi_{N_X}\}$ – множество научных журналов и конференций.

Сотрудники организации Ω выполняют НИР по тематикам T , создают научные статьи и доклады P для публикации их в журналах и выступлениях на конференциях X. При создании научных статей P используются обзоры материалов журналов и конференций X. Конференции и редакции журналов X устанавливают приоритетные тематики T и принимают статьи для публикации по определенному графику (тезисы, полный текст, замечания рецензентов, выступление, публикация) и от наиболее

квалифицированных и опытных научных работников. Научный работник обладает квалификациями по тематикам T, которые можно представить в виде n-мерного вектора $(c_1, \ldots, c_{N_T})$ и опытом написания статей $(e_1, \ldots, e_{N_E})$, где c_i,$e_i \in \mathbb{R}$. И квалификации, и опыт не нуждаются в нормировке. Квалификация растет при успешном выполнении НИР, а опыт растет с успешной публикацией статей по соответствующей тематике. Имитационное моделирование представляет собой статистический эксперимент. Его результаты должны основываться на соответствующих статистических проверках. Автор выбрал метод повторения для вычисления доверительных интервалов и проверки гипотез. Таким образом, каждое наблюдение представляется независимым прогоном модели, в котором переходный период не учитывается. Далее производится вычисление средних величин выборки. Так как, прогоны независимы, то применяется стандартная формула для дисперсии. Преимуществом данного метода является то, что каждый имитационный прогон модели определяется своей последовательностью случайных чисел из интервала [0;1], что действительно обеспечивает статистическую независимость получаемых наблюдений. Недостатком является то, что все наблюдения могут оказаться под сильным влиянием начальных переходных условий.

В качестве калибровки для моделирования взят Газпромнефть НТЦ. В рамках НТЦ выбраны шесть тематик исследований $T = \{t_1, \ldots, t_{N_T}\}$, где $N_T = 6$:

1. Разработка и эксплуатация нефтяных месторождений
2. Геология и геологоразведочные работы
3. Информационные технологии
4. Техника и технология добычи нефти
5. Проектирование обустройства месторождений
6. Бурение скважин

В качестве издателя χ_1 выбрана редакция "Нефтяное хозяйство" выпускающая одноименный журнала с 1933 года. Авторы выбрали номер журнала за декабрь 2016 (НХ,12-2016), полностью состоящий из статей сотрудников "Газпромнефть НТЦ". В качестве конференции χ_2 выбрана конференция 16RPTC (SPE Russian Petroleum Technology Conference and Exhibition), прошедшая 24 октября 2016 года в Москве. Таким образом, $X = \{\chi_1, \ldots, \chi_{N_X}\}$, где $N_X = 2$.

В настоящее время в анализе социальных коллабораций выделяются два подхода:

– Структурный подход акцентирует внимание на геометрической форме сети и интенсивности взаимодействий (весе ребер). Для интерпретации результатов в данном случае используются структурные теории и теории сетевого обмена.
– Динамический подход акцентирует внимание на изменениях в сетевой структуре с течением времени

Целью эксперимента на данном этапе было подтвердить достаточность структуры компонент модели $\mathbb{M}_\Omega$ на примере научно-технического центра из нефтегазовой от-

расли. При наблюдении визуализации поведения агентов у авторов не возникло необходимости в добавлении новых компонентов в модель.

Для поставленных условий на основании $\mathbb{M}_\Omega$ была создана частная модель $\mathbb{M}_{GPN}$ и проведена много прогонная симуляция модели (Рис. 4.13).

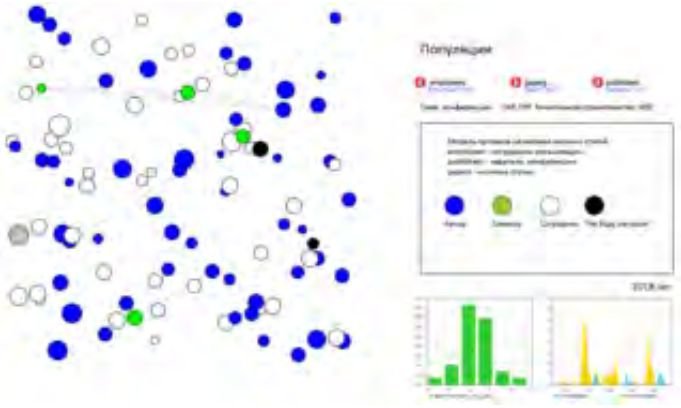


Рисунок 4.13 — Фрагмент визуализации прогона симуляции частной модели для НТЦ из нефтегазовой отрасли.

Далее на основании симуляций была создана база данных для последующего исследования процессов. База данных содержит следующие основные сущности (Рис. 4.14).

По результатам имитационного моделирования нами были получены следующие результаты для процесса создания и публикации научной статьи:

1. Среднее время написания статьи: 20 ± 2 недель.

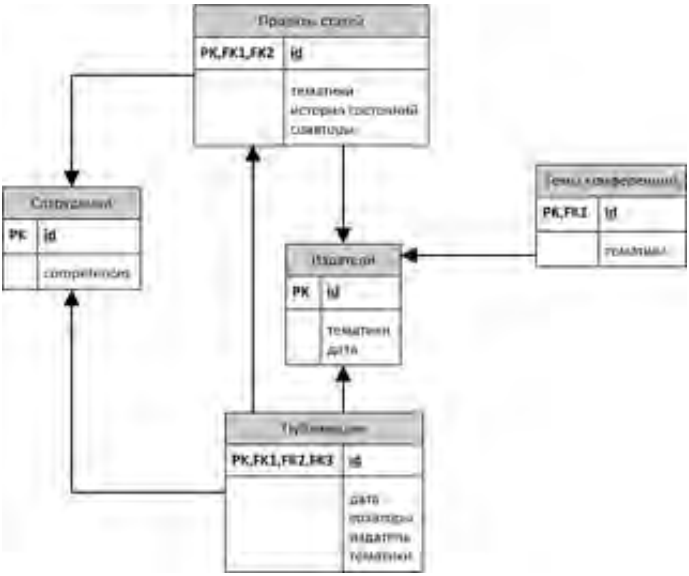


Рисунок 4.14 — Entity relationship diagram (ERD): Сущности процесса создания научных статей.

2. Среднее число соавторов: 3.5 ± 1.0
3. Максимальное число соавторов: 7 ± 3
4. Среднее количество статей на одного автора в год: 2 ± 0.5
5. Доля статей, не уложившихся в график публикации: 40 ± 10 %.

Данные результаты находятся в согласии с опытом автора, но нуждаются в дальнейшей проверке. Для оценки корректности результатов моделирования было проведено сопутствующее библиометрическое исследование реальных эмпирических данных по собственной методике автора [286]. В соответствии с поставленными условиями эксперимента

была создана база публикаций, содержащая следующие информационные поля:

1. Дата публикации статьи
2. Список авторов
3. Название статьи
4. Тематика статьи согласно классификатору тематик T
5. Издатель согласно классификатору X

В базу публикаций были собраны статьи издательства “Нефтяное хозяйство” и публикации из электронной библиотеки сообщества нефтегазовых инженеров SPE OnePetro, сделанные сотрудниками Газпромнефть НТЦ. Проведен анализ публикаций и построен граф соавторств (Рис.4.15).

Рисунок 4.15 — Синтетический граф соавторств для НТЦ "Газпромнефть".

На основании базы публикаций вычислены следующие параметры соавторства:

1. Среднее количество опубликованных статей на одного автора в год: 2 ± 0.5
2. Среднее количество соавторов: 2.8 ± 0.1
3. Максимальное количество соавторов: 10 ± 1

Полученные эмпирические результаты подтверждают результаты имитационного моделирования, что свидетельствует о перспективах применения имитационного моделирования как аналога для моделирования социальных процессов в организационной среде.

4.4.1 Настройка свободных параметров тематической модели на предметную область

Несмотря на то, что тематические модели используются для построения кластеров документов уже более 10 лет до сих пор существует проблема выбора оптимального количества тематик. Автор проанализировал ряд ключевых исследований, предпринятых на эту тему за последнее время [287]. Основная проблема состоит в отсутствии стабильной метрики качества тематик, полученных в ходе построения тематической модели. Автором проведён анализ внутренних метрик тематической модели: Когерентность, Контрастность и Чистота ядра тем для определения оп-

тимального количества тем и сделано заключение об их неприменимости для решения этой задачи.

Далее проанализирован подход к выбору оптимального количества тем на основе полученных кластеров. Для этого были рассмотрены поведения метрик валидации кластеров Davies Bouldin Index, Silhouette Coefficient и Calinski-Harabaz Index в зависимости от количества тематик. В основу предлагаемой автором новой методики определения оптимального количества тематик легли следующие принципы:

- Настройка тематической модели с последовательной регуляризацией (ARTM) для отделения шумовых тематик;
- В качестве векторного представления слов, входящих в тематики, авторы предложили использовать плотных представления (embedded) векторов (GloVe, FastText, Word2Vec);
- Для оценки расстояний авторы предложили использовать косинусную меру, которая на векторах с большой размерностью работает лучше, чем Евклидова мера расстояния.

Разработанная автором методика получения оптимального количества тем была опробована на коллекции научных статей из библиотеки OnePetro, отобранным определенным тематикам рубрикатора. Эксперимент показал, что предложенная авторами методика позволяет точно оценить оптимальное количество тематик для тематической

модели, построенной по небольшой коллекции англоязычных документов.

4.5 Результаты оптимизации процессов научной деятельности

Задача поиска оптимальных параметров команды соавторов для наиболее продуктивного написания научных статей относится к классу задач оптимизации. Функция, которую необходимо минимизировать будет зависеть от следующих параметров:

- Количество сотрудников в организационной среде (N_o);
- Скорость появления новых сотрудников ($Vemp_{new}$);
- Скорость увольнения сотрудников ($Vemp_{fired}$);
- Максимальное количество компетенций у сотрудника ($Cmax_{emp}$);
- Максимальное количество компетенций необходимых для достижения цели исследования ($Cmax_{pub}$).

Показателями производительности процесса написания статей, оптимальные значения которых необходимо найти, могут быть следующие:

- Время написания научной статьи (T_{pub})
- Доля сотрудников, опубликовавших статьи от всего количества сотрудников ($Frac_{pub}$)

– Доля несостоявшихся статей ($Frac_{notpub}$).

Параметрами организационной среды будут следующие:

– Минимальное и максимальное количество сотрудников в организации (N_{omax},N_{omin})
– Скорость появления потенциальных целей исследований (V_{pub})
– Временные ограничения на написание статьи (T_{eoc})
– Скорость встреч для заведения знакомств между сотрудниками ($V_{friending}$)
– Скорость встреч участников с потенциальными целями (V_{go})

Исходя из вышеописанных параметров фитнесс-функция $\mathcal{F}$ для оптимизации может быть записана в следующем виде (4.3):

$$\mathcal{F}\left\{\frac{1}{Frac_{pub}}, T_{pub}, Frac_{notpub}\right\} \rightarrow min \tag{4.3}$$

При выполнении системы основных условий:

$$\begin{cases} N_o \in [N_{omin}, N_{omax}] \\ Cmax_{emp} \leqslant Cmax_{pub} \in [1, N_{comp}] \\ Vemp_{new} \geqslant Vemp_{fired} \geqslant 0 \end{cases} \tag{4.4}$$

Оптимизационный эксперимент был проведен в среде AnyLogic для моделей, с применением Scrum и без [288]. Графы соавторств с применением Scrum не изменились. На основании оптимизационного эксперимента

была произведена калибровка имитационной модели соавторства разработанной автором данного исследования. Были найдены оптимальные параметры N_o,$Vemp_{new}$, $Vemp_{fired}$,$Cmax_{emp}$,$Cmax_{pub}$ для Газпромнефть НТЦ. Оптимальные значения параметров приведены в Таблице 10.

Калиброванная модель стала основой для исследования эффекта от введения Scrum ролей в процесс написания научных статей. Для выбранных в разделе показателей производительности T_{pub} и $Frac_{notpub}$ была проведена много прогонная симуляция двух типов: с использованием методики Scrum и без Scrum. Анализ данных был произведен в статистической среде R.

Результаты попрогонного изменения T_{pub} и $Frac_{notpub}$ приведены на рисунках 4.16 и 4.17 соответственно.

Для оценки влияния Scrum на время написания статей T_{pub}, автор сравнил время написания статей для двух выборок методом t-теста для сравнения двух независимых выборок. Результаты показали, что на уровне 1% значимости длительность написания статей с использованием Scrum не изменяется.

- Среднее время написания научной статьи со Scrum составило 19.90 недель со стандартным отклонением 3.33 недели.
- Среднее время написания научной статьи без Scrum составило 19.90 недель со стандартным отклонением 0.77 недели

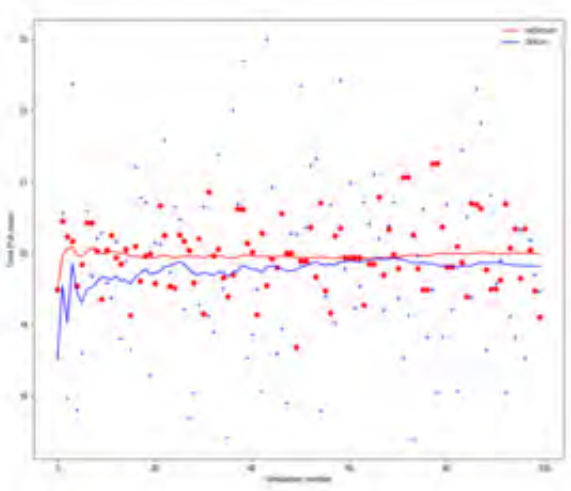


Рисунок 4.16 — Среднее время публикации статей в зависимости от номера прогона. Линиями нарисована зависимость скользящего среднего.

Автор также дополнительно использовал непараметрический критерий U Манна-Уитни в случае, при котором распределении признаков не соответствует нормальному распределению, результаты которого оказались аналогичны t-тесту. Полученные результаты свидетельствуют о том, что использование Scrum не ускоряет написание статей, даже при условии того, что функция написания статей не подчиняется нормальному распределению.

Другим показателем, который может быть использован для оценки продуктивности Scrum, является доля *несостоявшихся научных статей*. Мы оценили долю *несо-*

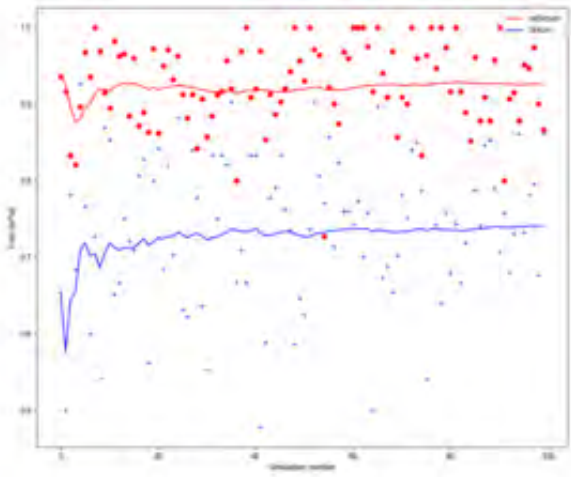


Рисунок 4.17 — Доля несостоявшихся научных статей в зависимости от номера прогона. Линиями нарисована зависимость скользящего среднего.

стоявшихся научных статей для команд, использующих Scrum и не использующих.

– Доля *несостоявшихся научных статей* в командах, использующих Scrum 0.74 со стандартным отклонением 0.02
– Доля *несостоявшихся научных статей* в командах, не использующих Scrum составляет 0.92 со стандартным отклонением 0.01

Другими словами, из 100% начатых статей в командах, использующих Scrum, успешными будут 26% статей. В случае, если Scrum не используется, во временные рамки

публикационного процесса с требуемым качеством уложатся 8% статей.

4.6 Прогнозирование соавторства

Коллективное соавторство в написании научных статей имеет детерминированную и случайную структурные составляющие. Кроме рациональных аспектов при образовании коллектива соавторов отдельной научной статьи существуют и эмоциональные составляющие. Во временной перспективе складываются и распадаются рабочие группы исследователей, обновляется трудовой коллектив и состав подрядчиков, которые участвуют в совместных отраслевых коллаборациях для проведения исследований.

Несмотря на всю сложность соавторства, существуют несколько классов моделей для симуляции образования соавторства. В их числе модели на основании случайных графов и модели образования соавторств на основе компетенций соавторов. Оба математических аппарата разработаны и применяются в течении нескольких десятков лет по отдельности. Но практических применений моделей соавторств в корпоративной практике не так много.

Автор выдвинул гипотезу о том, что необходимо объединить несколько различных типов моделей для того,

чтобы лучше понять природу научных коллабораций в отдельной организации.

Автор данного исследования поставил задачу разработать методику построения модели соавторства для научно-технического центра, учитывающую различные структурные составляющие соавторства.

В результате автор разработал модель с использованием методов машинного обучения, случайных графов и модели компетенций. На основании разработанной модели сделан прогноз развития соавторства в написании научных статей научно-технического центра Газпромнефть [289].

Практическая ценность результатов данного исследования состоит в следующем: Количественно оценен вклад различных структурных составляющих в формировании соавторств при написании научных статей.

Прогнозирование развития соавторства в написании научных статей позволяет осуществить планирование корпоративных ресурсов для поддержания роста научных публикаций. Понимание кластерной структуры соавторства позволяет производить выравнивание направлений научной деятельности в соответствии со стратегическим планом развития научно-технического центра [290].

Измерение деятельности научно-исследовательских организаций на основании графа соавторства является хорошо зарекомендовавшей себя практикой. Исследователи показывают возможности выявления наиболее

производительных авторов (“Highly Productive Authors”) и влиятельных авторов (“Influential Authors”).

В качестве объекта исследования была выбрана публикационная активность НТЦ Газпромнефть. Данные были получены из открытой электронной библиотеки OnePetro международного сообщества нефтегазовых инженеров (SPE). После очистки было получено 172 статьи. Распределение авторов по годам отображено на рисунке (Рис. 4.18).

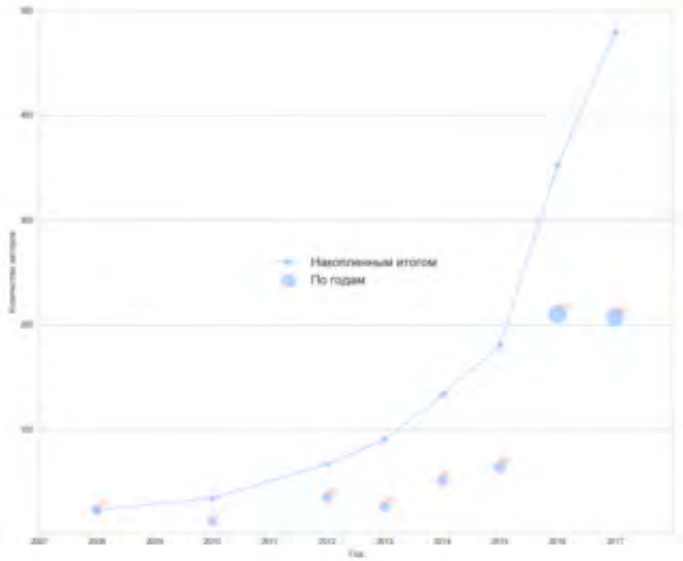


Рисунок 4.18 — Распределение авторов по годам.

Прямолинейным ответом на поставленный исследовательский вопрос может быть интерполяция кривой роста количества авторов. В результате такой оценки получим следующую зависимость $y = 13.3816e^{0}.3422x$ с достоверностью $R^2 = 0.98$, дающей прогноз 585 авторов в 2018 году. Но из

графика 4.18 мы так же видим, что количество авторов в 2017 году (207) меньше чем в 2016 году (210), что может оказаться насыщением роста и повлиять на прогноз.

Построим прогноз на основании графа соавторства [291]. Для этого построим двудольный граф соавторства с вершинами: автор (479) и статья (171). Авторы обладают техническими компетенциями, статьи характеризуются названием, годом издания и ключевыми словами.

Полученный граф соавторства имеет 26 связанных компонент наибольшая из которых содержит 556 вершин, а остальные – не более 8. Малые связанные компоненты относятся к авторам, написавшим свою первую статью. Наличие малых связанных компонент можно рассматривать, как одну из составляющих роста графа соавторств. В таблице (Таб. 11) приведены количества и размеры связанных компонент за каждый год нарастающим итогом.

Таким образом мы видим, что граф соавторства прогрессирует в сегменте малых связанных компонент по количеству и вместе с тем граф становится более связанным – увеличивается количество узлов в главной связанной компоненте. Для уточнения прогнозирования целесообразно будет учесть такое строение. На Рис. 4.19 приведена инкрементальная динамика прироста графа соавторства по годам. Уточним, что граф соавторств 2017 года является суммой всех изображенных на (Рис. 4.19).

Из рисунка (Рис. 4.19) можно сделать качественной вывод об увеличении ежегодно прибавляемых к графу со-

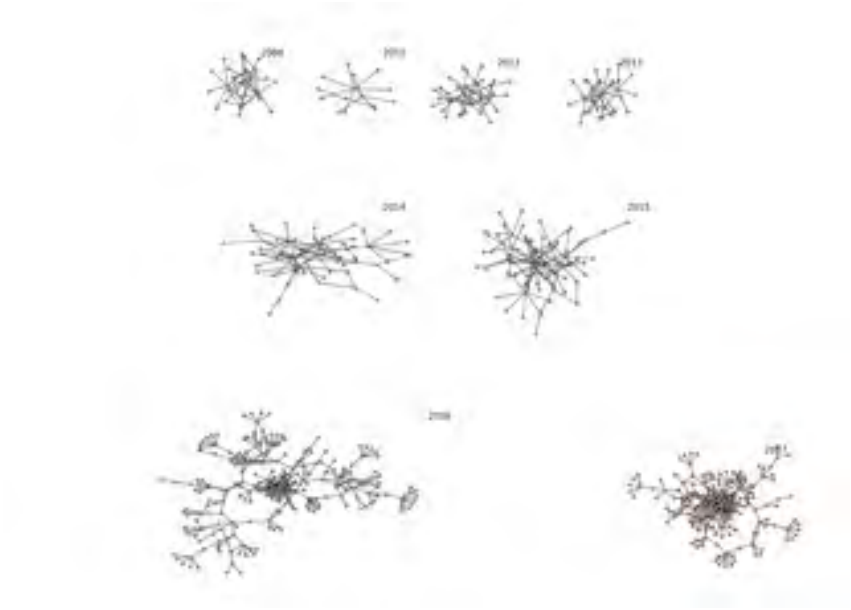


Рисунок 4.19 — Динамика прироста развития графа соавторств по годам.

авторства связей. Изменение роста ведет к усложнению структуры графа соавторства в 2016 году, что можно констатировать как “эффект локтя”.Для прогнозирования авторства будем использовать следующие метрики вершин графа:

- Degree centrality
- Betweenness centrality
- Closeness centrality
- Harmonic centrality
- Clustering

Распределения метрик вершин графа соавторства приведены на рисунке (Рис.4.20).

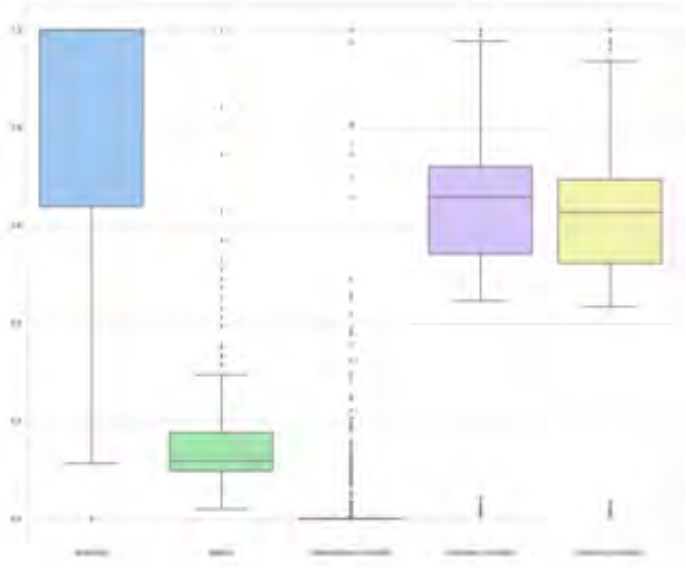


Рисунок 4.20 — Метрики вершин графа соавторства.

Для прогнозирования авторства будем использовать модель бинарной классификации. Выбор модели будем производить на основе ROC-кривой. Обучения модели будем производить на метриках 2016 года. Параметры моделей оптимизированы с помощью кросс-валидации с 5-кратным фолдингом. В результате сравнения различных классификаторов были получены следующие результаты (Таб. 12).

Лучшее значение метрики ROC AUC показал классификатор на основе нейронной сети (Multi-layer perceptron) с

одним слоем из 10 персептронов. Отчет о выполнении прогноза авторства в 2018 году на основе графа соавторства 2017 года приведен в таблице (Таб. 13).

В результате прогнозирования определено 406 авторов в 2018 году. Если добавить к этому прогнозу сотрудников, которые напишут свою первую статью в 2018 году, то на основании оценки по динамике роста связанных компонент получим прибавку в 15%. Итого в 2018 году авторами станут 467 сотрудников.

4.6.1 Распределение научных направлений на основе соавторств

Планирование научно-технического развития исследовательской организации должно быть увязано с реальным положением дел. Такие факторы, как организационная инертность, диверсификация исследований и увлечение созданием ИТ-продуктов, могут существенно исказить любые стратегии и планы развития. Тем не менее, исполнимость планов является важной характеристикой развития, существенно повышающей мотивацию персонала к достижению результата. Поэтому постановка реально, а не только “на бумаге” выполнимых задач, необходима.

Не может быть достаточно только количественных инструментов оценки выполнения научно-исследовательских

работ. Формально-бумажная отчетность по НИР не способна отразить увлеченность и вовлеченность исследователей в работу [292]. В то время как малые формы исследовательских работ, такие как презентация на научно-технической конференции или научная статья в рецензируемом периодическом издании, требуют намного более неформального отношения со стороны исследователей. Анализ развития научно-технической организации на основе публикационной активности является распространенной практикой. Многие исследования анализируют корпусы текстов научных статей и делают заключения о трендах развития. Текстовые данные обладают высоким уровнем шума и даже современные методы анализа на основании word embedding выдают точные прогнозы только на основании огромных объемов текстов, которые не всегда имеются у небольших организаций. При этом именно небольшие научно-исследовательские организации в наибольшей степени страдают от неточности планирования научной деятельности.

Современный фокус применения научных подходов к управленческим решениям приобретает все большую актуальность. С увеличением объемов данных традиционные аналитические средства руководителей организаций становятся все менее эффективными. С другой стороны, необходимого объема данных для устойчивой работы современных алгоритмов часто бывает недостаточно. В авангарде этой тенденции возникает задача адаптации и создания новых

эвристик для таких классических задач, как кластеризация для применения в организационной среде.

Кластеризация данных на основании статической модели получила развитие с открытием таких алгоритмов, как PAM, CLARANS, DBSCAN, CURE и ROCK. Тем не менее, в последнее время особое внимание привлечено к алгоритмам кластеризации на основании динамической модели, например, CHAMELEON. Основная идея алгоритма CHAMELEON заключается в использовании метрик близости графа, построенного на основании набора кластеризуемых данных с помощью метода “k наиболее близких соседей” (KNN). Метрики графа оказываются более эффективными для разбиения данных “сверху вниз” в случае сложных объектов.

Разнообразие алгоритмов кластеризации не умаляет важность задачи оценки их качества. Но в условиях ограниченного количества данных и для обеспечения управленческих решений качество кластеризации должно иметь не только математически обоснованную, но и уверенную наглядную составляющую. Другими словами, чтобы “с одного взгляда было понятно” и не нужно было вникать в формулы. Таковы требования современного бизнеса.

С формальной точки зрения необходимо решить задачу обучения без учителя (unsupervised machine learning) для графа соавторств, отнести кластеры к определенным тематикам и выявить изменения в кластерах со временем.

Кластеризация графа соавторства может быть осуществлена на основании различных метрик вершин:

- Degree centrality
- Betweenness centrality
- Closeness centrality
- Harmonic centrality
- Clustering

Рассмотрим содержательный смысл метрики Betweenness centrality применительно к задаче кластеризации графа соавторств научно-технической организации. Метрика Betweenness centrality характеризует то, насколько данный узел важен для связанности графа. Связи в графе соавторств отражают совместную исследовательскую работу. Графы соавторств не всегда являются связанными, обычно это несколько связанных компонент разного размера.

Связанные компоненты являются естественными кластерами. Небольшие связанные компоненты отражают начальные инициативы – это первые статьи сотрудников. Но главная связанная компонента может содержать 90% вершин графа соавторства и нуждается в отдельном подходе к кластеризации.

Для выделения кластеров из главной связанной компоненты графа соавторств возможно использовать методику искусственного удаления вершин с наибольшей метрикой Betweenness centrality. При каждом таком удалении вершины граф может распадаться на несколько несвязанных

компонент. На рисунке (Рис.4.21) приведена модель такого разделения.

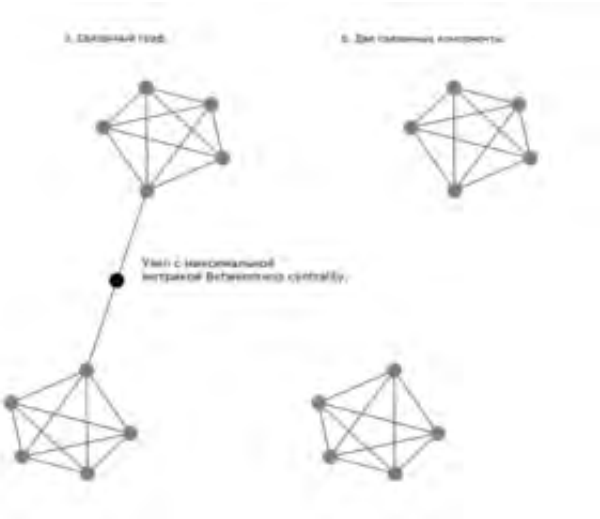


Рисунок 4.21 — Модель разделения графа. а.Связанный изначальный граф. б.Тот же граф, но после удаления вершины с наибольшей метрикой Betweenness centrality уже представляет две связанных компоненты.

Каждую из получившихся при таком распаде связанных компонент можно анализировать на однородность тематики на основании текстов статей, которыми она образована. В результате нескольких итераций мы получим набор кластеров. Предложенный автором метод является эвристическим и нуждается в проверке по определенному формальному критерию. Для задач кластеризации таким критерием принято считать метрики близости объектов в

кластере и расстояния между объектами в разных кластерах.

Сходимость данного метода обеспечивается путем поиска минимума функционала ошибок определения кластеров:

$$|WSS - BSS| \to min, \tag{4.5}$$

где WSS - это функция связанности кластера C_i:

$$WSS = \sum_i \sum_{x \in C_i} (x - m_i)^2, \tag{4.6}$$

а BSS это функция разделения кластеров C_i:

$$BSS = \sum_i |C_i|(m - m_i)^2, \tag{4.7}$$

где $|C_i|$ - это размер кластера C_i.

Междисциплинарные исследования приводят к тому, что статьи будут относится к нескольким тематикам, так что полученные кластера будут пересекающимися – не эксклюзивными. Таккая кластеризания называется"мягкой".

В качестве объекта исследования была выбрана публикационная активность НТЦ “Газпромнефть”. Данные были получены из открытой электронной библиотеки OnePetro международного сообщества нефтегазовых инженеров (SPE). После очистки было получено 172 статьи. Построим прогноз на основании графа соавторства. Для этого построим двудольный граф соавторства с вершинами:

автор (479) и статья (171). Авторы обладают техническими компетенциями, статьи характеризуются названием, годом издания и ключевыми словами. Полученный граф соавторства имеет 26 связанных компонент наибольшая из которых содержит 556 вершин, а остальные не более 8. Связанные компоненты с количеством узлов до 8 являются считать первыми статьями сотрудников. Рассмотрим наибольшую связанную компоненту (556 вершин). Выделим подграф из основного графа соавторства на основании узлов, относящихся к наибольшей связанной компоненте. Получившийся подграф отображен на рисунке (Рис. 4.22).

Рисунок 4.22 — Подграф наибольшей связанной компоненты графа соавторства.

Рассчитаем для полученного подграфа метрику Betweenness centrality. Полученные значения Betweenness centrality отображены на рисунке (Рис. 4.23). Нулевые значения Betweenness centrality не отображены.

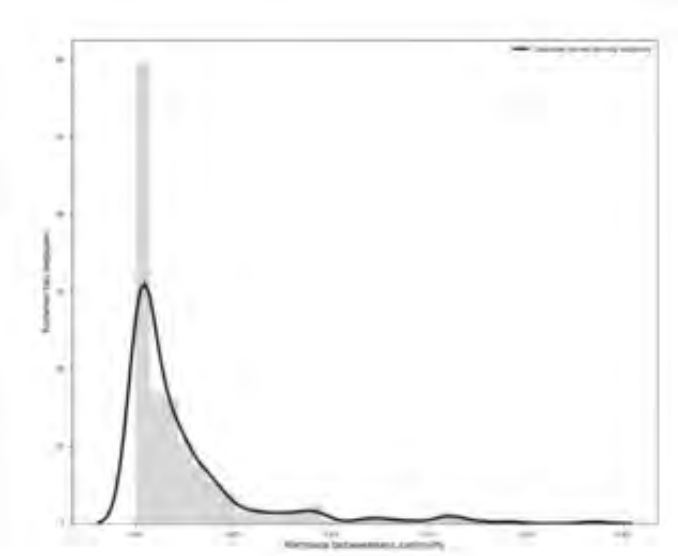


Рисунок 4.23 — Гистограмма значений Betweenness centrality для подграфа наибольшей связанной компоненты графа соавторств.

Как мы видим из Рис. 4.23 значения метрики Betweenness centrality в третьем квартиле принадлежат всего 23 вершинам, что составляет менее 5 % от всего количества вершин. Применим алгоритм искусственного удаления вершин с наибольшим значением метрики Betweenness centrality. На рисунке (Рис. 4.24) отображена

зависимость количества связанных компонент от количества искусственно удаленных вершин

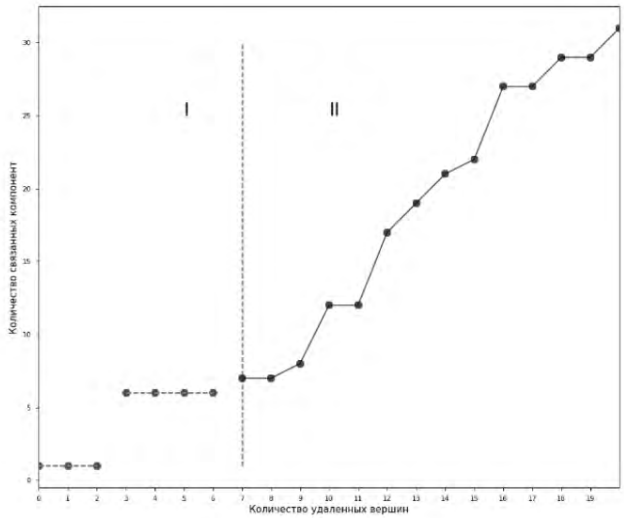


Рисунок 4.24 — Зависимость количества связанных компонент от количества искусственно удаленных вершин.

При удалении вершин поведение графа происходит в двух режимах:

1. Удержание связанности
2. Экспоненциальный распад

Отличительной чертой режима I является то, что граф остается связанным при удалении вершин с высокими значениями метрики Betweenness centrality. Это означает, что удаляемые вершины не являются единственными связующими между кластерами. Отличительной особенностью режима II является следование степенной модели распада

графа, когда удаление каждого узла вызывает степенной рост появления новых связанных компонент. Рассмотрим более подробно вторую половину режима I алгоритма, когда граф разделился на 6 связанных компонент. Размеры этих компонент составляют 511, 34, 1, 1, 1, 1. И среди них ярко выраженное направление исследований по Теме 1 представлено именно компонентой с 34 узлами, представленной на рисунке (Рис. 4.25).

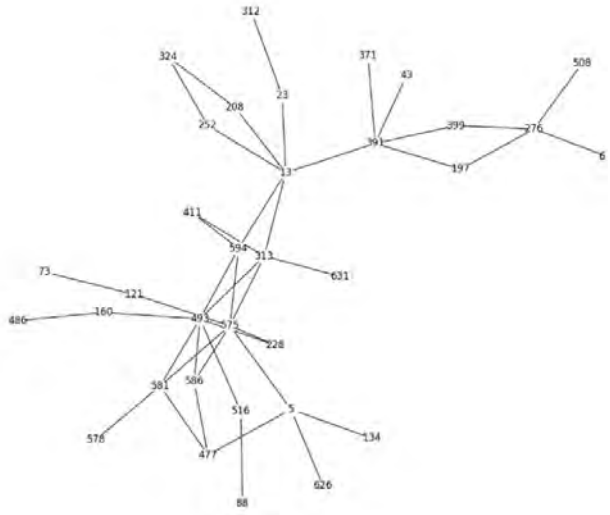


Рисунок 4.25 — Кластер исследователей по Теме 1, выделенный в результате применения метода удаления вершин в наибольшим значением метрики Betweenness centrality.

Мы рассмотрели выделение одного кластера подробно. Полный алгоритм выделения кластеров будет состоять из следующих шагов:

1. Построение двудольного графа соавторств: G
2. Определение метрики Betweenness centrality для графа G
3. Определение вершины с максимальной метрикой $N_{max(Betweennesscentrality)}$
4. Удаление вершины $N_{max(Betweennesscentrality)}$ из графа G
5. Получение списка связанных компонент графа G
6. Вычисление метрики качества полученных кластеров BSS и WSS
7. Далее алгоритм повторяется для каждой связанной компоненты
8. Алгоритм завершается, когда все связанные компоненты представляют кластеры удовлетворительного качества.

Для выбранного графа соавторства были выделены 16 кластеров. Для вычисления значений и W на основании текстов статей было использовано векторное представление текста статьи (VSM). Каждая статья представлена в виде вектора со значениями метрики BM25 для каждого слова. Статьи рассматривались как “мешок слов” (bag of words). Для измерения дистанции между векторными представлениями статей была использована косинусная мера. На

рисунке (Рис. 4.26) изображена матрица раздельности кластеров .

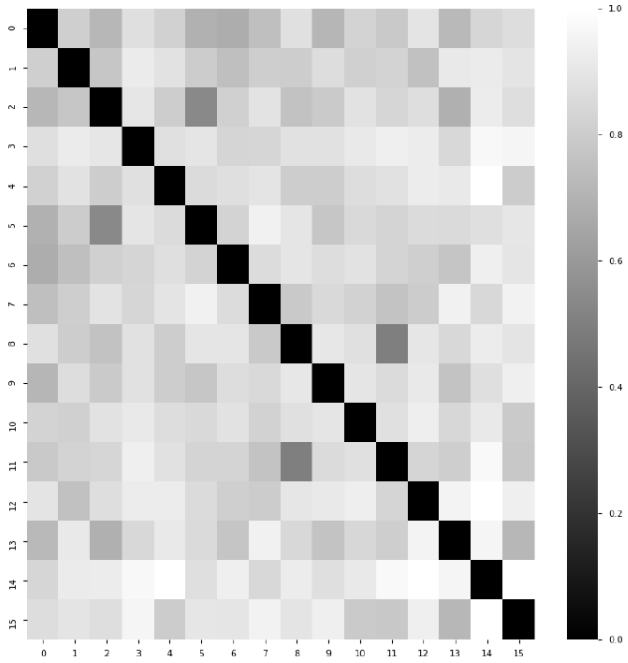


Рисунок 4.26 — Матрица раздельности кластеров. По осям отображены номера кластеров. В ячейках значения функции BSS.

Для сравнения полученной кластеризации статей была проведена кластеризация с помощью алгоритма KMeans, показавшая схожие результаты (Рис. 4.27).

С помощью KMeans была произведена кластеризация статьей, а затем из графа соавторства были определены кластеры авторов на основании полученных кластеров статей.

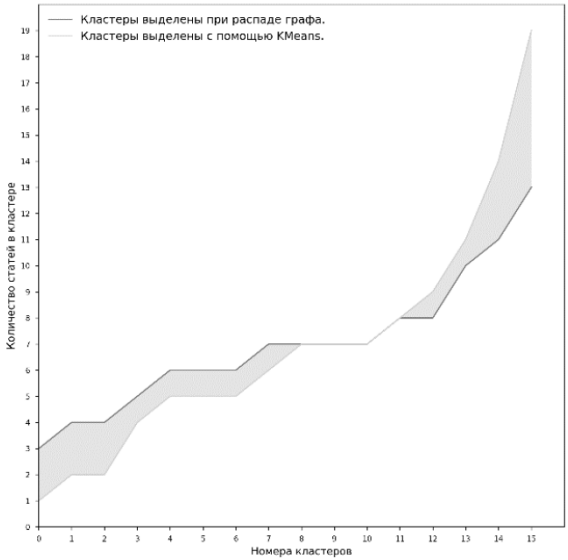


Рисунок 4.27 — Сравнение предлагаемого в статье алгоритма кластеризации и алгоритма KMeans.

4.7 Вероятностная модель скрытых тем на основе архива журнала “Нефтяное Хозяйство”

Вопрос о том, по какому пути движется прикладная наука и технологии, является ключевым для любой научно-технической области. Традиционно определение векторов развития производилось и производится экспертами по предметному направлению, однако значительный рост объе-

мов информации и увеличение числа направлений развития свидетельствуют о необходимости доработки и совершенствования этого инструментария и выработке дополнительных методов исследования индустриальных трендов. В данном исследовании автором проанализированы тренды нефтяной индустрии посредством автоматизированной обработки текстов научных статей в отраслевом журнале “Нефтяное хозяйство”. Выявляя наиболее часто встречаемые темы в журнале за период с 2008 по 2016 годы, мы сделали вывод об увеличении значимости трудноизвлекаемых запасов и росте интереса к методам разработки подобных месторождений.

Журнал “Нефтяное хозяйство” посвящен нефтегазовой проблематике. В нем публикуются статьи, посвященные широкому кругу вопросов нефтегазового сектора: экономических, технических, технологических, экологических и информационных. Издание насчитывает почти вековую историю и выходит каждый месяц с 1920 года. Все публикуемые статьи проходят процедуру рецензирования. Журнал включен в Российский индекс научного цитирования (РИНЦ) и международную систему индексирования Scopus. Материалы журнала находятся в закрытом доступе и распространяются по подписке. Основные рубрики журнала:

– новости нефтегазовых компаний;
– нефтяная и газовая промышленность;
– экономика, управление, право;

– геология и геологоразведочные работы;
– бурение;
– разработка и эксплуатация нефтяных месторождений;
– проектирование и обустройство месторождений;
– техника и технологии добычи нефти;
– нефтепромысловое оборудование;
– транспорт и подготовка нефти;
– экологическая и промышленная безопасность;
– информационные технологии.

Как видно, журнал подробно рассматривает практически все аспекты функционирования нефтяных компаний – от экономико-правовых вопросов, до технологических аспектов и тонкостей.

Для проведения исследования редакцией были любезно предоставлены архивы статей журнала “Нефтяное хозяйство” за период 2008-2016 гг. В выборке содержится 108 выпусков журналов, со статьями от 3517 авторов. В каждом из выпусков журналов содержатся все статьи номера, таким образом, была получена сплошная выборка, в которой содержались материалы по самым различным содержательным направлениям. В среднем, в номере журнала “Нефтяного хозяйства” около 20-25 статей. Были рассмотрены именно выпуски журнала, так как они являлись единицей анализа. Авторами статей журнала являются научные сотрудники, инженеры и отраслевые эксперты, многие из них кандидаты и доктора наук.

Процесс исследования имел следующие этапы:

1. Изначально архивы представлены в виде файлов в формате PDF. Иногда это был единый файл (биндер) со статьями за весь год, а иногда разрозненные файлы с отдельными статьями. В обоих случаях файлы были предназначены для печати, то есть содержали оглавления, номера страниц, тематические вставки и другие редакторские элементы. Для анализа нужны были только тексты в виде предложений поэтому автором был реализован программный модуль для приведения всех данных к такому формату. Отметим, что, исходя из выбранной методики, важно было сохранить порядок слов и разделение на предложения, при этом необходимо было сохранять принадлежность к выпуску, а не к статье, так как минимальной единицей временного анализа выбран один выпуск.
2. На втором этапе анализа происходило приведение слов к основным формам. Для анализа и сравнения слов методами частотного и вероятностного анализа необходимо сузить возможные варианты употребления словоформ. Существуют несколько алгоритмов для решения этой задачи (нормализации текста), в данном случае был использована стемминг. Стеммингом называют процедуру нахождения основы слова, при этом основа и корень слова могут различаться между собой.

Одним из наиболее распространенных инструментов является стеммер Портера, который, однако, часто обрезает слово больше необходимого, что затрудняет получение правильной основы слова, например, кровать->крова. Также стеммер Портера не справляется со всевозможными изменениями корня слова (например, выпадающие и беглые гласные), характерными для русского языка. Поэтому автор остановился на использовании технологии стемминга компании “Яндекс” - MyStem. Данная программа производит морфологический анализ текста на русском языке. Она умеет строить гипотетические разборы для слов, не входящих в словарь и предлагает несколько вариантов основ слова. Тем не менее, автор сочел необходимым поддерживать обратный словарь для полученных словоформ, чтобы сохранять связь между изначальным словом и полученной словоформой. Отдельной веткой обработки подвергались аббревиатуры, широко распространенные в нефтегазовой отрасли. Определение аббревиатур производилось на основе словаря аббревиатур, созданного и поддерживаемого в компании ГазпромНефть в рамках проекта Корпоративной Википедии [1].

3. На третьем этапе исследования проводилось формирование словаря. Известно, что наибольшую смысловую нагрузку несут не одиночные слова,

а сочетания слов, в частности пары слов – биграммы. Для выделения биграммов автором был использован эвристические алгоритмы. Была составлена матрица биграммов в окрестности 5 слов для каждого из предложений. Затем были рассчитаны частоты использования каждого из биграммов, после чего были зафиксированы 5% наиболее встречаемых словосочетаний.

4. На четвертом этапе словари отдельных слов и биграммов были объединены для общей обработки алгоритмами выделения тематик. Получившийся словарь был проанализирован, на предмет выделения высоко- и низкочастотных слов для их фильтрации. Традиционно окончательное формирование словаря производится с помощью стоп-слов. Алгоритмы выделения стоп-слов не использовались автором в данной статье. Это решение было обусловлено тем, что добавление словаря стоп-слов не добавляло точности и вносило субъективный характер исследования.
5. На заключительном пятом этапе производилось построение модели тематик. Для этого был использован инструмент BigARTM [233]. На этом этапе были получены матрицы распределение тем для документа (Θ) и распределение слов для темы (Φ). Для повышения точности алгоритма автором был

применен аналитический подход, уточняющий регуляризационные параметры на основании метрик.

Основной метрикой для выявления факта сходимости модели тем является метрика Perplexity. График зависимости Perplexity от количества проходов по корпусу текстов отображен на (Рис. 4.28).

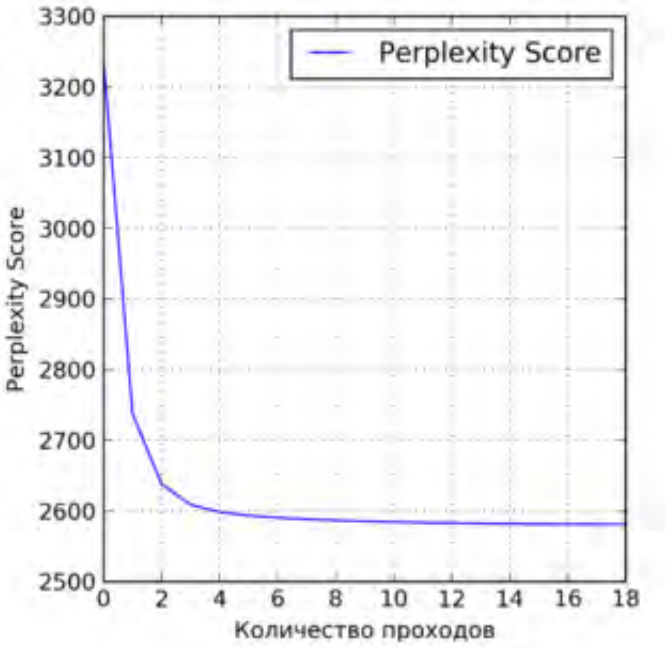


Рисунок 4.28 — Зависимость метрики Perplexity от количества проходов по корпусу текстов.

Из (Рис. 4.28) видно, что за три прохода модель показала приемлемую сходимость и не нуждается в дальнейшей оптимизации.

Важными метриками качества модели тем являются степень разрежённости матриц Φ и Θ. Повлиять на эти метрики можно с помощью параметров τ, соответствующих регуляризаторов. На рисунках Рис. 4.29 и Рис. 4.30 отображены зависимости для разрежённости матриц Φ и Θ.

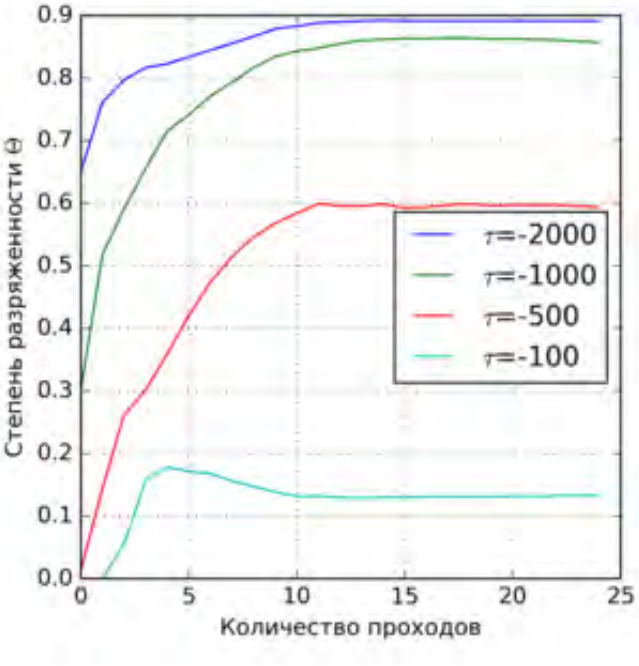


Рисунок 4.29 — Зависимость разрежённости матрицы Θ от параметра регуляризации τ.

На основании зависимостей отображенных на Рис. 4.29 и Рис. 4.30 автором были выбраны параметры регуляризации модели тем, позволяющие достичь оптимального соотношения между значимыми терминами и шумовыми.

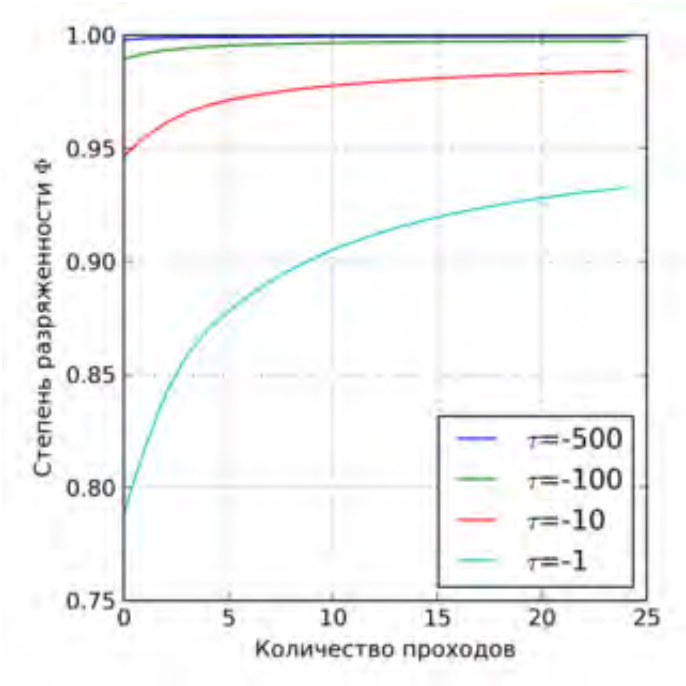


Рисунок 4.30 — Зависимость разрежённости матрицы Φ от параметра регуляризации τ.

Тематическая модель, полученная в результате данного исследования, может быть представлена в различных формах. Уровень шумовых терминов мешает интерпретировать результаты, поэтому от запланированных 12 тем содержательных осталось шесть. В Таблице (Таб. 14) представлены темы выделенные с помощью модели.

Можно с уверенностью сказать, что термины, собранные в столбце 1 характеризуют тематику управления знаниями. В столбце 2 представлена тема добычи. Остальные

столбцы тоже могут быть достаточно однозначно проинтерпретированы. А для машинной обработки набор терминов важнее чем обобщающая его тема.

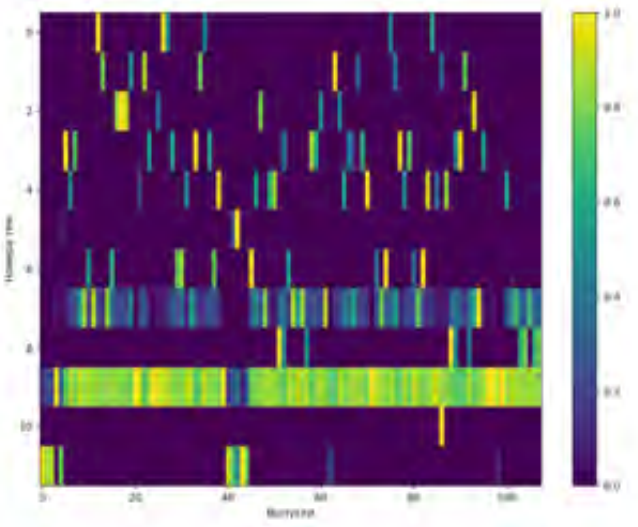


Рисунок 4.31 — Матрица Θ: Распределение тем для документов.

На Рис. 4.31 представлена матрица Θ, дающая представление как полученные тематики распределены в каждом из анализируемых выпусков. Можно увидеть, что тема с 9 представлена во всех выпусках – это общая информация, поздравления, реклама. Полученное представление так же позволяет выбирать наиболее релевантные выпуски с определенной темой.

Важно отметить, что выбранный автором метод показал высокую скорость анализа, что делает его возможным для применения в онлайновых процессах поиска. Например,

на сайте издательства в качестве средства улучшающего поиск и дающего рекомендации читателям по статьям со схожей тематикой. Также следует отметить, что данная методика может быть в дальнейшем усовершенствована и адаптирована для анализа существенно больших массивов динамических данных и выделения ключевых направлений технологического развития как в более широких, так и в более узких областях.

4.8 Разведка скрытых направлений научных исследований в нефтегазовой отрасли

По оценкам автора более 6 тысяч научно-практических статей публикуется ежегодно на основном нефтегазовом портале https://OnePetro.org. Большинство лиц, принимающих решения в нефтяной индустрии желают быть в курсе основных технологических трендов. Но лишь единицы из них имеют время на то, чтобы прочитать одну-две научных статьи в неделю. Драматически важно чтобы это время было использовано с максимальной эффективностью и выбранные научные статьи представляли действительно сфокусированные исследования высокого качества, а не вторичное перемалывание известных фактов.

Автор выбрал 1696 статей с сайта OnePetro для углубленного анализа. Эти документы были в формате PDF и нуждались в трансформации в формат пригодный для текстового анализа. Автор использовали библиотеку Apache TIKA для конвертации PDF в текст. В процессе трансформации была восстановлена пунктуация. После получения корпуса текстов необходимо было создать словарь для терминов.

На Рисунке 4.32 изображена гистограмма частот терминов (слов), которые употреблялись в выбранных статьях и доля выбранных терминов для дальнейшего анализа. С помощью такой выборки автор избавился от слов с низкими и высокими частотами употребления в коллекции текстов.

В соответствии с методикой, изложенной в разделе 3.11, в начале автор произвел тренировку PLSA topic model чтобы определить скорость сходимости по метрике *Perplexity*. Зависимость *Perplexity* от количества проходов отображена на Рисунке 4.33.

Из рисунка 4.33 видно, что модель хорошо сходится уже на 20 прогонах.

Для дальнейшего обучения к модели были добавлены следующие регуляризаторы:

1. Sparse Theta – для увеличения разрежённости матрицы Θ для основных тематик,
2. Sparse Phi – для увеличения разрежённости матрицы Φ для основных тематик,

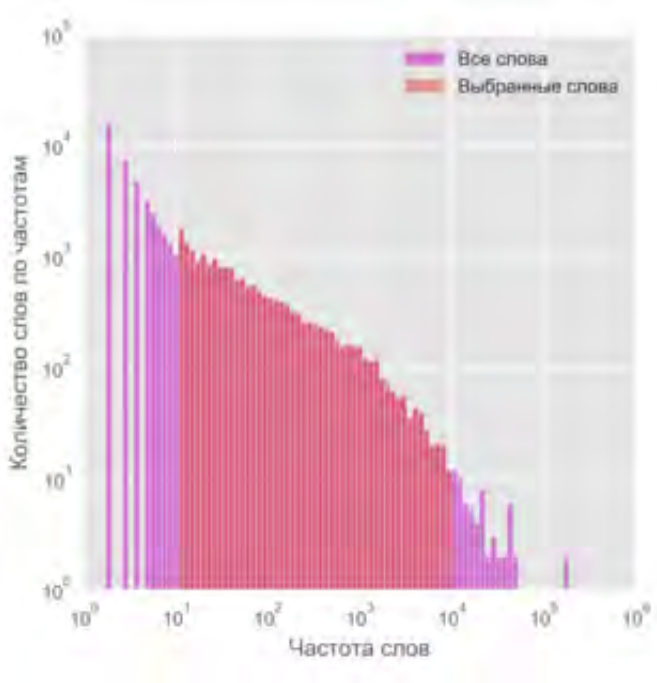


Рисунок 4.32 — Распределение частот терминов в корпусе текстов.

3. Smooth Theta – для уплотнения матрицы Θ для шумовых тематик,
4. Smooth Phi – для уплотнения матрицы Φ для шумовых тематик.

Для определения параметров регуляризаторов были проведены следующие пробные эксперименты по обучению модели. Результаты этих пробных экспериментов оценивались по метрикам разрежённости матриц Φ и Θ . Для уплотнения матриц и предполагается использовать полученные параметры τ с обратным знаком.

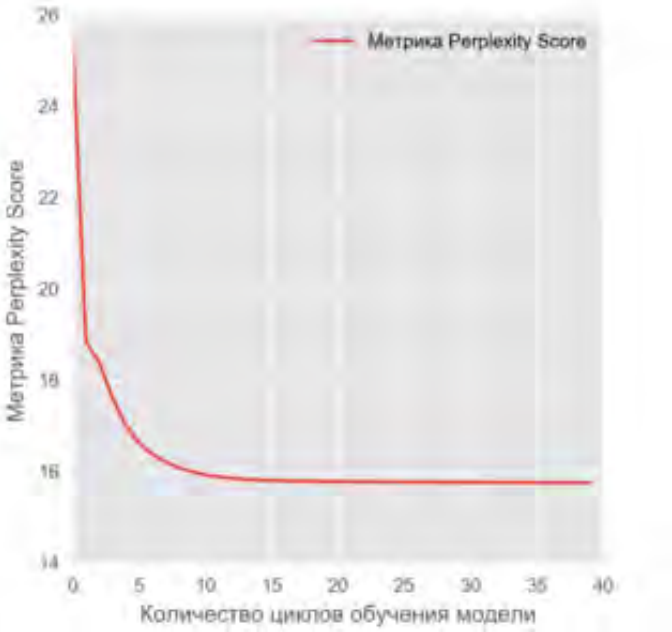


Рисунок 4.33 — Зависимость Perplexity от количества проходов.

На рисунке 4.34 приведена зависимость разрежённости матрицы для нескольких значений параметра регуляризации τ.

На основании зависимости отображенной на рисунке 4.34 для дальнейших экспериментов было выбрано значение $\tau = -10$. При таком значении τ после 25 прогонов в матрице Φ остается 92% нулевых значений.

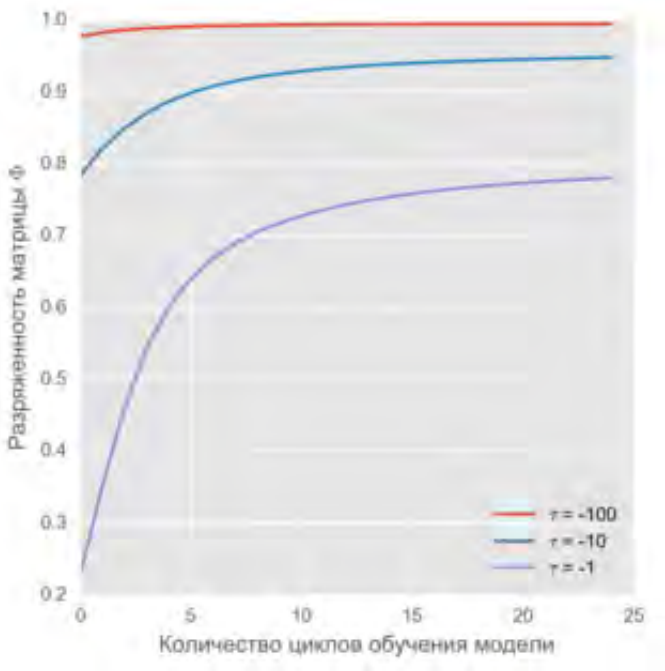


Рисунок 4.34 — Зависимость разрежённости матрицы Φ для нескольких значений параметра регуляризации τ.

На рисунке 4.35 приведена зависимость разрежённости матрицы Θ для нескольких значений параметра регуляризации τ.

На основании зависимости отображенной на рисунке 4.35 для дальнейших экспериментов было выбрано значение $\tau = 10$. При таком значении τ после 25 прогонов в матрице остается 78% нулевых значений.

Для того, чтобы качественно оценить полученную тематическую модель автор применил метод визуальных

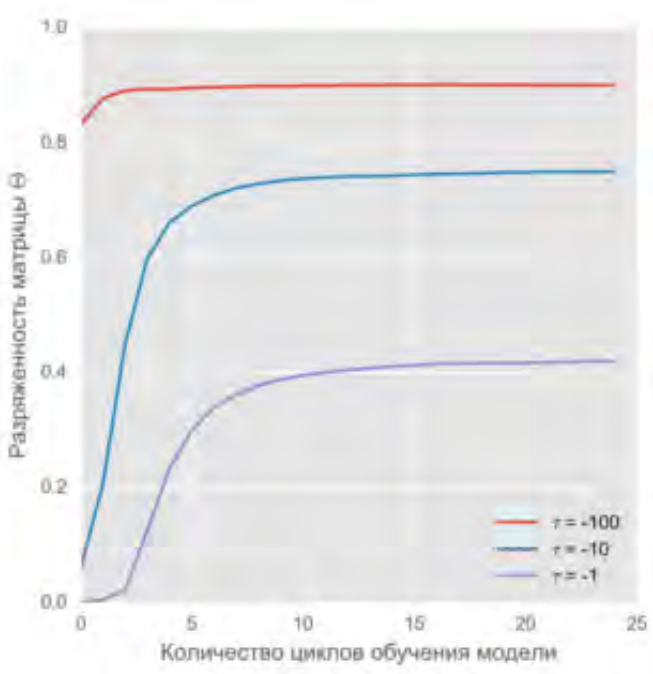


Рисунок 4.35 — Зависимость разрежённости матрицы Θ для нескольких значений параметра регуляризации τ.

оценок качества кластеризации. Рассмотрим темы в тематической модели как кластеры. Тогда хорошо выделенная тема должна описываться “близкими” словами и отстоять “далеко” от слов, образующих другие темы.

Чтобы сравнивать “близость” и “удаленность” слова были представлены в виде векторов с помощью алгоритмов FastText и GloVe. Для отображения полученных векторов были использованы два алгоритма уменьшения размерности: TSNE и MDS.

Алгоритм TSNE имеет несколько значимых параметров, таких как метрика, perplexity и learning rate. Автор рассмотрел значения perplexity от 5 до 50 с шагом 5 и перебрал следующие метрики расстояний: cosine и euclidean. Наиболее наглядные результаты представлены на 4.36 и 4.37.

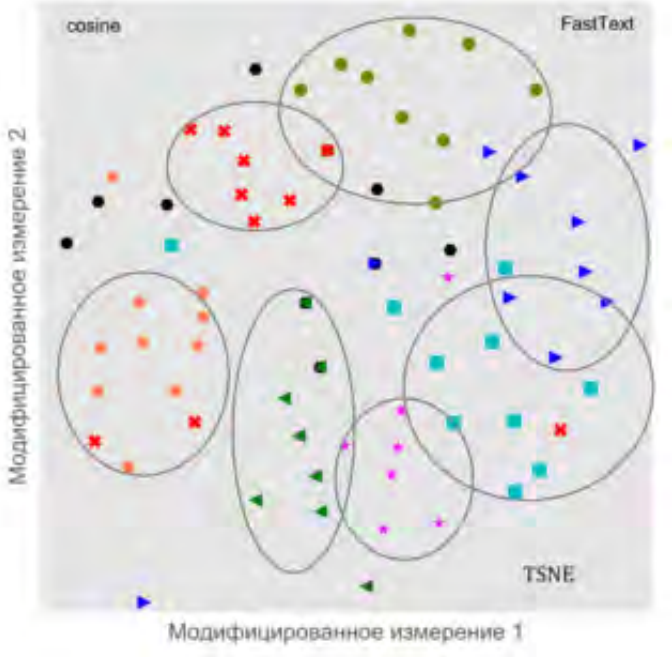


Рисунок 4.36 — Кластеры слов по тематикам. Векторы слов получены с помощью FastText. Визуализация получена с помощью TSNE с параметрами preplexity=30 и метрикой cosine.

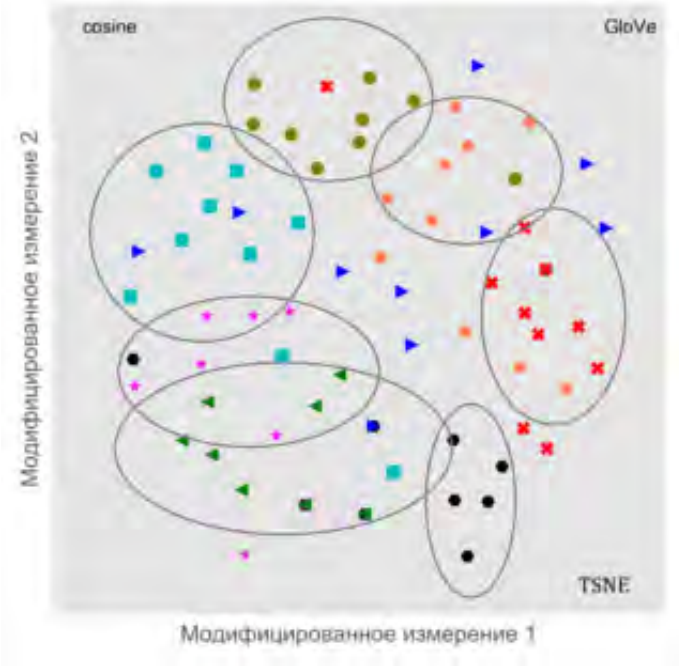


Рисунок 4.37 — Кластеры слов по тематикам. Векторы слов получены с помощью GloVe. Визуализация получена с помощью TSNE с параметрами preplexity=30 и метрикой cosine.

На основе 4.36 и 4.37 можно наблюдать группировки слов, образующих тематики. Отметим, что расстояния при трансформации векторного пространства методом TSNE не сохраняются, но сохраняются пропорции расстояний.

Преобразование векторного пространства с помощью метода MDS отображены на 4.38 и 4.39.

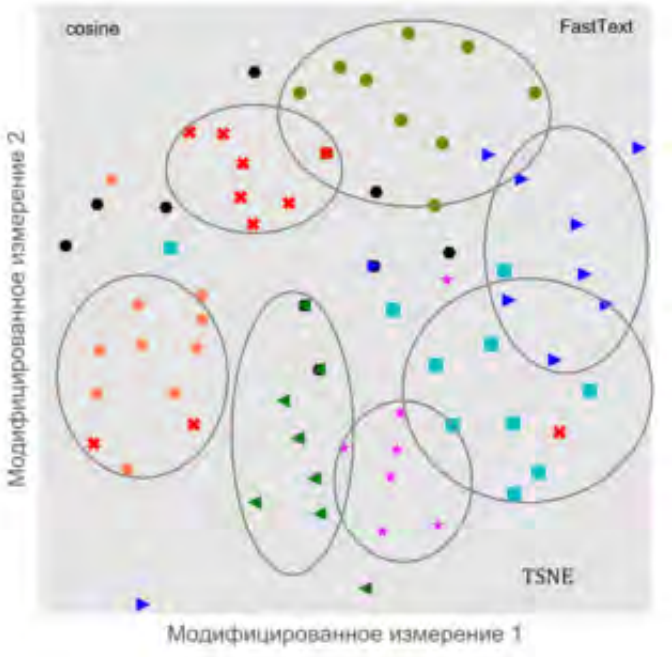


Рисунок 4.38 — Кластеры слов по тематикам. Векторы слов получены с помощью FastText. Визуализация получена с помощью MDS.

На рисунках 4.38 и 4.39 можно видеть группировку слов, образующих тематики. Алгоритм MDS использует евклидову метрику для вычисления расстояний.

Полученные с помощью MDS и TSNE результаты для FastText и GloVe показывают наличие кластеров слов , соответствующих тематикам. Так же мы видим наличие шумовых слов в тематиках.

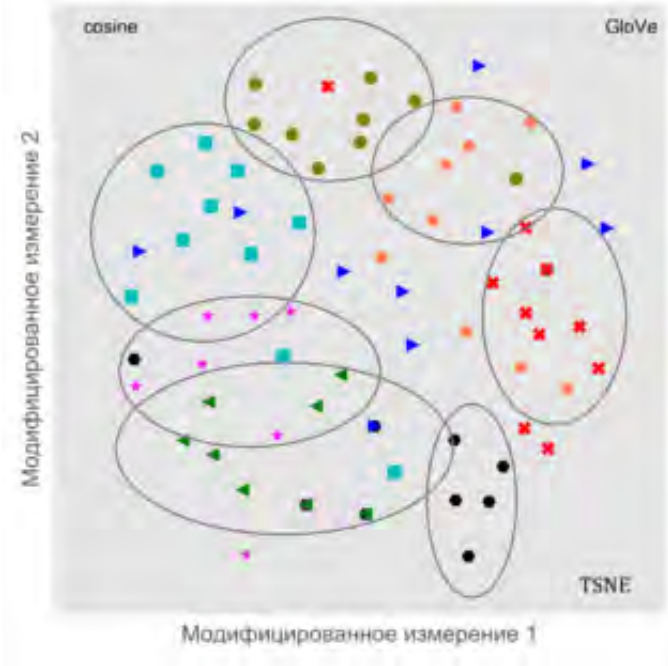


Рисунок 4.39 — Кластеры слов по тематикам. Векторы слов получены с помощью GloVe. Визуализация получена с помощью MDS.

В таблице 15 представлены top10 терминов, образующих основные тематики до регуляризации.

В таблице 16 представлены top10 терминов, образующих основные тематики после применения обучения с регуляризацией.

Из таблиц 15 и 16 мы видим, что основные термины, формирующие тематики устойчивы к процессам регуляризации. Качество интерпретируемости тематик улучшается

с регуляризацией за счет появления более конкретных терминов.

Так же представляет интерес поведение тематик для отбора шумовых терминов. В таблице приведены шумовые тематики до и после регуляризации.

Примечательно, что в шумовые тематики попала тема Сейсмики (nz1). Согласно мнению эксперта сейсмике относятся слова seismic, wave, velocity, elastic, seg, frequency и amplitude. Статьи по сейсмике мало представлены в библиотеке OnePetro и действительно могут быть отнесены к второстепенным. После обучения с регуляризацией в nz1 добавились несколько терминов связанных с вычислениями, но тема сейсмики осталась. В частности, термин offshore ушел в основные тематики. С тематикой nz0 все достаточно однозначно. В нее попали часто употребляемые слова, которые не являются на самом деле шумом.

Рассмотрим результат отнесения моделью документов к определенным тематикам. Распределение тематик для каждого документа отражены в матрице Θ . Для получения более общего представления о происходящем преобразовании матрицы авторы представили ее вид до (Рис. 4.40) и после (Рис. 4.41) регуляризации в виде 2D-карт.

Как мы видим из рисунков 4.40 и 4.41 матрица Θ в процессе регуляризации становится более разряженной на основных тематиках (sbj0-sbj10) и более плотной на шумовых тематиках (nz0-nz1).

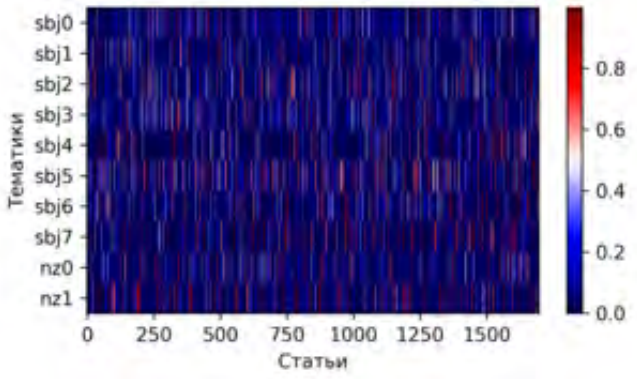


Рисунок 4.40 — Матрица Θ до регуляризации. По оси x отложены номера документов из коллекции.

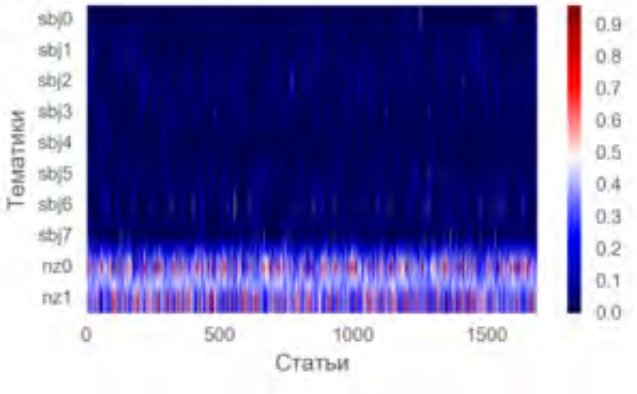


Рисунок 4.41 — Матрица Θ после регуляризации. По оси x отложены номера документов из коллекции.

Например, документ № 555 обладает самым большим весом тематики 0.72 (sbj6). Вероятности других основных тематик для этого документа равны нулю. Таким образом этот документ согласно модели, полностью по-

священ тематике sbj6, представленной словами (Таб. 16): recovery, injection, steam, core, viscosity, flooding, solvent, heavy, saturation, surfactant. При помощи эксперта sbj6 дано название: “Chemical enhanced oil recovery”.

Но с другой стороны, мы можем проверить по корпусу текстов нашей выборки, что данный документ соответствует статье с названием “Low Tension Gas Process in High Salinity and Low Permeability Reservoirs”. Вот фрагмент из публичной аннотации этой статьи с сайта OnePetro.org[1]:

Abstract *Chemical enhanced oil recovery (EOR) in carbonate reservoirs has always been technically and economically challenging. Conventional Alkaline-Surfactant-Polymer (ASP) flooding has limited application in low permeability (2-20 mD) and high salinity formations (200,000 ppm TDS) with a large concentration of divalent cations. Also into such low permeability reservoirs can be a significant problem with polymer solutions (. . .).*

Как мы видим из этого общедоступного фрагмента статьи тематика определена с высокой точностью. Но более того, из модели мы знаем, что эта статья действительно сфокусирована на этой тематике. Приобретя данную статью можно быть достаточно уверенным, что в ней не будет других тематик.

Важно так же отметить, что можно было и не прибегать к помощи эксперта для определения названия тематики

[1] https://www.onepetro.org/conference-paper/SPE-179839-MS

sbj6, а воспользоваться тем, что данный документ представлен единственной тематикой и взять название из аннотации статьи.

4.9 Выявление негативных отзывов о технологиях в текстах

Использование искусственных нейронных сетей для анализа текстов получило развитие в середине 90-х годов в работах [293—295].

Но из-за высоких требований к вычислительным ресурсам для обучения нейронных сетей оставалось академической дисциплиной. Ускорение исследований в этом направление связано с ростом скорости вычислений и с появлением таких новых архитектур искусственных нейронных сетей как сверточные нейронные сети [296] и рекуррентные нейронные сети [297]. Для обучения нейронных сетей всегда были нужны значительные массивы размеченных данных. А с появлением большего количества слоев с нейронами потребность в размеченных данных выросла в разы. Для примера, чтобы обучить искусственную нейронную сеть со ста тысячью коэффициентами нужны десятки тысяч размеченных текстов. А для архитектуры глубоких нейронных сетей количество обучаемых коэффициентов составляет миллионы [298; 299]. Поэтому обучение искусственной ней-

ронной сети на собственных данных означает выделение определенного времени и ресурсов на разметку. Другими словами, каждый классифицируемый объект человек должен отнести к одному из классов «вручную». Не так давно появились размеченные корпусы текстов с открытым доступом, например, UMICH SI650 [2] , TreeBank [3], Twitter Sentiments [4] , MPQA Opinion Corpus [5] и работы по их анализу [300—302].

Для данного эксперимента автор применил методику Transfer learning. В качестве размеченного набора данных автором были выбраны отзывы о кинофильмах [300]. В этом наборе данных содержится 25 тысяч положительных и 25 тысяч отрицательных отзывов. Набор данных таким образом сбалансирован для обучения и валидации модели классификации. Длинна отзывов варьируется от 5 до 977 слов и отображена на рисунке (Рис. 4.42).

При составлении словаря по отзывам были отброшены низкочастотные слова, то есть слова, встречающиеся в документах редко. Распределение частот слов по документам отображено на диаграмме (Рис. 4.43).

В качестве набора данных для тестирования были выбраны 1696 научно-практических статей с портала OnePetro.org.

---

[2]https://www.kaggle.com/c/si650winter11

[3]http://nlp.stanford.edu/sentiment/treebank.html

[4]http://www.sananalytics.com/lab/twitter-sentiment/

[5]http://mpqa.cs.pitt.edu

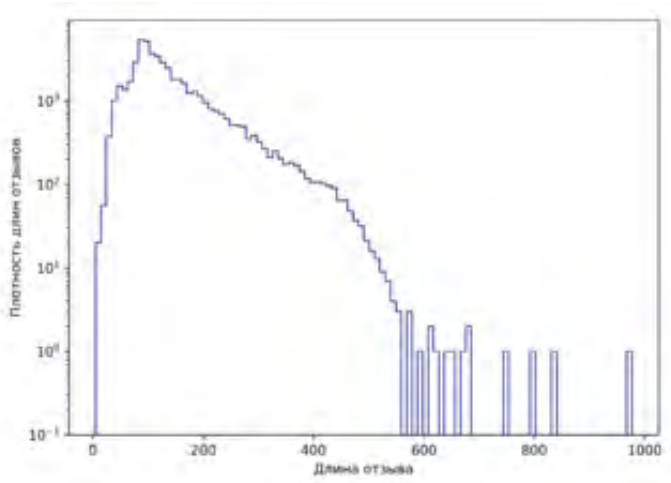


Рисунок 4.42 — Распределение длин отзывов.

Для построения векторной модели текста была использована обученная модель GloVe. Были использованы вектора с размерностью 100 и 300. Преимущества от использования обученной векторной модели текста состоит в существенном сокращении объема вычислений. Количество обучаемых параметров для создания векторной модели текста в разы превосходит количество параметров для выбранных автором архитектур моделей классификации.

Автор ограничил себя классом моделей, построенных на основе искусственных нейронных сетей. Среди архитектур искусственных нейронных сетей используемых для классификации текстов можно выделить CNN-LSTM и Stacked LSTM. Автором были выбраны следующие три

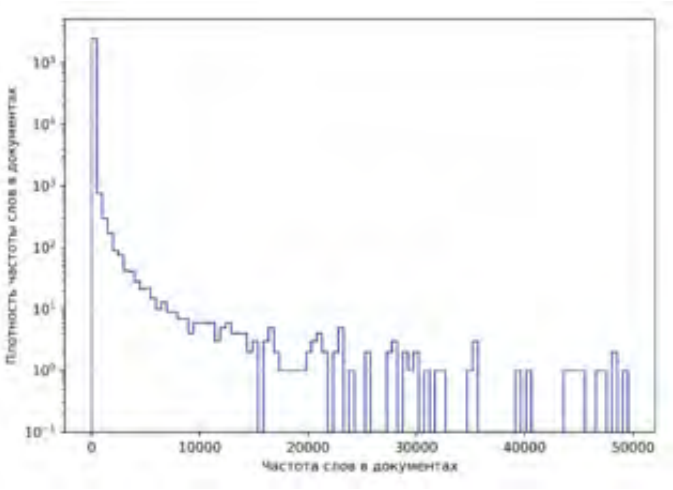


Рисунок 4.43 — Распределение частот слов по документам.

варианта архитектур моделей для классификации с использованием искусственных нейронных сетей.

1. Рекуррентная нейронная сеть из одного слоя LSTM. Далее будем называть эту архитектуру RNN и отдельно указывать количество элементов в LSTM слое.
2. Сверточная нейронная сеть из одного слоя Dropout-Conv1D-Conv1D-MaxPooling и рекуррентная нейронная сеть из одного слоя элементов LSTM. Далее будем называть эту архитектуру CNN-LSTM и отдельно указывать количество элементов и параметры сверточных слоев.
3. Рекуррентная нейронная сеть из двух слоев LSTM. Далее будем называть эту архитектуру RNN-2 и

отдельно указывать количество элементов в LSTM слое.

Для рассматриваемых архитектур моделей классификации автор выбрал следующие существенные гиперпараметры:

- Тип модели классификации: RNN, CNN-LSTM, RNN-2.
- Размерность словаря. В зависимости от фильтров низкочастотных слов размерность словаря изменялась от 2000 до 200000 слов.
- Размерность векторной модели текста: 100 и 300
- Длинна фрагмента текста: 80, 128 и 196.

Обучение моделей классификации производилось параллельно на нескольких серверах. Набор данных содержал равное количество положительный и отрицательных отзывов поэтому для оценки качества обучения была выбрана метрика Accuracy. Оптимизация параметров модели для классификации производилась на основании функции перекрестной энтропии (Cross Entropy). Для ускорения обучения автором был применен метод ранней остановки обучения на основании метрики Accuracy по валидационному набору данных. Кривые обучения для модели классификации типа RNN отображены на рисунках (Рис. 4.44 , Рис. 4.45).

Из зависимости метрики Accuracy для тренировочного и валидационного набора данных (Рис. 4.44) видно, что в районе 42-й итерации обучения метрика Accuracy перестает увеличиваться для валидационного набора данных. Это

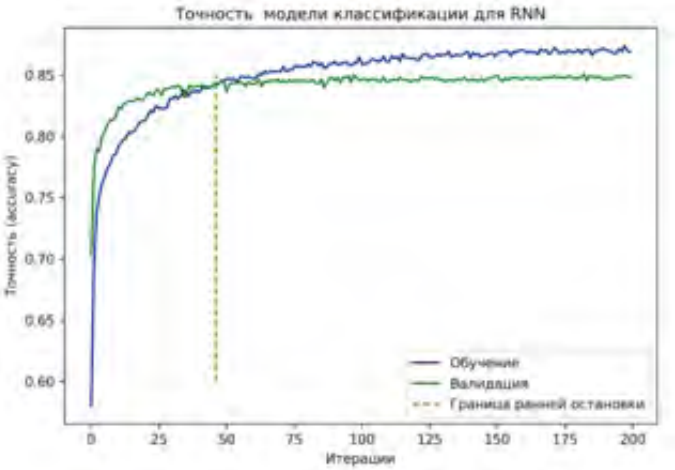


Рисунок 4.44 — Кривая обучения метрики Accuracy для модели классификации типа RNN.

явление означает, что модель начинает переучиваться по метрике Accuracy и обучение следует остановить. Данная архитектура модели для классификации не позволяет повышать точность на этом наборе данных.

Отметим что из зависимости, отображенной на Рис. 4.45 видно, что значение перекрестной энтропии на валидационном наборе данных начинает не убывать в районе 37 итерации. То есть, немногим раньше, чем начинает деградировать метрика Accuracy.

На основании изложенных выше методик было проведено обучение моделей классификации с различными архитектурами и гиперпараметрами. Результаты обучения приведены в таблице (Таб. 18).

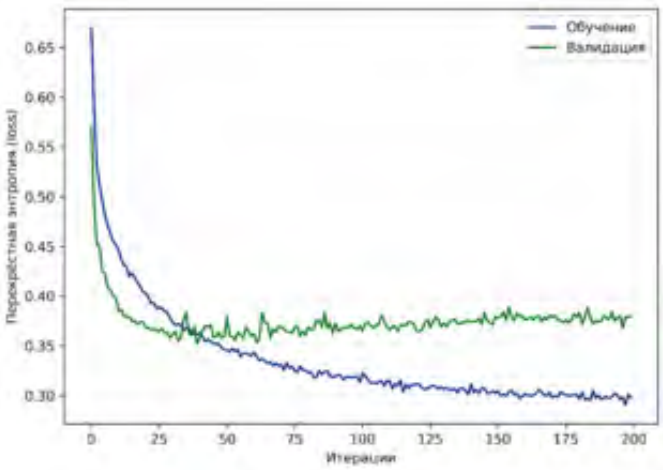


Рисунок 4.45 — Кривая функции потерь для модели классификации типа RNN.

Лучшее значение метрики Accuracy на валидационном наборе данных показала модель RNN со словарем из 23 тысяч слов и размерностью векторной модели текста равной 300. Отметим, что на тестовом наборе данных значение метрики Accuracy для данной модели составило 88%.

Полученная модель была использована для предсказания тональности научных статей портала OnePetro.org [303]. Каждая научная статья разбивалась на фрагменты длинной 196 слов для оценки эмоциональной окраски. Затем фрагменты статей собирались обратно для получения эмоциональной карты всей статьи. Таким образом, можно было определить фрагменты статьи, обладающие аномальными

эмоциональными окрасками, такими как разочарование и удовлетворенность.

Данное исследование не принимает в расчет семантику текста, поэтому предмет эмоциональной окраски автоматически не определялся. Выбранные фрагменты статьи необходимо проанализировать с помощью эксперта. Но такой подход к аннотированию статьи позволил найти сложно обнаруживаемые фрагменты. В Таблице 19 приведены примеры эмоциональных фрагментов статей.

Так же автор разработал цветовое представление эмоциональной окраски статей в зависимости от вероятности отнесения фрагмента текста к положительной или отрицательной эмоциональной окраске.

На рисунке (Рис. 4.46) эмоциональность фрагментов статей отображена в виде карты. Для каждой статьи на оси X цветом отображена эмоциональность каждого фрагмента последовательно по оси Y.

Научные статьи используют академическую лексику и ожидать в них градус эмоций сравнимый с отзывами на кинофильмы было бы наивно. Но современные концепции обработки текста, основанные на анализе контекста, позволяют выделять и классифицировать изменения эмоциональности достаточно точно для того, чтобы обрабатывать даже научные статьи. Автор считает, что проведенное исследование открывает возможности по созданию дополнительных инструментов для аннотации и классификации научных текстов.

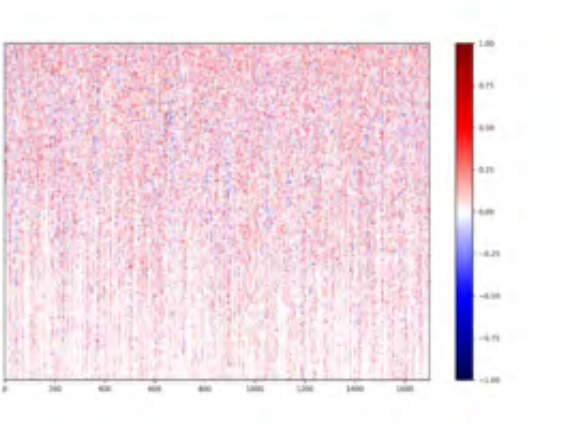


Рисунок 4.46 — Карта полярности эмоциональной окраски статей 1696 статей. По оси x отложен порядковый номер статьи, по оси y эмоциональная окраска фрагментов статьи. На цветовой шкале отображена цифровая характеристика эмоциональности: негативная (-1), позитивная (+1).

4.10 Сравнение корпусов научных статей

Для экспериментального подтверждения предложенной методики T4C (3.20) авторами была выбрана коллекция патентов [304]. Патент так же как и научная статья является слабо-структурированным документом. Патент обладает следующими формализованными полями: название, описание, автор, страна, дата. Для исследования были выбраны 9992 патента из Китая(CN) и 22502 из США (US). Для вы-

бора использовалось ключевое слово *seism*. Таким образом, на входе было две коллекции документов по направлению сейсмика.

Размерность словаря получившейся коллекции составляет 16 млн единиц. В словарь входят биграмы. Представление текста в виде биграммной вероятностной модели выбрано для более емкой передачи смыслов. Для уменьшения редко встречаемых слов были выбраны фильтры: минимальная частота слова более 4 и максимальное количество документов в которых встречается слово не более 40% от общего количества документов. После фильтрации словарь составил 4 млн единиц.

Для проверки гипотезы о классификации нам нужно вычислить точность с которой документ может быть отнесён к своей коллекции. Для этого авторы использовали подход на основе "мешка слов"(Bag of Words, BoF). Для частотного анализа была применена библиотека *sklearn*. В качестве начальной оценки авторы применили классификатор на основе алгоритма Байеса [305]. Так как метки для классификации представлены в коллекциях неравномерно, то при разделении набора данных на обучающий и тестируемый была использована стратификация. Для подготовки признаков (features) использовался конвеер (pipiline) по трансформации текста в частоты и коэффициенты TF-IDF.

В результате обучения модель показала точность (*Accuracy*) равную 97.5% на отложенной выборке с размером

0.3 от всей выборки. Такая высокая точность классификации означает, что каждый документ очень характерен для своей коллекции. Китайский патент по составу текста описания заявки принципиально отличается от Американского патента.

Содержание патентов не должно повторяться со временем. Суть научной разработки в том, чтобы предлагать что-то новое. Поэтому в рамках методики T4C авторы проверили насколько точно классифицируется патенты по годам для двух коллекций. Распределение патентов по годам представлено на рисунке 4.47.

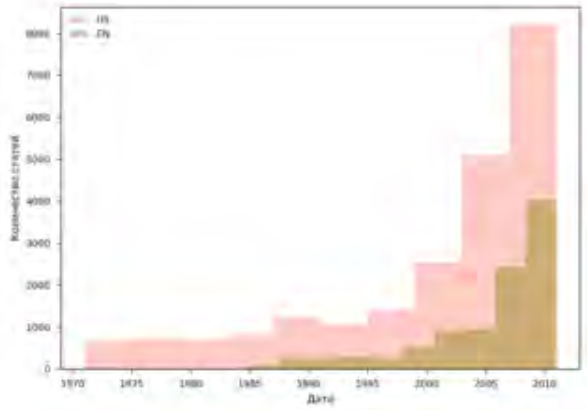


Рисунок 4.47 — Гистограмма распределения коллекции по годам.

Для проверки гипотезы о миграции контента во времени был использован тот же конвеер. В результате патенты сделанные до 2004 года отличаются от патентов сделанных в 2010 году с точностью 85% (US) и 67% (CN).

Следующим шагом была построена модальная тематическая модель. На рисунке 4.48 отображены качественные характеристики (метрики Контраст и Чистота ядра тематик) обучения тематической модели, показывающие, что сходимость достигнута:

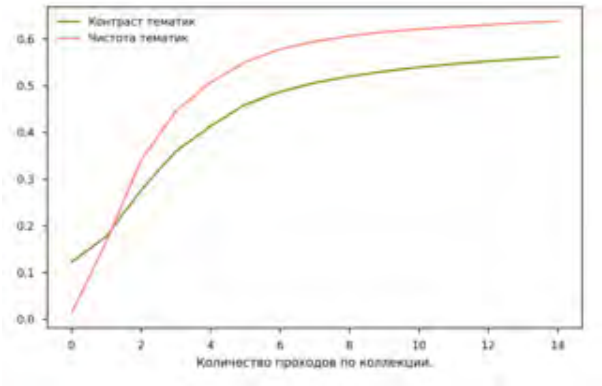


Рисунок 4.48 — Метрики качества обучения тематической модели.

В качестве модальностей были выбраны текст описания, страна и год. Веса модальностей были подобраны в соответствии с методикой, описанной в работе [233] и составили 1:3:3. Количество тематик модели так же является свободным параметром. С помощью методики, разработанной автором работах [306; 307] было определено, что для данной коллекции достаточно 30 тематик.

На рисунке 4.49 представлена матрица θ, описывающая распределения вероятностей в координатах "тематика–документ"для каждой страны.

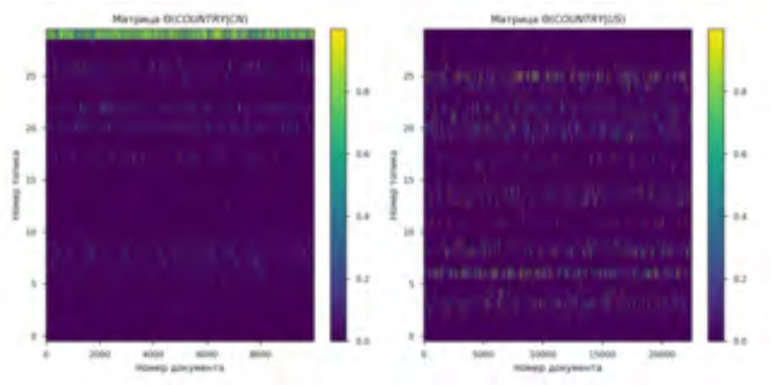


Рисунок 4.49 — Распределение "тематика–документ"для обоих коллекций.

С помощью рисунка 4.49 авторы провели визуальный анализ распределений "тематика–документ"(θ_{td}) для различных тематик. Полосы на рисунке соответствуют тематикам, которые имеют большие вероятности для многих документов. Для Китайских патентов это тематики представленные следующими термами из матрицы Ф (Таблица 20.).

Для тематик Американских патентов нет таких характерных полос. Это означает, что выделенные тематики распределены более равномерно. В таблице 21 показаны тематики Американских патентов с высокими вероятностями.

Для дальнейших шагов эксперимента важно отметить, что из пяти тематик, приведенных в таблицах 20 и 21, тематики *sbj20* и *sbj8* являются важными для обоих коллекций, остальные тематики отличаются.

Для более точного определения степени тематического пересечения авторами в рамках методики T4C была рассчитана точность классификации патентов по матрице θ. Значение метрики Accuracy было вычислено с помощью метода RandomForest и получилось 65%. Оптимальные параметры метода RandomForest были подобраны с помощью поиска по сетке (GridSearch) с осреднением по 5 прогонам и составили: количество эстиматоров – 200, максимальная глубина дерева – 5. Полученный результат согласуется с визуальными наблюдениями матриц θ , сделанными на основании рисунка 4.49. Действительно, из визуального наблюдения видно, что распределение тематик по документам для каждой из меток (CN,US) недостаточно различается, чтобы точность классификации была высокой.

4.11 Классификация перевода статей на английский язык на основании мультимодальной тематической модели

Для проверки предложенной гипотезы авторы собрали двуязычный корпус из 242 статей на английском и 242 статей на русском с портала OnePetro.org международного сообщества нефтегазовых инженеров (SPE). Соответствие статей на разных языках установлено по индексу DOI.

При создании словарей была применена лемматизация и отброшены высоко и низко частотные слова. Размер словарей для русского и английского корпуса подобран одинаковый около $0.5 * 10^5$ слов.

Обучение модели останавливалось при достижении пологого характера изменений метрики перплексии

$$\mathcal{P} = \exp\left(-\frac{1}{n}\sum_{d\in D}\sum_{w\in W} n_{dw} \ln \sum_{t\in T} \varphi_{wt}\theta_{td}\right), \qquad (4.8)$$

характеризующей информационную энтропию модели.

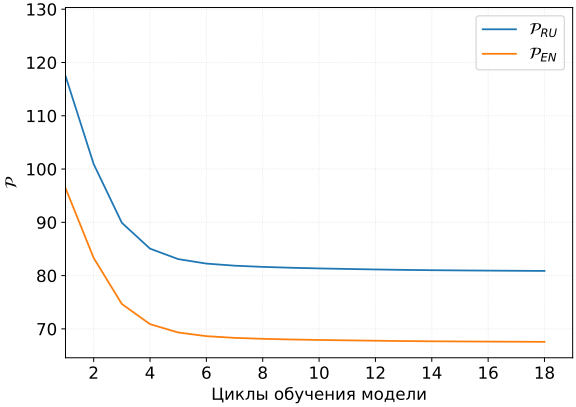


Рисунок 4.50 — Зависимость метрики перплексия от циклов обучения модели.

На рисунке 4.50 мы наблюдаем,что энтропия русского текста выше энтропии английского. Сравнение значений *Perplexity* для разных языков находятся в согласии с результатами, опубликованными в работе [308]. Метрика *Perplexity* достаточно сильно зависит от редких слов в словаре.

Для регуляризации были использованы подобранные в работе [237] коэффициенты μ. После двенадцати итераций обучения с последовательной регуляризацией полученная матрица θ_{dt} представлена на рисунке 4.51.

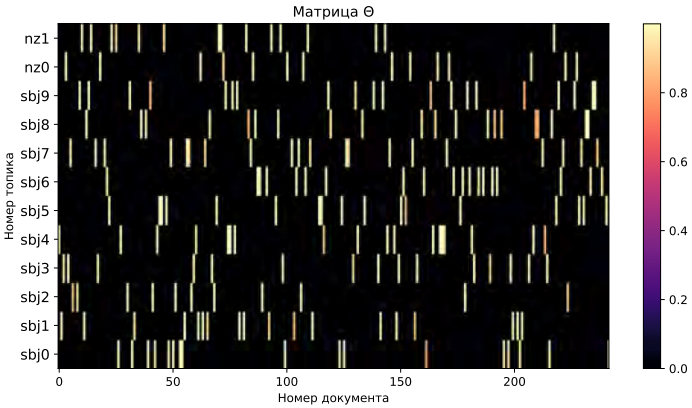


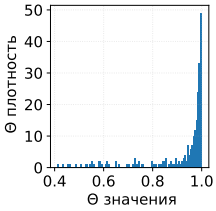
Рисунок 4.51 — Матрица значений θ_{td} после обучения модели.

На рисунке 4.51 мы видим по оси y тематики: sbj_{0-9} – основные и nz_{0-1} – шумовые. За счет разнонаправленной регуляризации пространство основных тематик разряжено, а пространство шумовых тематик сглажено. Рассмотрим примеры соответствий тематик для русского и английского корпусов в таблице 22.

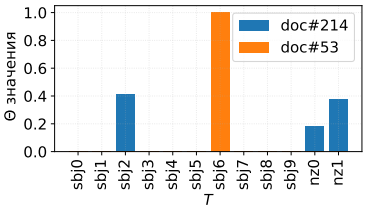
Из таблицы 22 видно, что тематики на английском и русском для многих t_i полностью совпадают. Этот результат визуального анализа говорит о том, что тематическая модель настроена на существующую зависимость в данных. Но еще есть расхождения, которые надо анализировать.

Рассмотрим плотность максимальных значений θ_{td} для каждого документа, отображенную на рисунке 4.11:

Из рисунка 4.11 мы видим, что существует два характерных класса документов, которые можно разделить по значению максимальной θ_{td}. Рассмотрим подробнее документы, с диаметрально противоположными значениями максимальной θ_{td} : 0.4101 документ №214 и 0.9998 документ №53.



Плотность значений θ_{td} после обучения модели.



Значения θ_{td} для разнотипных документов.

Рисунок 4.52 — Значения θ_{td}.

Из рисунка 4.11 видно, что документ №214 укладывается в одну тематику sbj6, а документ №53 распределён по трём тематикам: sbj2,nz0 и nz1. Сравним веса термов каждой из этих тематик, чтобы понять насколько они коррелированы.

На рисунке 4.53 отображены десять наибольших значений из матрицы φ_{wt} для тематик sbj6,sbj2,nz0 и nz1. Вероятности термов для тематики sbj6 значительно выше,

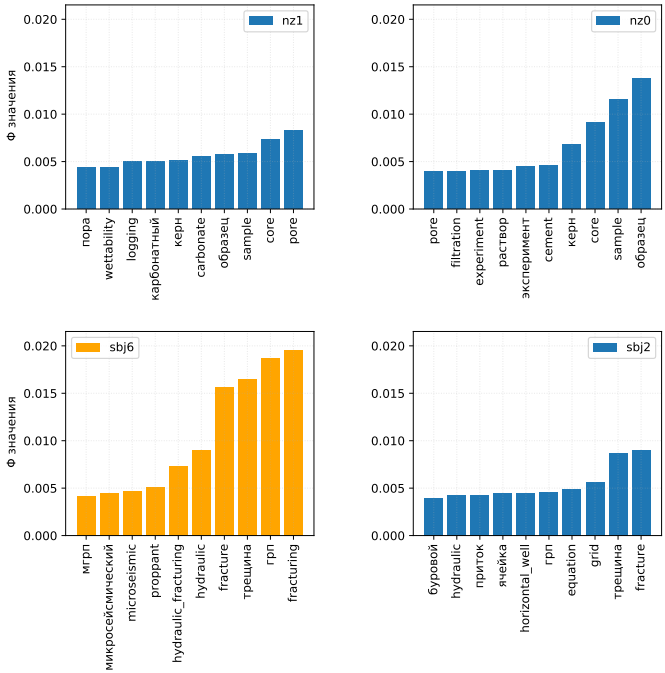


Рисунок 4.53 — Значенияφ_{wt} для двух разнотипных документов.

чем для sbj2,nz0 и nz1. Это наблюдение свидетельствует о различном характере переводов документов.

Таким образом, исследуемая коллекция содержит документы двух разных типов, которые можно разделить по характеру распределений в них тематик. Большинство составляют документы, с содержащие одно значение θ_{td} большее 0.95.

4.12 Оценка оптимального количества тематик в тематической модели

Для эксперимента был использован корпус научно-технических статей по темам, связанным с разработкой нефе-газовых месторождений. Всего было выбрано 1695 статей на английском языке по 10 направлениям исследований согласно рубрикатору. Создание словаря для выбранного корпуса подробно описано в предыдущем исследовании автора [237]. Для построения тематической модели была использована библиотека BigARTM, позволяющая производить настройку тематической модели путём последовательной регуляризации. Выбор и настройка параметров регуляризации тематической модели сделаны автором в предыдущем исследовании [237]. Для преобразования разряженного пространства векторов слов, составляющих тематики, была выбрана библиотека GloVe [273]. Для получения визуального представления о виде плотного представления тематик была сделана проекция на двумерное пространство с сохранением расстояний с помощью библиотеки MDS [309]. Полученный таким образом вид кластеров тематик представлен на рисунке 4.54.

На рисунке 4.54 различными маркерами выделены двумерные проекции слов из тематик. Овалы сделаны для того, чтобы подчеркнуть уверенную визуальную сгруппированность слов в тематиках.

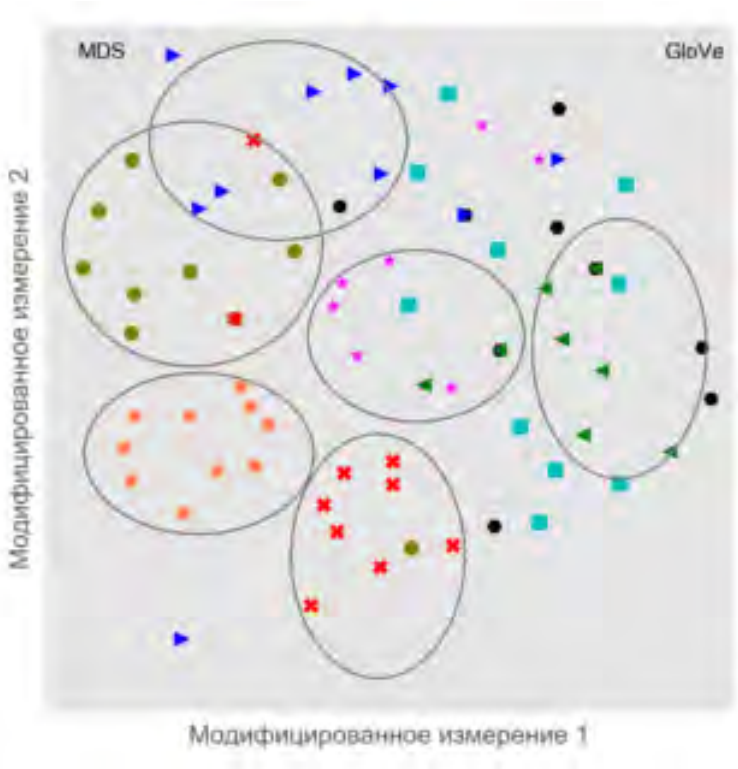


Рисунок 4.54 — Проекция плотного представления тематик с сохранением расстояний.

На рисунке 4.55 представлены предварительные расчёты поведения основных метрик тематической модели, настроенной в соответствии с предложенной авторами методикой, в зависимости от количества тем.

Как мы видим из рисунка 4.55, характер зависимостей носит монотонный характер и не позволяет определить оптимальное количество тем. Измерения основных внутренних метрик сделаны для 1000 различных случайных

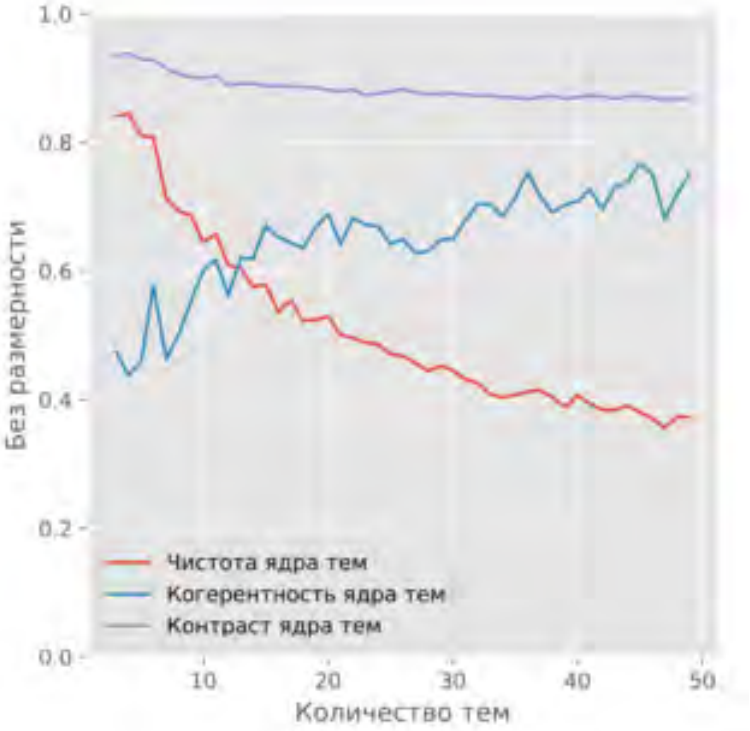


Рисунок 4.55 — Зависимости основных внутренних метрик качества тематической модели от количества тем.

порядков документов. По оси y отложено значение одного стандартного отклонения. Мы видим, что для метрики *Контраст ядра тем* отклонения минимальны. Для метрик *Чистота ядра* тем и *Когерентность ядра тем* большее значения характеризуют лучшее качество тематической модели. Характерной точкой можно считать количество тем равное 12, когда кривые изменения метрики *Чистота ядра тем* и *Когерентность ядра тем* пересекаются.

Рассмотрим зависимости метрик Calinski-Harabaz Index [310], Silhouette Coefficient [272], используемых для валидации количества кластеров.

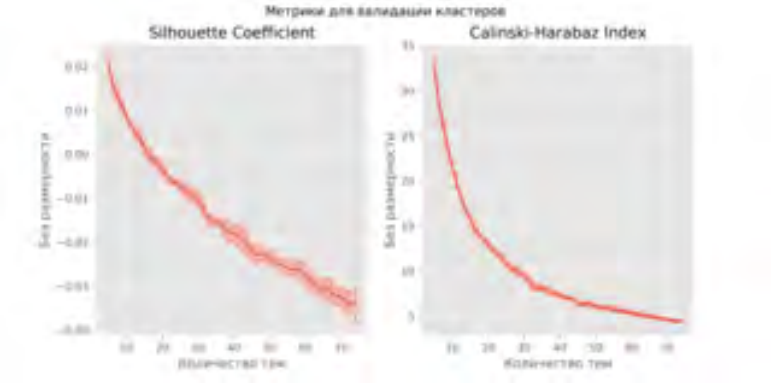


Рисунок 4.56 — Метрики валидации кластеров.

Как можно увидеть из рисунка 4.56 метрики Calinski-Harabaz Index и Silhouette Coefficient не дают возможности определить оптимальное количество тем. С ростом количества тем значения этих метрик уменьшаются, что говорит о том, что с точки зрения этих метрик кластеры становятся хуже. По-другому в зависимости от количества тем ведёт себя метрика cDBI, разработанная авторами и изображённая на рисунке 4.57.

На рисунке 4.57 можно видеть явно выраженный максимум при количестве тематик равному 16. Алгоритм для расчёта метрики cDBI основан на идеологии метрики Davies Bouldin Index, предложенной в работе [269] и модифицированной в работах [270; 271].

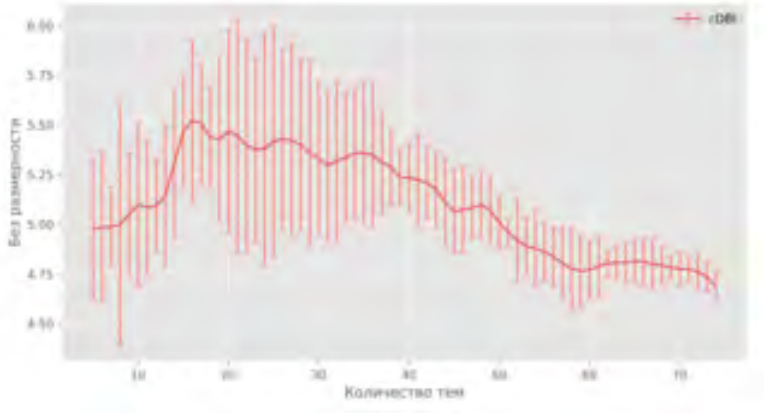


Рисунок 4.57 — Метрика cDBI.

Algorithm 1: Алгоритм вычисления метрики *Cosine Davies Bouldin Index (cDBI)* .

Result: $cDBI$

$V := GloVe(ARTM(tn, \mu, (corpus\,of\,texts))$

for $t \in W$: **do**

$C_t := \sum_{i \in t} V_t^{(i)}$

$D_t := \frac{1}{\dim t} \sum_{i \in t} \frac{C_t \cdot V_t^{(i)}}{|C_t| \cdot |V_t^{(i)}|}$

end

$cDBI := \frac{1}{\dim W} \sum_{t \in T} \frac{D_t}{C_t}$

В приведённом выше Алгоритме 1 T обозначает количество выбранных тем, а μ - это регуляризующие коэффициенты. Таким образом, с помощью метрики $cDBI$ возможно найти оптимальное количество тем для коллекции документов.

Таблица 5 — Динамические переменные модели численности персонала

Название переменной	**Обозначение**	**Формула**
Общее количество сотрудников	Total Employees	RookieEmployees + ExperiencedEmployees
Эффективное количество сотрудников	Effective Workforce	ExperiencedEmployees + RookieProductivityFraction * RookieEmployees
Средняя продуктивность	Average Productivity	EffectiveWorkforce / TotalEmployees
Устойчивая доля новичков	Steady State Rookie Fraction	AdaptationTime * (ExperiencedQuitFraction + GrowthRate) / (1 + AdaptationTime * (ExperiencedQuitFraction + GrowthRate))
Скорость увольнений	Total Quit Rate	RookieQuitRate + ExperiencedQuitRate

Таблица 6 — Формулы потоков модели численности персонала.

Название потока	**Формула**
Набор новичков	RookieHireRate = TotalQuitRate + TotalEmployees * GrowthRate
Увольнение новичков	RookieQuitRate = RookieEmployees * RookieQuitFraction
Адаптация новичков в опытных сотрудников	AdaptationRate = RookieEmployees / AdaptationTime
Увольнение опытных сотрудников	ExperiencedQuitRate = ExperiencedEmployees * ExperiencedQuitFraction

Таблица 7 — Свободные параметры модели выполнения заданий.

Название параметра	**Обозначение параметра**	**Значение параметра**
Стандартная рабочая неделя	Standard Workweek	40 часов
Целевая задержка исполнения задания	Target Delivery Delay	0.2 недель
Начальное время выполнения задания	Initial Standard Time Per Task	1 человек*час/задание
Минимальная задержка выполнения задания	Minimum Delivery Delay	0.05 недель

Таблица 8 — Динамические переменные модели выполнения заданий.

Название переменной	**Обозначение**	**Формула**
Необходимое количество сотрудников	Desired Workforce	DesiredCompletionRate * StandardTimePerTask / StandardWorkweek
Потенциальная скорость выполнения заданий	Potential Completion Rate	EffectiveWorkforce * Workweek / TimePerTask
Время затраченное на выполнение задания	Time Per Task	StandardTimePerTask * EffectOfWorkPressureOnTimePerTa WorkPressure)
Требуемая скорость выполнения заданий	Desired Completion Rate	ServiceBacklog / TargetDeliveryDelay
Степень нагрузки на персонал	WorkPressure	DesiredWorkforce / EffectiveWorkforce
Рабочая неделя	Workweek	StandardWorkweek * EffectOfWorkPressureOnWorkweek(WorkPressure)

Таблица 9 — Динамические переменные, объединяющие модель численности персонала и модель выполнения заданий.

Название переменной	**Обозначение**	**Описание действия**
Эффективное количество сотрудников	Effective Workforce	Вычисляется в модели численности персонала для модели выполнения заданий. Описывает количество сотрудников для выполнения заданий.
Рабочая неделя	Workweek	Вычисляется в модели выполнения заданий для модели численности персонала. Описывает количество часов требуемых для выполнения поступающих заданий.

Таблица 10 — Оптимальные значения параметров

Название параметра	**Значение параметра**
Количество сотрудников в организационной среде (N_o)	136
Скорость появления новых сотрудников ($Vemp_{new}$)	1 сотрудник в неделю
Скорость увольнения сотрудников ($Vemp_{fire}$)	1 сотрудник в месяц
Максимальное количество компетенций у сотрудника ($Cmax_{emp}$)	4
Максимальное количество компетенций необходимых для достижения цели ($Cmax_{pub}$)	5

Таблица 11 — Размеры связанных компонент графа соавторств по годам нарастающим итогом.

Год	Размеры связанных компонент	Доля малых компонент
2017	556, 8, 8, 8, 6, 5, 5, 5, 4, 4, 4, 4, 4, 3, 3, 3, 2, 2, 2, 2, 2, 2, 2, 2, 2	15%
2016	367, 8, 8, 8, 8, 8, 6, 5, 5, 5, 5, 4, 4, 4, 4, 3, 3, 3, 2, 2, 2, 2, 2, 2, 2	23%
2015	89, 22, 21, 15, 12, 12, 8, 8, 8, 8, 6, 6, 5, 4, 3, 3, 2, 2, 2, 2, 2	63%
2014	46, 18, 15, 12, 12, 10, 8, 8, 8, 7, 6, 5, 4, 4, 3, 2, 2, 2, 2, 2	74%
2013	23, 15, 12, 11, 10, 8, 8, 7, 5, 4, 4, 4, 2, 2, 2	80%
2012	15, 14, 12, 11, 8, 8, 7, 4, 4, 4	83%
2010	12, 9, 8, 8, 4, 3	73%
2008	12, 8, 7, 3	60%

Таблица 12 — Сравнение классификаторов по метрике ROC AUC.

Модель	ROC AUC
KNeighborsClassifier	0.66
RidgeClassifier	0.73
RandomForestClassifier	0.72
SVM	0.70
Multi-layer perceptron	0.75

Таблица 13 — Отчет о выполнении прогноза авторства на 2018 г.

	precision	recall	f1-score	support
not author	0.80	0.98	0.88	66
author	0.80	0.20	0.32	20
Avg / Total	0.80	0.80	0.75	86

Таблица 14 — Фрагмент матрицы Ф для терминов с максимальными вероятностями.

Тема1	Тема2	Тема3	Тема4	Тема5	Тема6
электронный	ЭЦН	сдвиг	почва	нефтегазоносность	ингибитор
знание	УЭЦН	сигнал	добавка	свод	разлом
автоматизация	сероводород	окисление	композиция	компания	деформация
интегрировать	фациальный	разрушение	знание	впадина	трещиноватость
пользователь	гамма	деформация	агрегат	сепарация	исследовательский
архив	доломит	реологический	загрязнение	миграция	известняк
хранение	замер	песчаный	ПЗП	прогнозный	порода
доступ	депрессия	осадки	надежность	активность	политехнический
подразделение	агент	капиллярный	камень	филиал	штанга
платформа	каротаж	сечение	окисление	цемент	приемистость

Таблица 15 — Top10 терминов, образующих тематики до регуляризации.

Топик	sbj0	sbj1	sbj2	sbj3	sbj4	sbj5	sbj6	sbj7
Терм1	liquid	sand	stress	injection	corrosion	casing	injection	safety
Терм2	equation	shale	fractures	history	nace	mud	recovery	management
Терм3	velocity	porosity	hydraulic	matrix	samples	cement	steam	risk
Терм4	pipe	logging	fracturing	optimization	concentration	hole	viscosity	assessment
Терм5	experimental	pore	proppant	recovery	acid	tubing	core	human
Терм6	eq	samples	shale	porosity	treatment	string	heavy	health
Терм7	coefficient	core	stage	linear	steel	drill	injected	company
Терм8	multiphase	log	treatment	matching	ph	bit	polymer	team
Терм9	equations	sample	conductivity	match	inhibitor	completion	flooding	equipment
Терм10	mass	logs	stimulation	cumulative	chemical	mpd	solvent	environmental

Таблица 16 — Top10 терминов, образующихосновные тематики после применения обучения с регуляризацией.

Топик	sbj0	sbj1	sbj2	sbj3	sbj4	sbj5	sbj6	sbj7
Терм1	liquid	shale	fracturing	injection	corrosion	casing	recovery	safety
Терм2	pipeline	porosity	proppant	fractures	nace	cement	injection	management
Терм3	pipe	logging	hydraulic	shale	concentration	mud	steam	risk
Терм4	velocity	sand	stress	matrix	samples	hole	core	human
Терм5	multiphase	pore	fractures	hydraulic	inhibitor	mpd	viscosity	health
Терм6	slug	samples	stage	recovery	acid	bit	flooding	business
Терм7	friction	core	shale	fractured	ph	drill	solvent	assessment
Терм8	bhr	spwla	treatment	bakken	steel	string	heavy	training
Терм9	group	clay	conductivity	porosity	houston	pipe	saturation	company
Терм10	holdup	symposium	stages	unconventional	iron	liner	surfactant	activities

Таблица 17 — Top10 терминов, образующих шумовые тематики до и послеприменения обучения с регуляризацией.

До регуляризации		**После регуляризации**	
nz0	**nz1**	**nz0**	**nz1**
pump	wave	pump	stress
pipeline	seismic	sand	equation
esp	seg	completion	seismic
power	frequency	injection	wave
subsea	velocity	tubing	velocity
operating	waves	equipment	numerical
lift	amplitude	operating	x
equipment	x	downhole	pore
installation	elastic	power	our
liquid	offshore	esp	direction

Таблица 18 — Результаты обучения моделей классификации с различными гиперпараметрами.

Архитектура модели	Количество параметров, тыс.	Длина фрагмента текста	Размерность векторного пространства текста	Словарь, количество слов	Точность валидации
CNN+RNN	63	128	100	2 300	0,85
CNN+RNN	63	196	100	2 300	0,87
CNN+RNN	63	196	100	23 400	0,86
RNN	69	128	100	2 300	0,87
RNN	722	196	300	23 400	0,88
RNN	81	128	100	47 969	0,85
RNN-2	161	128	100	2 300	0,86
RNN-2	161	196	100	2 300	0,87
RNN-2	1 443	196	300	23 000	0,87
RNN-2	1 443	196	300	248 739	0,85
RNN-2	1 443	80	300	248 739	0,85

Таблица 19 — Выявленные эмоциональные фрагменты статей.

the results from pilot tests which were using as injectant are disappointing and the results from pilot tests which were using natural gases are encouraging
to sum up diffusion mechanism for in pilot tests had not been well recognized which in turn did not enhance oil production rate in those wells
the outstanding result from this study
using the other forward model result dramatically bad

Таблица 20 — Тематики Китайских патентов с высокими вероятностями.

nz1	**sbj26**	**sbj20**	**sbj22**	**sbj8**
earthquake	polymer	spring	wave	user
model	fiber	plate	phase	communication
implementation	resin	mass	vibration	network
steel	weight	axis	response	node
carry	layer	tube	mass	module
utility	composition	vibration	transducer	sensor
plate	temperature	contact	detector	computer

Таблица 21 — Тематики Американских патентов с высокими вероятностями.

sbj6	**sbj25**	**sbj3**	**sbj20**	**sbj8**
member	trace	output	spring	user
panel	seismic_data	input	plate	communication
plate	receiver	filter	mass	network
assembly	velocity	circuit	axis	node
frame	survey	digital	tube	module
building	reflection	pulse	vibration	sensor
structural	equation	block	contact	computer

Таблица 22 — Таблица с тематиками из матриц φ_{wt}^{rus} и φ_{wt}^{eng}.

sbj0:RU	реакция, колонна, приток, насос, раствор, обсадный, концентрация
sbj0:EN	reaction, casing, pump, logging, log, mixture, string
sbj1:RU	долото, управление, дкс, наземный, интегрировать, мощность, пара
sbj1:EN	bit, steam, network, integrated, bcs, run, risk
sbj2:RU	трещина, грп, ячейка, приток, буровой, сетка, раствор
sbj2:EN	fracture, grid, equation, horizontal well, hydraulic, liner, boundary
sbj3:RU	трещина, напряжение, грп, модуль, геомех-ский, упругий, трещиноватость
sbj3:EN	fracture, stress, medium, hydraulic, elastic, geomechanical, fracturing
sbj4:RU	приток, колонна, установка, оборудование, заканчивание, песок, труба
sbj4:EN	sand, esp, completion, inflow, pump, failure, tubing
sbj5:RU	кислотный, воздействие, отложение, кислота, залежь, разрез, карбонатный
sbj5:EN	acid, treatment, stimulation, carbonate, acid treatment, pilot, seismic
sbj6:RU	грп, трещина, микросейсмический, мгрп, проппант, порт, гнкт
sbj6:EN	fracturing, fracture, hydraulic, hydraulic fracturing, proppant, microseismic, port
sbj7:RU	потеря, интегрировать, сбор, расход, управление, нагнетательный, ппд
sbj7:EN	pipeline, integrated, flow rate, condensate, unit, gathering, reservoir pressure
sbj8:RU	раствор, буровой, колонна, буровой раствор, геомех-ский, строит-во, риск
sbj8:EN	mud, casing, geomechanical, risk, weight, history, stress
sbj9:RU	неопред-сть, оторочка, рабочий, залежь, адаптация, конденсат, вытеснение
sbj9:EN	uncertainty, composition, history, condensate, mineral, sample, assessment
nz0:RU	образец, керн, эксперимент, раствор, частица, фильтрация, заводнение
nz0:EN	sample, core, cement, experiment, filtration, pore, strength
nz1:RU	образец, керн, карбонатный, пора, смачиваемость, гис, поровый
nz1:EN	pore, core, sample, carbonate, logging, wettability, space

Глава 5. Выводы

В последние годы вопрос о том, по какой траектории происходит развитие нефтегазового комплекса, как и всей энергетической системы, приобретает все больший интерес, как со стороны экспертов, так и со стороны широкой общественности [311; 312]. Этому способствует несколько факторов.

- Во-первых, темпы экономического развития приводят к значительному росту мирового энергопотребления. Как отмечается в докладе Аналитического Центра при Правительстве РФ , значительный рост потребления энергоресурсов происходит за счёт развивающихся стран, преимущественно Азиатско-тихоокеанского региона, в то время как в развитых странах объем выработки электроэнергии стабилен, а динамика потребления схожа с тенденциями общеэкономических приростов и спадов.
- Во-вторых, наблюдается изменение структуры запасов углеводородов. Как отмечается в «Энергетической стратегии России на период до 2035 года» (сформулированной в 2015 году), отечественная нефтяная отрасль сталкивается с такой проблемой, как «увеличение себестоимости добычи вследствие преобладания труднодоступных запасов нефти (да-

лее по тексту ТРиЗ) и большой выработанности действующих месторождений, что усложняет удержание достигнутых уровней добычи нефти». При этом одной из задач, ставящейся перед нефтяным сектором, является освоение ТРиЗ в объёмах до 17% от общей добычи, которая может быть решена путём развития добывающих технологий.

– Наконец, в-третьих, всё большую роль в энергетическом секторе играют источники возобновляемой энергии (т.н. ВИЭ), что сказывается на структуре энергетических рынков . Эксперты, политики и граждане всё больше озабочены экологическими и климатическими вызовами, что свидетельствует о необходимости диверсификации энергоносителей. Дополнительно стоит отметить негативное влияние внешних экономических и политических ограничений на сырьевой сектор российской экономики.

Таким образом, энергетическая сфера находится в процессе постоянной трансформации [190], а одним из важных вопросов повестки дня нефтяного сообщества является оптимизация методов геологоразведки [313], добычи и использования энергоносителей.

Анализировать, по какой траектории движется изменение научно-технических и технологических процессов нефтедобычи, можно несколькими способами [314]. Наиболее очевидным видится опрос экспертов, специализирующихся на вопросах добычи.

Методы экспертных опросов (также называемые методами экспертных оценок) широко используются в различных исследованиях, в которых невозможны или труднодоступны другие формы исследований ввиду отсутствия объективных данных. Таким образом реализуется подавляющее большинство форсайт-исследований. К достоинствам экспертного опроса можно отнести их относительную простоту и доступность, а также возможность применения в случае отсутствия информации об изучаемом явлении.

В то же время очевидным недостатком экспертного опроса являются возможный субъективизм и ограниченность экспертов, их приверженность определенной точке зрения. Как отмечается в работе Бахтина с соавторами [311], в течение последних лет объёмы экспертно-аналитической и научной литературы, а также информации в целом, стремительно растут (по некоторым оценкам объёмы информации удваиваются каждые два года), так что задача получения, фильтрации, переработки и рефлексивного восприятия всей информации становится фактически невозможной. При этом эксперту необходимо развиваться и совершенствоваться в различных содержательных направлениях, что требует ещё больших трудовых и временных инвестиций. Это свидетельствует о необходимости разработки и формирования дополнительной обратной связи, которая призвана помочь экспертному и профессиональному сообществу анализировать огромные объемы информации и

выделять из нее наиболее релевантные аспекты, в частности – выявлять технологические тренды.

С развитием автоматизированных методов обработки неструктурированных данных, в частности текстовых данных, популярность набирает тематическое моделирование научных текстов [107]. Как было продемонстрировано в работе Блея и Лафферти, тематическое моделирование оказывается перспективным инструментов отслеживания трендов в таких научных направлениях, как ядерная физика и нейронауки [315], технологии агропромышленного комплекса [311] и так далее. Изучение автоматически выделенных тематик во временной перспективе иллюстрирует изменение интереса научного сообщества к различным объектам и предметам исследования. Достоинством этого метода является возможность автоматизированной обработки огромных массивов информации и выявления латентных (скрытых) тематик текстов. При этом тематическое моделирование нельзя назвать исключительно автоматизированным методом, так как полученные в результате машинной классификации тематики в дальнейшем должны быть вручную просмотрены и проработаны экспертами-специалистами предметной области. Таким образом, тематическое моделирование может рассматриваться как метод, заключающий в себе достоинства и автоматизированной обработки текста, и экспертной оценки. Реализация подобного метода в приложении к различным содержательным задачам позволит сформировать диалог между наукой

и стратегией на принципиально новом уровне. Тематическое моделирование позволяет оперативно обрабатывать значительные объемы текстов для сужения найденных понятий до небольших значимых фрагментов текста - топиков. Каждый топик представляется набором слов и от качества этого представления зависит возможная интерпретация.

Автор показал результативность подхода к улучшению интерпретируемости тематик на основе последовательной регуляризации.

Примененные методы управление отношением «плотность-разрежённость» открывают возможности настройки модели на предметную область текстов. Автор показал принципы создания и настройки модели тематик, которые позволяют вести интеллектуальный поиск (разведку) высоко сфокусированныхисточников знаний.

Кластеризация топиков была проверена с помощью двух методов для векторизации слов (FastText, GloVe) и двух методов для уменьшения размерности векторного пространства (TSNE, MDS). Результаты представлены в виде диаграмм и уверено показывают наличие кластеров.

Подход к анализу текстовой информации на основе моделирования тематик теперь широко используется во внутренних процессах компании ООО «Газпромнефть НТЦ» для оптимизации процессов управления знаниями, выявления наиболее перспективных направлений исследований и поиска opinion leaders в определенных научных направлениях.

Важно отметить, что выбранный автором метод показал высокую скорость анализа, что делает его возможным для применения в онлайновых процессах поиска. Например, на сайте издательства в качестве средства улучшающего поиск и дающего рекомендации читателям по статьям со схожей тематикой.

Также следует отметить, что разработанная методика может быть в дальнейшем усовершенствована и адаптирована для анализа существенно больших массивов динамических данных и выделения ключевых направлений технологического развития как в более широких, так и в более узких областях.

Существующие прогнозы научно-технического развития (с том числе форсайт-прогнозы) в большинстве своем экстраполируют существующие тренды на долгосрочную перспективу. Таким образом, большой интерес приобретают работы, в которых становится возможным выявление новых технологических направлений, способных существенно видоизменить структуру рынков.

Сами по себе отдельные технологии не следует рассматривать как оторванные и изолированные друг от друга инициативы. В действительности многие технологические направления развиваются параллельно, что является результатом венчурной политики, технологического развития и других сопутствующих факторов.

Ввиду этого важным направлениям анализа технологических трендов выглядит изучение коэволюции развития

сразу нескольких технологий [23]. Именно изучение совокупности научно-технических инициатив позволит содержательно проанализировать направление технологического развития.

В разделе 4.2 автором представлен новый взгляд на процесс публикации научных статей. Определены показатели продуктивности и стратегии управления продуктивностью процесса публикаций.

Организационная среда должна служить инструментом для повышения эффективности основных производственных процессов. Признание научно-исследовательской организацией того факта, что публикация научных статей является одним из основных производственных процессов означает, что необходимо создавать специальные подразделения, нацеленные на поддержку эффективности этого процесса. Мерой зрелости процесса служит степень разделения труда его участников. Учёный должен заниматься своими прямыми обязанностями - исследованиями и не обязан вникать в детали процессов оформления командировок, эргономичности презентаций и тонкостей общения с издателями и т.п.

Автором разработана ролевая модель, которая позволит разгрузить учёных от формальных трудозатрат по публикации результатов исследований и в некоторых случаях избежать появления «гостевых» соавторов.

Из-за ограничения по объёму публикаций в выпуске издателя организациям необходимо расширять список изда-

тельств, в которых публикуются их исследователи, чтобы поддерживать темп роста количества опубликованных статей.

Показатель продуктивности выражающий долю отвергнутых издательством статей является важной характеристикой процесса публикации результатов исследований не только на организационном, но и на отраслевом уровне. Возможность анализа этого показателя позволяет оценить достаточность ёмкости рынка научных издательств и степень конкуренции за публикацию в изданиях с высоким импакт-фактором.

В разделе 4.5 автором обобщена и проработана формализация процесса самоорганизации команд для достижения определённой цели – написания научных статей. В исследовании разработан детальный алгоритм образования графа соавторств широко используемого в различных исследованиях. Сформулированы основные теоретические утверждения, даны определения *укомплектованности команды* и *несостоявшейся научной статьи*. Сформулирована гипотеза (Гипотеза 3.10.3) об инвариантность графа соавторства относительно введения Scrum ролей в процесс написания статей. В результате проведённого автором оптимизационного эксперимента найдены оптимальные значения параметров для построенной модели написания статей. По результатам, сделанным на оптимизированной модели соавторства разработанной автором, эффект от введения гибких методик

(Scrum) в процесс написания научных статей небольшими командами соавторов состоит в следующем:

- Среднее время написания научной статьи (T_{pub}) не изменяется
- Средняя доля несостоявшихся научных статьей ($Frac_{notpub}$) уменьшается

Общее влияние Scrum на процесс написания научных статей командой соавторов является положительным. То, что T_{pub} не изменяется может служить экспериментальным подтверждением Гипотезы 3.10.3.

Продуктивность команд, образованных по комплементарному принципу, становится выше от применения гибких методик и Scrum, в частности.

В эксперименте описанном в разделе 4.9 подтверждена гипотеза о возможности выделения эмоционально-окрашенных фрагментов текста из научных статей. Научные статьи используют академическую лексику и ожидать в них градус эмоций сравнимый с отзывами на кинофильмы было бы наивно. Но современные концепции обработки текста, основанные на анализе контексте, позволяют выделять и классифицировать изменения эмоциональности достаточно точно для того, чтобы обрабатывать даже научные статьи. Автор считает, что проведённое исследование открывает возможности по созданию дополнительных инструментов для аннотации и классификации научных текстов.

Наилучшее оценку по качеству выделения эмоционально окрашенных фрагментов текста показали рекуррентные

нейронные сети. Точность по метрике Accuracy для них составила 88%. Важно отметить, что по скорости обучения рекуррентные сети существенно проигрывают свёрточным сетям. Автор видит объяснение разности в производительности в том, что для обучения сверточных нейронных сетей возможна более высокая степень параллельных вычислений. Тогда как для рекуррентных нейронных сетей необходимо поддерживать последовательность предыдущих состояний нейронов.

В дальнейших исследованиях автор планирует исследовать применимость эмоционально окрашенных фрагментов текста для задач классификации текстов в качестве признаков. Так же на взгляд авторов, научный интерес представляет анализ синтаксиса эмоционально окрашенных фрагментов текста.

В эксперименте описанном в разделе 4.6.1 проведён анализ динамики графа соавторства для одной организации на основании публичных данных о публикациях. Основным аналитическим инструментом был выбран двудольный граф соавторства, методически обоснованный автором в разделе 3.9.

В работе применён много компонентный подход к прогнозированию изменению свойств графа соавторства. Анализ малых связанных компонент позволил выявить их долю в ежегодном увеличении количества авторов. Отметим, что доля малых компонент в рассматриваемом графе

соавторства уменьшается со временем, что является структурным ограничением роста рассматриваемой организации.

В графе соавторства за 2016 год обнаружен «Эффект локтя» - резкое усложнение характера роста графа соавторства по годам. Автором сделан прямой прогноз роста на основании тренда роста авторов по годам и уточненный прогноз роста графа соавторства на основе моделирования с помощью методов машинного обучения.

Проведённое сравнение точности классификаторов определило классификатор на основе нейронной сети как наиболее точный для данной задачи.

Прогноз, сделанный на основе модели, показал результат (467) существенно меньший чем результат на основе тренда (585).

В результате проведённого исследования автор сделал вывод о наличии в структуре графа соавторств важной информации о развитии графа соавторств, которая определяет прогноз роста. Что позволяет определить значимые признаки образования новых коллабораций, а также регрессионного предсказания новых связей между уже сформировавшимися исследовательскими коллективами.

Использование методов векторизации графовых моделей в комбинации с извлечением признаков позволит улучшить точность предсказания появления новых связей, а также качественно измерить публикационную активность на основе публично доступных метрик журналов и конференций .

Так же автором предложен метод выделения направлений научных исследований на основе графа соавторства. Содержательно предложенный метод относится к top-down алгоритмам кластеризации. В качестве критерия выделения кластеров выбрана метрика *Betweenness centrality*.

В качестве критерия проверки качества кластеров выбрана метрика близости членов кластера и метрика удалённости различных кластеров на основе тематик научных статей, входящих в граф соавторства.

Результатом применения предложенного метода является укрупненное виденье научных направления развития организации, сделанное на основе публичных данных о публикационной активности сотрудников.

Разработанный автором метод выделения направлений научных исследований на основе графа соавторства опробован на Научно-техническом центра ГазпромНефть. В результате выделены 16 кластеров, характеризующих деятельность организации. Важными особенностями разработанного метода выделения направлений научных исследований на основе графа соавторства являются следующие:

- Рекурсивность алгоритма позволяет работать с графами различных порядков.
- «Жадный» алгоритм определения качества кластеров позволяет корректировать оптимизацию на каждом шаге.
- Применение двудольного построения графа соавторства позволяет анализировать различные проекции.

– Работа на основании публичных данных даёт широкие возможности для применения в бизнес разведке.

Новизна предложенного автором метода выделения направлений научных исследований на основе графа соавторства состоит в использовании двудольного построения графа соавторства и в динамической модели кластеризации, использующей структурные метрики графа соавторства и метрики близости текстов научных статей.

Автор создал действующие модели движения персонала в организации и модель выполнения заданий. На основе взаимодействия этих моделей автор построил модель продуктивности, которая, отражает для научно-исследовательской организации изменения ИК. Экспериментальные результаты представлены в разделе 4.3.

Согласно мнению многих исследователей ИК сложно измерить. Предложенный автором драйвер ИК в виде производительности научно-исследовательской организации имеет самостоятельную ценность и характеризует ИК, как комплексный показатель организации.

Автор построил зависимости ИК от различных времён адаптации новичков и различной сложности поступающих заданий, показали асимптотическое поведение ИК. Что позволяет моделировать ситуации разных видов задач, особенностей организации (текучесть, скорость адаптации, сложность задач и пр).

В исследовании проанализировано как на ИК влияет нагрузка на персонал. Показано как со временем уменьшает-

ся продуктивность при высоких нагрузках и необходимости работать дольше 40 часов в неделю. Автором смоделирован эффекты «выгорания» и «усталости» персонала от длительной высокой нагрузки.

Результаты, полученные в исследовании, обладают научной новизной и практической ценностью, дают возможность детального исследования и моделирования динамики продуктивности.

Созданная автором и описанная в разделе 4.4 частная модель $\mathbb{M}_{GPN}$ организации оправдала себя как метод исследования социальных явлений и процессов организации посредством их воспроизведения в менее сложных формах и проведения необходимых операций с полученными таким образом аналогами реальных процессов в организационной среде.

Формальная математическая модель $\mathbb{M}_{\Omega}$ организации дает ответы на вопросы о ключевых компонентах деятельности по написанию и публикации научных статей.

Выбранный автором метод создания частной модели $\mathbb{M}_{GPN}$ показал результаты согласующиеся с эмпирическим исследованием публикаций конкретной научно-исследовательской организации.

Был проведен эксперимент по многоагентному симулированию, в котором в качестве агентов выступали научные сотрудники лабораторий, взаимодействующие друг с другом и производящие в качестве результата своей работы научные статьи. Создана частная модель $\mathbb{M}_{GPN}$ путем

калибровки на данных НТЦ «Газпромнефть». В работе использовано программное обеспечение Anylogic.

На основании созданной частной модели автор пришел к необходимости дальнейшего изучения чувствительности от следующих свободных параметров:

1. Максимальное количество соавторов
2. Среднее количество соавторов в статьях
3. Распределение количества соавторов
4. Количество статей в год на одного сотрудника

Полученные в результате симуляционного эксперимента результаты согласуются с эмпирическими наблюдениями. Исходя из этого, автором сделан вывод о том, что работа исследователей может быть смоделирована с использованием агентного подхода. Решение подобной задачи является важным шагом на пути к идентификации механизмов коллективной работы и формирования коллективного наукоемкого продукта.

Процессы контроля качества и управления содержанием научного журнала представляют важное направление деятельности для редакционных коллегий. Существует множество факторов, влияющих на контент научного журнала. Автор в работе выделены факторы жизненного цикла и факторы научной экосистемы.

Предложенная авторами методика T4C является удобным подручным инструментом для выявления глубинных свойств коллекций текстов. Отличительная особенность ме-

тодики T4C состоит в том, что она позволяет сравнивать коллекции текстов [316].

Для проверки методики T4C авторами был выбран экспериментальный подход. Минимально необходимые аналитические выкладки сделаны авторами в разделе описывающем методику T4C. В настоящей работе авторами приведено экспериментальное подтверждение методики в части гипотезы о классификации и модальной тематической модели. Метрики, связанные с авторами в данном эксперименте не вычислялись.

Патенты являются хорошим аналогом научных статей, имеют четкую временную динамику и языковые особенности. А в смысле временной миграции контента должны вести себя лучше научных статей. Кроме того, в результате эксперимента было выяснено, что тексты заявок на патенты содержат культурные особенности.

В результате применения T4C к коллекции слабо структурированных текстов были получены следующие результаты:

1. Гипотеза о классификации подтверждена с точностью 97.5%. Китайский (Американский) патент может быть идентифицирован с указанной точностью на основании текста описания. Анализ величин коэффициентов модели позволяет сделать заключение о характерных особенностях описаний. Американские патенты содержат больше картинок, что отражает визуальную национальную

особенность. Китайские патенты содержат больше шумовых тематик (опечаток, редких идиом, устаревших оборотов).

2. Тематическая однородность оценена визуально, с помощью карт "тематика–документ"для выборки Американских и Китайских патентов.
3. Метрика тематической чистоты (Purity) 4.48 показывает, что разработанная авторами последовательность регуляризации позволяет выделять шумовые и основные тематики.
4. Классификация коллекций тематик по странам (CN,US) позволяет достичь точности 60%, что находится в согласии с визуальной оценкой.
5. Корреляции между патентами из разных лет (2010 vs. $<$2004) составляют 85% и 67% для Американских и Китайских патентов, соответственно. Данный факт показывает эволюцию патентных заявок во времени. И может качественно характеризовать скорость изменений.
6. Визуальная оценка слов с наибольшим весом из разных лет согласуется с вычисленными значениями корреляции.

На основании проведённого эксперимента автор сделал следующие качественные выводы и наблюдения:

– Обучение модели с весами модальностей (1:3:3) позволяет сделать заключение о характерных научных направлениях для каждой страны.

– Для Китая наибольшую значимость имеют патенты направленные на изучение землетрясений .
– В то время как для Америки патентование в области сейсмики идёт более равномерно по всем тематикам.

Автор исследовал структурные отличия научных статей, возникающие при переводе с русского на английский. Для своих исследований автор использовал методику модального тематического моделирования. В собранной коллекции каждый документ был представлен двумя модами, на английском и русском языках. В результате построения тематической модели были получены бимодальные матрицы Φ и Θ. Анализ матрицы Φ показал, что тематики разделились по степени соответствия между русскими и английскими терминами при рассмотрении слов в порядке убывания вероятности. Для 90% тематик английские слова полностью соответствовали русским. Анализ матрицы Θ показал, что для 99% документов существует тематика со значением больше *0.95*. Таким образом большинство документов являются монотематичными. И это большинство не зависит от языка документа.

Особенностью данного исследования является количество документов - 484 документа, по 242 на каждом языке. Размер общего словаря составил после удаления высокочастотных и редких слов 10^5 термов для биграмной модели. Выбор такого корпуса документов обусловлен тематической сфокусированностью. Все статьи согласно рубрикатору сосредоточены на одном научном направлении: гидродина-

мический разрыв пласта. Новизна данного исследования заключается в распространении стратегий аддитивной регуляризации тематической модели в направлении машинного перевода.

Проделанная авторами работа позволяет с уверенностью сказать, что разработанная методика T4C может быть успешна применена для решения бизнес задач по сравнению коллекций слабо структурированных текстов.

Автором исследован двуязычный корпус документов. Основная цель исследования состояла в выделении особенностей перевода, которые по гипотезе авторов состояли в разных подходах к переводам текстов с русского на английский. В случае использования автором для перевода средств СМП соответствие между английскими и русскими словами в тематиках должно быть более точным, чем при авторском переводе.

Высказанная автором гипотеза была подтверждена проведенным экспериментом: Переводная статья написана на «русском английском» языке может быть точно идентифицирована на основе тематической модели.

В исследовании обнаружена кластеризация тематик аддитивной мультимодальной модели для корпуса двуязычных документов. Документы с типом перевода «русский английский» отличимы по значениям φ_{wt} и θ_{td} от документов с более творческим подходом к переводу.

Методика, предложенная автором в данной статье, подтверждает возможности использования стратегий

регуляризации тематических моделей для получения компактных представлений тематик, подчеркивающих определенные характеристики исследуемой коллекции документов.

Полученный автором результат может быть использован для автоматизированного определения качества перевода научных текстов.

Автором исследован вопрос выбора оптимального количества тематик для построения тематической модели для заданного корпуса текстов. Результатом данного исследования стала методика, позволяющая определить оптимальное количество тематик для небольшого корпуса однородных англоязычных документов.

Важным методическим приёмом автора является подготовка тематической модели с помощью последовательной регуляризации. При формировании коллекции текстов были заданы условия, ограничивающие число тематик научных статей согласно тематическому рубрикатору до 10. Суть эксперимента состояла в том, чтобы подтвердить выбранное число тематик с помощью оптимизационного подхода на основе разработанной авторами метрики качества тематической модели — *cDBI*. В результате эксперимент показал, что максимальное значение метрики *cDBI* достигается при количестве тем равному 16. Данный результат получен при много прогонном обучении модели, чтобы исключить эффект зависимости от порядка документов в коллекции.

Результаты исследования

1. Автором представлен новый взгляд на процесс публикации научных статей. Определены показатели продуктивности и стратегии управления продуктивностью процесса публикаций.
2. Автором обобщена и проработана формализация процесса самоорганизации команд для достижения определённой цели – написания научных статей.
3. В исследовании разработан детальный алгоритм образования соавторств, широко используемого в различных исследованиях.
4. Сформулированы основные теоретические утверждения, даны определения *укомплектованности команды* и *несостоявшейся научной статьи*.
5. Сформулирована и подтверждена гипотеза об инвариантности графа соавторства относительно введения Scrum ролей в процесс написания статей.
6. В результате проведённого автором оптимизационного эксперимента найдены оптимальные значения параметров для построенной модели написания статей.

   По результатам, сделанным на оптимизированной модели соавторства, разработанной автором, эффект от введения гибких методик (Scrum) в процесс

написания научных статей небольшими командами соавторов состоит в следующем:

- Среднее время написания научной статьи (T_{pub}) не изменяется
- Средняя доля несостоявшихся научных статьей ($Frac_{notpub}$) уменьшается

Общее влияние Scrum на процесс написания научных статей командой соавторов является положительным. Экспериментально подтверждена гипотеза автора о том, введение ролей Scrum в процесс соавторства не изменяет вид графа соавторства.

Продуктивность команд, образованных по комплементарному принципу, становится выше в результате применения гибких методик и, в частности, Scrum.

7. Подтверждена гипотеза о возможности машинного выделения эмоционально-окрашенных фрагментов текста из научных статей. Научные статьи используют академическую лексику и ожидать в них градус эмоций, сравнимый с отзывами на кинофильмы, было бы наивно. Но современные концепции обработки текста, основанные на анализе контекста, позволяют выделять и классифицировать изменения эмоциональности;этого достаточно для того, чтобы обрабатывать даже научные статьи.

8. Автор показал, что проведённое исследование открывает возможности по созданию дополнительных инструментов для аннотации и классификации научных текстов.
9. Автором проведён анализ динамики графа соавторства для одной организации на основании публичных данных о публикациях. Основным аналитическим инструментом был выбран двудольный граф соавторства, методически обоснованный автором.
10. В работе применён многокомпонентный подход к прогнозированию изменения свойств графа соавторства. Анализ малых связанных компонентов позволил выявить их долю в ежегодном увеличении количества авторов. Отметим, что доля малых компонентов в рассматриваемом графе соавторства уменьшается со временем, что является структурным ограничением роста рассматриваемой организации.
11. При исследовании временной зависимости структуры соавторств был обнаружен «Эффект локтя» - резкое усложнение характера роста графа соавторства по годам. Автором сделан прямой прогноз роста на основании тренда роста авторов по годам и уточненный прогноз роста графа соавторства на основе моделирования с помощью методов машинного обучения. Проведённое сравнение точности клас-

сификаторов определило классификатор на основе нейронной сети как наиболее точный для данной задачи.

12. В результате проведённого исследования автор сделал вывод о наличии в структуре графа соавторств важной информации о развитии этого графа, которая определяет прогноз роста. Это позволяет определить значимые признаки образования новых коллабораций, а также регрессионного предсказания новых связей между уже сформировавшимися исследовательскими коллективами.
13. Так же автором предложен метод выделения направлений научных исследований на основе графа соавторства. Содержательно-предложенный метод относится к top-down (нисходящим) алгоритмам кластеризации. В качестве критерия выделения кластеров выбрана метрика *Betweenness centrality*. В качестве критерия проверки качества кластеров выбрана метрика близости членов кластера и метрика удалённости различных кластеров на основе тематик научных статей, входящих в граф соавторства.
14. Результатом применения предложенного метода является укрупненное виденье научных направления развития организации, сделанное на основе публичных данных о публикационной активности сотрудников.

15. Разработанный автором метод выделения направлений научных исследований на основе графа соавторства опробован на в Научно-техническом центре Газпромнефть . В результате апробации выделены 16 кластеров, характеризующих деятельность организации.
16. Важными особенностями разработанного метода выделения направлений научных исследований на основе графа соавторства являются следующие:
    - Рекурсивность алгоритма позволяет работать с графами различных порядков.
    - «Жадный» алгоритм определения качества кластеров позволяет корректировать оптимизацию на каждом этапе.
    - Применение двудольного построения графа соавторства позволяет анализировать различные проекции.
    - Работа на основании публичных данных даёт широкие возможности для применения в бизнес разведке.

    Новизна предложенного автором метода выделения направлений научных исследований на основе графа соавторства состоит в использовании двудольного построения графа соавторства и в динамической модели кластеризации, использующей структурные метрики графа соавторства и метрики близости текстов научных статей.

17. Автор создал действующие модели движения персонала в организации и модель выполнения заданий. На основе взаимодействия этих моделей автор построил модель продуктивности, которая отражает изменения ИК для НИ организации.
18. Автор построил зависимости ИК от различных сроков адаптации новичков и различной сложности поступающих заданий, показали асимптотическое поведение ИК, что позволяет моделировать ситуации разных видов задач, особенностей организации (текучесть, скорость адаптации, сложность задач и пр.).
19. В исследовании проанализировано, как на ИК влияет нагрузка на персонал. Показано как со временем уменьшается продуктивность при высоких нагрузках и необходимости работать дольше 40 часов в неделю. Автором смоделирован эффекты «выгорания» и «усталости» персонала от длительной высокой нагрузки.
20. Созданная автором частная модель $\mathbb{M}_{GPN}$ организации оправдала себя как метод исследования социальных явлений и процессов организации посредством их воспроизведения в менее сложных формах и проведения необходимых операций с полученными таким образом аналогами реальных процессов в организационной среде. Формальная математическая модель $\mathbb{M}_{\Omega}$ организации дает отве-

ты на вопросы о ключевых компонентах деятельности по написанию и публикации научных статей.

21. Выбранный автором метод создания частной модели $\mathbb{M}_{GPN}$ показал результаты, согласующиеся с эмпирическим исследованием публикаций конкретной НИ организации.
22. Был проведен эксперимент по многоагентному симулированию, в котором в качестве агентов выступали научные сотрудники лабораторий, взаимодействующие друг с другом и производящие в качестве результата своей работы научные статьи.
23. Помимо общей математической модели НТЦ $\mathbb{M}_{\Omega}$ автором была создана частная модель $\mathbb{M}_{GPN}$ путем калибровки общей модели на данных НТЦ Газпромнефть .

    Для создания компактного описания компьютерного эксперимента для прогнозирования эффективности НТЦ на основании частной модели $\mathbb{M}_{GPN}$ автором было использовано программное обеспечение Anylogic. На основании созданной частной модели автор выделил следующие свободные параметры:

    а) Максимальное количество соавторов
    б) Среднее количество соавторов в статьях
    в) Распределение количества соавторов
    г) Количество статей в год на одного сотрудника

24. Полученные в результате симуляционного эксперимента [317; 318] результаты согласуются с эмпирическими наблюдениями. Исходя из этого, автором сделан вывод о том, что работа исследователей может быть смоделирована с использованием агентного подхода. Решение подобной задачи является важным шагом на пути к идентификации механизмов коллективной работы и формирования коллективного наукоемкого продукта [319].

В заключение автор выражает благодарность и большую признательность научному консультанту Дегтяреву А.Б. за поддержку, помощь, обсуждение результатов и научное руководство. Также автор благодарит Хасанова М.М. за плодотворное научное сотрудничество и поддержку научных идей автора.

Список сокращений и условных обозначений

W	Словарь, уникальное множество слов.
D	Коллекция документов.
n_d	Количество документов в коллекции.
n_w	Количество слов в словаре.
n_dw	Количество слов документе.
w_i	Слово словаря с порядковым номером i.
d_i	Документ коллекции с номером i.
T	Набор тематик.
t_i	Тематика из набора тематик с номером i
Θ	Матрица “тематики–документы” с размерностью TxD.
θ_{td}	Элемент матрицы Θ для тематики t и документа d, число от 0 до 1.
Φ	Матрица “слова–тематики” с размерностью WxT.
φ_{wt}	Элемант матрицы Φ для слова w и тематики t, число от 0 до 1.
$\mathbb{M}_\Omega$	Модель социальной системы
MAP	Принцип максимизации апостериорной вероятности
MLE	Метод оценки максимального правдоподобия
DC	Degree centrality
BC	Betweenness centrality
CC	Closeness centrality

HC	Harmonic centrality
CN	Common Neighbours
SI	Salton Index
JI	Jaccard Index
HPI	Hub Promoted Index
HDI	Hub Depressed Index
LHN1	Leicht-Holme-Newman Index
PA	Preferential Attachment Index
AAI	Adamic-Adar Index
RAI	Resource Allocation Index
ПКК	Полный Командный Код
ОКК	Остаточный Командный Код
LDA	Latent Dirichlet Allocation
$\mathcal{KL}$	Дивергенция Кульбака-Лейблера
TSNE	t-distributed Stochastic Neighbor Embedding
DBI	Davies Bouldin Index
cDBI	Cosine Davies Bouldin Index
GloVe	Global Vectors representation
FastText	Библиотека для получения плотных векторных представлений текста

Словарь терминов

Maximum probability of improvement (MPI) : $\mu(x) = P(\hat{f}(x) \geqslant f^* + \varepsilon = \Phi\left(\frac{\mathbb{E}\hat{f}(x) - f^* - \varepsilon}{Var[\hat{f}(x)]}\right)$

Upper confidence bound (UCB) : $\mu(x) = \mathbb{E}\hat{f}(x) + \eta Var[\hat{f}(x)]$

Expected improvement (EI) : $\mu(x) = \mathbb{E}\max(f(x) - f^*,0) = Var[\hat{(}f(x)] \cdot [z\Phi(z) + \varphi(z)]$, где $z = \frac{\mathbb{E}\hat{f}(x) - m(x)}{Var[\hat{f}(x)]}$

Scrum : Одна из распространённых гибких методик командной работы является. Scrum предназначен для получения наилучших из возможных результатов для командной разработки сложных интеллектуальных продуктов.

Несостоявшейся научной статьей (ННС) : Несостоявшейся будем считать статью, не уложившуюся во временные рамки публикационного процесса с требуемым качеством.

Косинусная мера расстояния (Cosine) : $\frac{v_1 \cdot v_2}{\|v_1\|_2 * \|v_2\|_2}$

Евклидово расстояние (Euclidean) : $\|v_1 - v_2\|_2$

ROC AUC : "кривая ошибок". График, позволяющий оценить качество бинарной классификации, отображает соотношение между долей объектов от общего количества носителей признака, верно классифицированных как несущих признак и долей объектов от общего количества объектов, не несущих признака, ошибочно классифициро-

ванных как несущих признак при варьировании порога решающего правила.

T4C : Авторская методика для выявления глубинных свойств коллекций текстов.

Список литературы

2. *Nadhan, D.* Drilling with Digital Twins / D. Nadhan, M. G. Mayani, R. Rommetveit // IADC/SPE Asia Pacific Drilling Technology Conference and Exhibition. — Bangkok, Thailand : Society of Petroleum Engineers, 2018. — C. 18. — ISBN 978-1-61399-574-7. — DOI: 10.2118/191388-MS. — URL: https://doi.org/10.2118/191388-MS.

4. Drilling Automated Realtime Monitoring Using Digital Twin / M. Gholami Mayani, R. Rommetveit, S. I. Oedegaard, M. Svendsen. — Abu Dhabi, UAE, 2018. — DOI: 10.2118/192807-MS. — URL: https://doi.org/10.2118/192807-MS.

5. *Van Os, J.* The Digital Twin throughout the Lifecycle / J. Van Os. — Providence, Rhode Island, USA, 2018. — URL: https://doi.org/.

6. *Follesdal Tjonn, A.* Digital Twin Through the Life of a Field / A. Follesdal Tjonn. — Abu Dhabi, UAE, 2018. — DOI: 10.2118/193203-MS. — URL: https://doi.org/10.2118/193203-MS.

8. *Poddar, T.* Digital Twin Bridging Intelligence Among Man, Machine and Environment / T. Poddar. — Kuala Lumpur, Malaysia, 2018. — DOI: 10.4043/28480-MS. — URL: https://doi.org/10.4043/28480-MS.

9. Accelerating Well Construction Using a Digital Twin Demonstrated on Unconventional Well Data in North America / G. Saini, P. Ashok, E. van Oort, M. R. Isbell. — Houston, Texas, USA, 2018. — DOI: 10 . 15530 / URTEC - 2018 - 2902186. — URL: https : //doi.org/10.15530/URTEC-2018-2902186.

10. RB-FEA Based Digital Twin for Structural Integrity Assessment of Offshore Structures / P. Sharma, D. Knezevic, P. Huynh, G. Malinowski. — Houston, Texas, USA, 2018. — DOI: 10.4043/29005-MS. — URL: https://doi.org/10.4043/29005-MS.

11. Managing the flow of technology: Technology transfer and the dissemination of technological information within the R&D organization / T. J. Allen [и др.] // MIT Press Books. — 1984. — Т. 1.

12. Human resource management / R. A. Noe, J. R. Hollenbeck, B. Gerhart, P. M. Wright. — China People's University Press, 2006.

13. *Cooley, C. H.* Social organization / C. H. Cooley. — Transaction Publishers, 1956.

15. *Block, P.* Multidimensional homophily in friendship networks / P. Block, T. Grund // Network Science. — 2014. — Т. 2, № 2. — С. 189—212.

16. *De Nooy, W.* Exploratory social network analysis with Pajek / W. De Nooy, A. Mrvar, V. Batagelj. — Cambridge University Press, 2018.

17. *Moreno, J. L.* Who shall survive? Foundations of sociometry, group psychotherapy and socio-drama / J. L. Moreno. — 1953.

18. *Mullins, N. C.* The development of specialties in social science: The case of ethnomethodology / N. C. Mullins // Science Studies. — 1973. — Т. 3, № 3. — С. 245—273.

19. Link prediction in co-authorship networks based on hybrid content similarity metric / P. M. Chuan, M. Ali, T. D. Khang, N. Dey [и др.] // Applied Intelligence. — 2018. — Т. 48, № 8. — С. 2470—2486.

20. Building and analyzing a global co-authorship network using Google Scholar data / Y. Chen, C. Ding, J. Hu, R. Chen, P. Hui, X. Fu // Proceedings of the 26th International Conference on World Wide Web Companion. — International World Wide Web Conferences Steering Committee. 2017. — С. 1219—1224.

37. *Портер, М.* Международная конкуренция: конкурентные преимущества стран / М. Портер. — Альпина Паблишер, 1993.

38. *Тикин, В.* Эффективность-не коэффициент / В. Тикин // Экономические науки. — 2009. — № 7. — С. 94—97.

39. *Клещева, И.* Оценка эффективности научно-исследовательской деятельности студентов / И. Клещева // СПб: НИУ ИТМО. — 2014.

40. *Левин, В.* Возможна ли правильная оценка вклада ученого в науку с помощью индекса хирша? примеры / В. Левин // Математические методы в технике и технологиях-ММТТ. — 2016. — № 6. — С. 100—102.

41. *Липчиу, Н.* Методология научного исследования: учебное пособие / Н. Липчиу, К. Липчиу // Краснодар: КубГАУ. — 2013.

44. *Мкртчян, М.* Фазы переходного периода от группового способа обучения к коллективному / М. Мкртчян // Коллективный способ обучения. — 1995. — № 2. — С. 8—11.

45. *Данилевская, Н.* Оценка как источник динамики текстообразования в научной коммуникации / Н. Данилевская // Международный научно – исследовательский журнал. — 2016. — № 12. — С. 27—30.

47. *Gary, M. S.* Unpacking mental models through laboratory experiments / M. S. Gary, R. E. Wood // System Dynamics Review. — 2016. — Т. 32, № 2. — С. 101—129.

48. *Сидоренков, А.* Групповая сплоченность и неформальные подгруппы / А. Сидоренков // Психологический журнал. — 2006. — Т. 27, № 1. — С. 44—53.

49. *Gentner, D.* Mental models / D. Gentner, A. L. Stevens. — Psychology Press, 2014.

50. *Taylor, F. W.* Scientific management / F. W. Taylor. — Routledge, 2004.

51. *Королева, Т.* Критерии оценки эффективности деятельности научных учреждений / Т. Королева, И. Васильев, И. Торжков // Труды Санкт-Петербургского научно-исследовательского института лесного хозяйства. — 2014. — № 2. — С. 94.

52. *Vonortas, N. S.* New directions for US science and technology policy: the view from the R&D assessment front / N. S. Vonortas // Science and Public Policy. — 1995. — Т. 22, № 1. — С. 19—28.

53. *Veugelers, R.* Collaboration in R&D: an assessment of theoretical and empirical findings / R. Veugelers // De economist. — 1998. — Т. 146, № 3. — С. 419—443.

54. *Faems, D.* Interorganizational collaboration and innovation: Toward a portfolio approach / D. Faems, B. Van Looy, K. Debackere // Journal of product innovation management. — 2005. — Т. 22, № 3. — С. 238—250.

56. *Wasserman, S.* Social network analysis: Methods and applications. Т. 8 / S. Wasserman, K. Faust. — Cambridge university press, 1994.

59. *Kurtz, C. F.* Collective Network Analysis / C. F. Kurtz. — 2009.

60. Analyzing social networks using FCA: complexity aspects / V. Snasel, Z. Horak, J. Kocibova, A. Abraham // Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology-Volume 03. — IEEE Computer Society. 2009. — С. 38—41.

61. Gaining insight in social networks with biclustering and triclustering / D. Gnatyshak, D. I. Ignatov, A. Semenov, J. Poelmans // international conference on business informatics research. — Springer. 2012. — С. 162—171.

62. *Kuznetsov, S.* Reducing the representation complexity of lattice-based taxonomies / S. Kuznetsov, S. Obiedkov, C. Roth // International Conference on Conceptual Structures. — Springer. 2007. — С. 241—254.

63. Semi-automated knowledge discovery: identifying and profiling human trafficking / J. Poelmans, P. Elzinga, D. I. Ignatov, S. O. Kuznetsov // International Journal of General Systems. — 2012. — Т. 41, № 8. — С. 774—804.

64. Formal concept analysis in knowledge processing: A survey on applications / J. Poelmans, D. I. Ignatov, S. O. Kuznetsov, G. Dedene // Expert systems with applications. — 2013. — Т. 40, № 16. — С. 6538—6560.

65. *Aufaure, M.-A.* Advances in FCA-based applications for social networks analysis / M.-A. Aufaure, B. Le Grand // International Journal of Conceptual Structures and

Smart Applications (IJCSSA). — 2013. — Т. 1, № 1. — С. 73—89.

66. *Obiedkov, S.* Social Network Analysis and Conceptual Structures: Exploring Opportunities: Proceedings, Clermont-Ferrand, France, February 2007 / S. Obiedkov, C. Roth. — Universite Blaise Pascal, Laboratoire Limos, 2007.

67. *Pensa, R. G.* Towards fault-tolerant Formal Concept Analysis / R. G. Pensa, J.-F. Boulicaut // Congress of the Italian Association for Artificial Intelligence. — Springer. 2005. — С. 212—223.

68. *Klimoski, R.* Team mental model: Construct or metaphor? / R. Klimoski, S. Mohammed // Journal of management. — 1994. — Т. 20, № 2. — С. 403—437.

69. *Morgeson, F. P.* The structure and function of collective constructs: Implications for multilevel research and theory development / F. P. Morgeson, D. A. Hofmann // Academy of management review. — 1999. — Т. 24, № 2. — С. 249—265.

70. *Walsh, J. P.* Managerial and organizational cognition: Notes from a trip down memory lane / J. P. Walsh // Organization science. — 1995. — Т. 6, № 3. — С. 280—321.

71. *Fiske, S. T.* Social cognition: From brains to culture / S. T. Fiske, S. E. Taylor. — Sage, 2013.

72. *Sims, H. P.* The thinking organization / H. P. Sims, D. A. Gioia. — Jossey-Bass Inc Pub, 1986.

73. *Brief, A. P.* Cognitive and organizational structures: A conceptual analysis of implicit organizing theories / A. P. Brief, H. K. Downey // Human relations. — 1983. — T. 36, № 12. — C. 1065—1089.

74. *Hall, M.* Shale vs tight / M. Hall // Agile Geoscience. — 2011.

76. *Lim, B.-C.* Team mental models and team performance: A field study of the effects of team mental model similarity and accuracy / B.-C. Lim, K. J. Klein // Journal of Organizational Behavior: The International Journal of Industrial, Occupational and Organizational Psychology and Behavior. — 2006. — T. 27, № 4. — C. 403—418.

77. The influence of shared mental models on team process and performance. / J. E. Mathieu, T. S. Heffner, G. F. Goodwin, E. Salas, J. A. Cannon-Bowers // Journal of applied psychology. — 2000. — T. 85, № 2. — C. 273.

78. *Harper, R. G.* Power, dominance, and nonverbal behavior: An overview / R. G. Harper // Power, dominance, and nonverbal behavior. — Springer, 1985. — C. 29—48.

79. *McPherson, M.* Birds of a feather: Homophily in social networks / M. McPherson, L. Smith-Lovin, J. M. Cook // Annual review of sociology. — 2001. — Т. 27, № 1. — С. 415—444.

80. *Snijders, T. A.* Introduction to stochastic actor-based models for network dynamics / T. A. Snijders, G. G. Van de Bunt, C. E. Steglich // Social networks. — 2010. — Т. 32, № 1. — С. 44—60.

81. *Steglich, C.* 8. Dynamic Networks and Behavior: Separating Selection from Influence / C. Steglich, T. A. Snijders, M. Pearson // Sociological methodology. — 2010. — Т. 40, № 1. — С. 329—393.

82. *Weizenbaum, J.* ELIZA—a computer program for the study of natural language communication between man and machine / J. Weizenbaum // Communications of the ACM. — 1966. — Т. 9, № 1. — С. 36—45.

83. *Kucera, H.* Computational analysis of present – day American English / H. Kucera, W. N. Francis. — Dartmouth Publishing Group, 1967.

84. *Kleene, S. C.* Representation of events in nerve nets and finite automata : тех. отч. / S. C. Kleene ; RAND PROJECT AIR FORCE SANTA MONICA CA. — 1951.

85. *Thompson, K.* Programming techniques: Regular expression search algorithm / K. Thompson // Communications of the ACM. — 1968. — Т. 11, № 6. — С. 419—422.

86. emoji2vec: Learning emoji representations from their description / B. Eisner, T. Rocktaschel, I. Augenstein, M. Bosnjak, S. Riedel // arXiv preprint arXiv:1609.08359. — 2016.

87. From tweets to polls: Linking text sentiment to public opinion time series. / B. O'Connor, R. Balasubramanyan, B. R. Routledge, N. A. Smith [и др.] // Icwsm. — 2010. — Т. 11, № 122—129. — С. 1—2.

88. *Bingel, J.* Identifying beneficial task relations for multi-task learning in deep neural networks / J. Bingel, A. Sgaard // arXiv preprint arXiv:1702.08303. — 2017.

89. NLP-driven citation analysis for scientometrics / R. Jha, A.-A. Jbara, V. Qazvinian, D. R. Radev // Natural Language Engineering. — 2017. — Т. 23, № 1. — С. 93—130.

90. *Lovins, J. B.* Development of a stemming algorithm / J. B. Lovins // Mech. Translat. & Comp. Linguistics. — 1968. — Т. 11, № 2. — С. 22—31.

91. *Segalovich, I.* A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine. / I. Segalovich // MLMTA. — Citeseer. 2003. — С. 273—280.

92. *Sharoff, S.* The proper place of men and machines in language technology: Processing Russian without any linguistic knowledge / S. Sharoff, J. Nivre // Proc.

Dialogue 2011, Russian Conference on Computational Linguistics. — 2011.

93. *Korobov, M.* Morphological analyzer and generator for Russian and Ukrainian languages / M. Korobov // International Conference on Analysis of Images, Social Networks and Texts. — Springer. 2015. — С. 320—332.

94. *Willett, P.* The Porter stemming algorithm: then and now / P. Willett // Program. — 2006. — Т. 40, № 3. — С. 219—223.

95. *Porter, M. F.* Snowball: A language for stemming algorithms / M. F. Porter. — 2001.

96. *Packard, D.* Computer–assisted morphological analysis of ancient Greek / D. Packard // COLING 1973 Volume 2: Computational And Mathematical Linguistics: Proceedings of the International Conference on Computational Linguistics. Т. 2. — 1973.

97. *Bird, S.* Natural language processing with Python: analyzing text with the natural language toolkit / S. Bird, E. Klein, E. Loper. — "O'Reilly Media, Inc.", 2009.

99. *Schwenk, H.* Connectionist language modeling for large vocabulary continuous speech recognition / H. Schwenk, J.-L. Gauvain // Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on. Т. 1. — IEEE. 2002. — С. I—765.

100. *Teahan, W. J.* The entropy of English using PPM-based models / W. J. Teahan, J. G. Cleary // dcc. — IEEE. 1996. — C. 53.

101. *Teahan, W.* Models of English text / W. Teahan, J. G. Cleary // Data Compression Conference, 1997. DCC'97. Proceedings. — IEEE. 1997. — C. 12—21.

102. *Bahl, L. R.* A maximum likelihood approach to continuous speech recognition / L. R. Bahl, F. Jelinek, R. L. Mercer // IEEE transactions on pattern analysis and machine intelligence. — 1983. — № 2. — C. 179—190.

103. Maximum mutual information estimation of hidden Markov model parameters for speech recognition / L. Bahl, P. Brown, P. De Souza, R. Mercer // Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'86. T. 11. — IEEE. 1986. — C. 49—52.

104. Experiments with the TANGORA 20,000 word speech recognizer / A. Averbuch, L. Bahl, R. Bakis, P. Brown, G. Daggett, S. Das, K. Davies, S. De Gennaro, P. De Souza, E. Epstein [и др.] // Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'87. T. 12. — IEEE. 1987. — C. 701—704.

105. Efficient estimation of word representations in vector space / T. Mikolov, K. Chen, G. Corrado, J. Dean // arXiv preprint arXiv:1301.3781. — 2013.

106. Distributed representations of words and phrases and their compositionality / T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, J. Dean // Advances in neural information processing systems. — 2013. — С. 3111—3119.

107. *Blei, D. M.* Dynamic topic models / D. M. Blei, J. D. Lafferty // Proceedings of the 23rd international conference on Machine learning. — ACM. 2006. — С. 113—120.

108. *Cover, T. M.* Entropy, relative entropy and mutual information / T. M. Cover, J. A. Thomas // Elements of information theory. — 1991. — Т. 2. — С. 1—55.

109. Asymptotic optimality and asymptotic equipartition properties of log-optimum investment / P. H. Algoet, T. M. Cover [и др.] // The Annals of Probability. — 1988. — Т. 16, № 2. — С. 876—898.

110. *ShannClaudeon, E.* Prediction and entropy of printed English / E. ShannClaudeon // Bell system technical journal. — 1951. — Т. 30, № 1. — С. 50—64.

111. *Cover, T.* A convergent gambling estimate of the entropy of English / T. Cover, R. King // IEEE Transactions on Information Theory. — 1978. — Т. 24, № 4. — С. 413—421.

112. An estimate of an upper bound for the entropy of English / P. F. Brown, V. J. D. Pietra, R. L. Mercer,

S. A. D. Pietra, J. C. Lai // Computational Linguistics. — 1992. — Т. 18, № 1. — С. 31—40.

113. *Cohen, W. W.* Learning to order things / W. W. Cohen, R. E. Schapire, Y. Singer // Advances in Neural Information Processing Systems. — 1998. — С. 451—457.

114. Natural language processing (almost) from scratch / R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, P. Kuksa // Journal of Machine Learning Research. — 2011. — Т. 12, № 8. — С. 2493—2537.

115. *Pennington, J.* Glove: Global vectors for word representation / J. Pennington, R. Socher, C. Manning // Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). — 2014. — С. 1532—1543.

116. *Maron, M. E.* Automatic indexing: an experimental inquiry / M. E. Maron // Journal of the ACM (JACM). — 1961. — Т. 8, № 3. — С. 404—417.

117. *Bayes, T.* An essay towards solving a problem in the doctrine of chances / T. Bayes, R. Price, J. Canton. — 1763.

118. *Mosteller, F.* Inference in an authorship problem: A comparative study of discrimination methods applied to the authorship of the disputed federalist papers / F. Mosteller, D. L. Wallace // Journal of the American

Statistical Association. — 1963. — Т. 58, № 302. — С. 275—309.

119. A Bayesian approach to filtering junk e-mail / M. Sahami, S. Dumais, D. Heckerman, E. Horvitz // Learning for Text Categorization: Papers from the 1998 workshop. Т. 62. — Madison, Wisconsin. 1998. — С. 98—105.

120. *Metsis, V.* Spam filtering with naive bayes-which naive bayes? / V. Metsis, I. Androutsopoulos, G. Paliouras // CEAS. Т. 17. — Mountain View, CA. 2006. — С. 28—69.

121. *Wang, S.* Baselines and bigrams: Simple, good sentiment and topic classification / S. Wang, C. D. Manning // Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2. — Association for Computational Linguistics. 2012. — С. 90—94.

122. *Pang, B.* Thumbs up?: sentiment classification using machine learning techniques / B. Pang, L. Lee, S. Vaithyanathan // Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10. — Association for Computational Linguistics. 2002. — С. 79—86.

123. A comparison of event models for naive bayes text classification / A. McCallum, K. Nigam [и др.] // AAAI-98 workshop on learning for text categorization. Т. 752. — Citeseer. 1998. — С. 41—48.

124. Opinion mining and sentiment analysis / B. Pang, L. Lee [и др.] // Foundations and Trends® in Information Retrieval. — 2008. — Т. 2, № 2. — С. 1—135.

125. *Liu, B.* A survey of opinion mining and sentiment analysis / B. Liu, L. Zhang // Mining text data. — Springer, 2012. — С. 415—463.

126. *Stamatatos, E.* A survey of modern authorship attribution methods / E. Stamatatos // Journal of the American Society for information Science and Technology. — 2009. — Т. 60, № 3. — С. 538—556.

127. *Schutze, H.* Introduction to information retrieval. Т. 39 / H. Schutze, C. D. Manning, P. Raghavan. — Cambridge University Press, 2008.

128. Starspace: Embed all the things! / L. Wu, A. Fisch, S. Chopra, K. Adams, A. Bordes, J. Weston // arXiv preprint arXiv:1709.03856. — 2017.

129. Enriching word vectors with subword information / P. Bojanowski, E. Grave, A. Joulin, T. Mikolov // arXiv preprint arXiv:1607.04606. — 2016.

130. *Pagliardini, M.* Unsupervised learning of sentence embeddings using compositional n-gram features / M. Pagliardini, P. Gupta, M. Jaggi // arXiv preprint arXiv:1703.02507. — 2017.

131. *Finley, G.* What analogies reveal about word vectors and their compositionality / G. Finley, S. Farmer, S. Pakhomov // Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (* SEM 2017). — 2017. — С. 1—11.

132. Making sense of word embeddings / M. Pelevina, N. Arefyev, C. Biemann, A. Panchenko // arXiv preprint arXiv:1708.03390. — 2017.

133. Unsupervised, knowledge-free, and interpretable word sense disambiguation / A. Panchenko, F. Marten, E. Ruppert, S. Faralli, D. Ustalov, S. P. Ponzetto, C. Biemann // arXiv preprint arXiv:1707.06878. — 2017.

135. Breaking sticks and ambiguities with adaptive skip-gram / S. Bartunov, D. Kondrashkin, A. Osokin, D. Vetrov // Artificial Intelligence and Statistics. — 2016. — С. 130—138.

136. Improving word representations via global context and multiple word prototypes / E. H. Huang, R. Socher, C. D. Manning, A. Y. Ng // Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1. — Association for Computational Linguistics. 2012. — С. 873—882.

137. *Gladkova, A.* Intrinsic evaluations of word embeddings: What can we do better? / A. Gladkova, A. Drozd //

Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP. — 2016. — C. 36—42.

138. *Schuster, S.* Sentences with Gapping: Parsing and Reconstructing Elided Predicates / S. Schuster, J. Nivre, C. D. Manning // North American Chapter of the Association of Computational Linguistics (NAACL). — 2018. — URL: https : / / nlp . stanford . edu / pubs / schuster2018gapping.pdf.

139. *Eric, M.* A copy-augmented sequence-to-sequence architecture gives good performance on task-oriented dialogue / M. Eric, C. D. Manning // arXiv preprint arXiv:1701.04024. — 2017.

140. Naturalizing a programming language via interactive learning / S. I. Wang, S. Ginn, P. Liang, C. D. Manning // arXiv preprint arXiv:1704.06956. — 2017.

141. Adversarial learning for neural dialogue generation / J. Li, W. Monroe, T. Shi, S. Jean, A. Ritter, D. Jurafsky // arXiv preprint arXiv:1701.06547. — 2017.

142. Data noising as smoothing in neural network language models / Z. Xie, S. I. Wang, J. Li, D. Levy, A. Nie, D. Jurafsky, A. Y. Ng // arXiv preprint arXiv:1703.02573. — 2017.

143. *Pawelczak, D.* Benefits and drawbacks of source code plagiarism detection in engineering education / D. Pawelczak // 2018 IEEE Global Engineering Education Conference (EDUCON). — IEEE. 2018. — С. 1048—1056.

144. *Jack, F.* Study on the Different Forms of Plagiarism in Textual Data and Image: Internal and External Detection / F. Jack // Advanced Metaheuristic Methods in Big Data Retrieval and Analytics. — IGI Global, 2019. — С. 75—90.

145. *Rayson, P.* Comparing corpora using frequency profiling / P. Rayson, R. Garside // Proceedings of the workshop on Comparing corpora-Volume 9. — Association for Computational Linguistics. 2000. — С. 1—6.

146. Basic statistical analysis of corpus and cross comparison among corpora / A. Bharati, K. P. Rao, R. Sangal, S. Bendre // Technical Report of Indian Institute of Information Technology. — 2000.

147. Measuring Ethnic Stratification and its Effect on Trust in Africa / R. Hodler, S. Srisuma, A. Vesperoni, N. Zurlinden. — 2018.

148. *Rayson, P.* Matrix: A statistical method and software tool for linguistic analysis through corpus comparison : дис. ... канд. / Rayson Paul. — Lancaster University, 2003.

149. *Perez, M. J. M.* Measuring the degree of specialisation of sub-technical legal terms through corpus comparison / M. J. M. Perez // Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication. — 2016. — Т. 22, № 1. — С. 80—102.

150. *Chung, T. M.* A corpus comparison approach for terminology extraction / T. M. Chung // Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication. — 2003. — Т. 9, № 2. — С. 221—246.

151. The mathematics of statistical machine translation: Parameter estimation / P. F. Brown, V. J. D. Pietra, S. A. D. Pietra, R. L. Mercer // Computational linguistics. — 1993. — Т. 19, № 2. — С. 263—311.

152. *Tiedemann, J.* Improved sentence alignment for movie subtitles / J. Tiedemann // Proceedings of RANLP. Т. 7. — 2007.

153. *Itamar, E.* Using Movie Subtitles for Creating a Large-Scale Bilingual Corpora. / E. Itamar, A. Itai // LREC. — 2008.

154. *Tiedemann, J.* Parallel Data, Tools and Interfaces in OPUS. / J. Tiedemann // Lrec. Т. 2012. — 2012. — С. 2214—2218.

155. *Mohammadi, M.* Building bilingual parallel corpora based on wikipedia / M. Mohammadi, N. GhasemAghaee // 2010 Second International Conference on Computer

Engineering and Applications. T. 2. — IEEE. 2010. — C. 264—268.

156. *Weaver, W.* Translation / W. Weaver // Machine translation of languages. — 1955. — T. 14. — C. 15—23.

157. Opennmt: Open-source toolkit for neural machine translation / G. Klein, Y. Kim, Y. Deng, J. Senellart, A. M. Rush // arXiv preprint arXiv:1701.02810. — 2017.

158. Massive exploration of neural machine translation architectures / D. Britz, A. Goldie, M.-T. Luong, Q. Le // arXiv preprint arXiv:1703.03906. — 2017.

159. *Sutskever, I.* Sequence to sequence learning with neural networks / I. Sutskever, O. Vinyals, Q. V. Le // Advances in neural information processing systems. — 2014. — C. 3104—3112.

161. *Vorontsov, K.* Additive regularization of topic models / K. Vorontsov, A. Potapenko // Machine Learning. — 2015. — T. 101, № 1—3. — C. 303—323.

162. *Tsvetovat, M.* Social Network Analysis for Startups: Finding connections on the social web / M. Tsvetovat, A. Kouznetsov. — "O'Reilly Media, Inc.", 2011.

163. *Atherley, S.* A Model of Policy Formation Through Simulated Annealing: The Impact of Preference Alignment on Productivity and Satisfaction / S. Atherley, C. Dillon, V. Kane // International Conference on Social Computing, Behavioral-Cultural

Modeling, and Prediction. — Springer. 2015. — C. 93—100.

164. *Porter, W.* Mapping the Navy Innovation Network Using Social Network Analysis / W. Porter, C. Warren, R. Schroeder. — 2018.

165. *Kurmukov, A.* Classification of normal and pathological brain networks based on similarity in graph partitions / A. Kurmukov, Y. Dodonova, L. Zhukov // Data Mining Workshops (ICDMW), 2016 IEEE 16th International Conference on. — IEEE. 2016. — C. 107—112.

166. *Zafarani, R.* Social media mining: an introduction / R. Zafarani, M. A. Abbasi, H. Liu. — Cambridge University Press, 2014.

167. *Makarov, I.* Co-author recommender system / I. Makarov, O. Bulanov, L. E. Zhukov // International Conference on Network Analysis. — Springer. 2016. — C. 251—257.

168. Recommending Co-authorship via Network Embeddings and Feature Engineering: The case of National Research University Higher School of Economics / I. Makarov, O. Gerasimova, P. Sulimov, L. E. Zhukov // Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries. — ACM. 2018. — C. 365—366.

169. *Watts, D. J.* Collective dynamics of 'small-world'networks / D. J. Watts, S. H. Strogatz // nature. — 1998. — Т. 393, № 6684. — С. 440.

170. *Barabasi, A.-L.* Emergence of scaling in random networks / A.-L. Barabasi, R. Albert // science. — 1999. — Т. 286, № 5439. — С. 509—512.

171. *Fortunato, S.* Community detection in graphs / S. Fortunato // Physics reports. — 2010. — Т. 486, № 3. — С. 75—174.

172. *Lancichinetti, A.* Community detection algorithms: a comparative analysis / A. Lancichinetti, S. Fortunato // Physical review E. — 2009. — Т. 80, № 5. — С. 056117.

173. From node embedding to community embedding / V. W. Zheng, S. Cavallari, H. Cai, K. C.-C. Chang, E. Cambria // arXiv preprint arXiv:1610.09950. — 2016.

174. Semantic Proximity Search on Heterogeneous Graph by Proximity Embedding. / Z. Liu, V. W. Zheng, Z. Zhao, F. Zhu, K. C.-C. Chang, M. Wu, J. Ying // AAAI. — 2017. — С. 154—160.

183. *Mazov, N. A.* Russian publications and journals on Earth sciences in international databases / N. A. Mazov, V. N. Gureev, M. I. Epov // Herald of the Russian Academy of Sciences. — 2015. — Т. 85, № 1. — С. 20—25.

192. *Щербаков, А. И.* Эффективность научной деятельности в СССР / А. И. Щербаков. — 1982.

193. *Овчинников, О.* К методологии оценки научной деятельности в научных и образовательных учреждениях Российской Федерации / О. Овчинников // Вестник

Московского университета МВД России. — 2009. — № 3. — С. 48—51.

194. *Фурсов, К.* Факторы результативности научной деятельности: микроуровневый анализ / К. Фурсов, Р. Яна, О. Балмуш // Форсайт. — 2016. — Т. 10, № 2.

195. *Шматко, Н.* Служба или служение? Мотивационные паттерны российских ученых / Н. Шматко, Г. Волкова // Форсайт. — 2017. — Т. 11, № 2.

196. *Moe, N. B.* Understanding self-organizing teams in agile software development / N. B. Moe, T. Dingsoyr, T. Dybaa // Software Engineering, 2008. ASWEC 2008. 19th Australian Conference on. — IEEE. 2008. — С. 76—85.

197. *Bavelas, A.* A mathematical model for group structures / A. Bavelas // Human organization. — 1948. — Т. 7, № 3. — С. 16—30.

198. *Новиков, Д. А.* Математические модели формирования и функционирования команд / Д. А. Новиков // М.: Издательство физико-математической литературы. — 2008.

199. *Бейльханов, Д. К.* Использование модели компетенций в процессе командообразования / Д. К. Бейльханов, И. Ю. Квятковская // Технические науки-от теории к практике. — 2014. — № 30. — С. 7—12.

200. *Rozewski, P.* Competence management in knowledge-based organisation: case study based on higher education organisation / P. Rozewski, B. Malachowski // Knowledge Science, Engineering and Management. — 2009. — С. 358—369.

201. Collaboration patterns in the German political science co-authorship network / P. Leifeld, S. Wankmuller, V. T. Berger, K. Ingold, C. Steiner // PloS one. — 2017. — Т. 12, № 4. — e0174671.

202. Authorship trends, collaboration patterns, and co-authorship networks in lodging studies (1990–2016) / M. A. Koseoglu, F. Okumus, E. D. Putra, M. Yildiz, I. C. Dogan // Journal of Hospitality Marketing & Management. — 2018. — Т. 27, № 5. — С. 561—582.

203. with basic network measures of 2008-2017 Scopus data / T. M. Ho, H. V. Nguyen, T.-T. Vuong, Q.-M. Dam, H.-H. Pham, Q.-H. Vuong. — 2017.

204. *Chang, H.-J.* The Hidden Power of Social-Linkage in the Office: A Co-authorship Network Analysis / H.-J. Chang, W.-M. Wang // Proceedings of the 4th Multidisciplinary International Social Networks Conference on ZZZ. — ACM. 2017. — С. 4.

205. Semantic Similarity versus Co-authorship Networks: A Detailed Comparison / I. C. Paraschiv, M. Dascalu, S. Trausan-Matu, N. Nistor, A. M. M. De Oca, D. S. McNamara // Control Systems and Computer

Science (CSCS), 2017 21st International Conference on. — IEEE. 2017. — C. 566—570.

206. Analysis of co-authorship in computer networks using centrality measures / T. Ahmed, A. Ahmed, M. Ali, M. Kamran // Communication, Computing and Digital Systems (C-CODE), International Conference on. — IEEE. 2017. — C. 54—57.

207. University-industry research collaboration in the Brazilian oil industry: the case of Petrobras / G. G. Gielfi, A. T. Furtado, A. S. de Campos, R. J. Tijssen // Rev. Bras. Inov. — 2017. — T. 16, № 2. — C. 325—350.

208. *Fowler, M.* The agile manifesto / M. Fowler, J. Highsmith // Software Development. — 2001. — T. 9, № 8. — C. 28—35.

209. *Hoda, R.* Multi-level agile project management challenges: A self-organizing team perspective / R. Hoda, L. K. Murugesan // Journal of Systems and Software. — 2016. — T. 117. — C. 245—257.

210. *Moe, N. B.* Overcoming barriers to self-management in software teams / N. B. Moe, T. Dingsoyr, T. Dybaa // IEEE software. — 2009. — T. 26, № 6.

211. *Alnuaimi, O. A.* Team size, dispersion, and social loafing in technology-supported teams: A perspective on the theory of moral disengagement / O. A. Alnuaimi, L. P. Robert, L. M. Maruping // Journal of Management

Information Systems. — 2010. — Т. 27, № 1. — С. 203—230.

212. Team assembly mechanisms determine collaboration network structure and team performance / R. Guimera, B. Uzzi, J. Spiro, L. A. N. Amaral // Science. — 2005. — Т. 308, № 5722. — С. 697—702.

213. *Berger, J.* Status characteristics and social interaction / J. Berger, B. P. Cohen, M. Zelditch Jr // American Sociological Review. — 1972. — С. 241—255.

214. *Berger, J.* Status organizing processes / J. Berger, S. J. Rosenholtz, M. Zelditch Jr // Annual review of sociology. — 1980. — Т. 6, № 1. — С. 479—508.

215. *Hamilton, B. H.* Team incentives and worker heterogeneity: An empirical analysis of the impact of teams on productivity and participation / B. H. Hamilton, J. A. Nickerson, H. Owan // Journal of political Economy. — 2003. — Т. 111, № 3. — С. 465—497.

216. *Prat, A.* Should a team be homogeneous? / A. Prat // European Economic Review. — 2002. — Т. 46, № 7. — С. 1187—1207.

217. *Boehm, B.* People factors in software management: lessons from comparing agile and plan-driven methods / B. Boehm, R. Turner // Crosstalk-The Journal of Defense Software Engineering,(Dec 2003). — 2003.

218. *Paasivaara, M.* Experiences in scaling the product owner role in large-scale globally distributed scrum / M. Paasivaara, V. T. Heikkila, C. Lassenius // Global Software Engineering (ICGSE), 2012 IEEE Seventh International Conference on. — IEEE. 2012. — С. 174—178.

219. *Tannenbaum, S. I.* Do team and individual debriefs enhance performance? A meta-analysis / S. I. Tannenbaum, C. P. Cerasoli // Human factors. — 2013. — Т. 55, № 1. — С. 231—245.

220. *Hill, G. W.* Group versus individual performance: Are N+ 1 heads better than one? / G. W. Hill // Psychological bulletin. — 1982. — Т. 91, № 3. — С. 517.

221. Job crafting at the team and individual level: Implications for work engagement and performance / M. Tims, A. B. Bakker, D. Derks, W. Van Rhenen // Group & Organization Management. — 2013. — Т. 38, № 4. — С. 427—454.

222. *Cropley, D. H.* Measuring functional creativity: Non-expert raters and the Creative Solution Diagnosis Scale / D. H. Cropley, J. C. Kaufman // The Journal of Creative Behavior. — 2012. — Т. 46, № 2. — С. 119—137.

223. *Jackson, P. W.* The person, the product, and the response: conceptual problems in the assessment of creativity 1 / P. W. Jackson, S. Messick // Journal of personality. — 1965. — Т. 33, № 3. — С. 309—329.

224. Sensible organizations: Technology and methodology for automatically measuring organizational behavior / D. O. Olguin, B. N. Waber, B. Taemie Kim, A. Mohan, K. Ara, A. Pentland //. — Institute of Electrical, Electronics Engineers. 2008.

225. *Sutherland, J.* The scrum guide / J. Sutherland, K. Schwaber // The definitive guide to scrum: The rules of the game. Scrum. org. — 2013. — T. 268.

226. Toward An integrative CSDS based model of industrial R&D division efficiency / F. Pereme, B. Rose, V. Goepp, J. P. Radoux, A. Belhaoua // IFAC-PapersOnLine. — 2016. — T. 49, № 12. — C. 1785—1790.

227. Relationships among team ability composition, team mental models, and team performance. / B. D. Edwards, E. A. Day, W. Arthur Jr, S. T. Bell // Journal of Applied Psychology. — 2006. — T. 91, № 3. — C. 727.

228. *Hofmann, T.* Probabilistic latent semantic indexing / T. Hofmann // ACM SIGIR Forum. T. 51. — ACM. 2017. — C. 211—218.

229. *Lu, X.* Latent semantic minimal hashing for image retrieval / X. Lu, X. Zheng, X. Li // IEEE Transactions on Image Processing. — 2016. — T. 26, № 1. — C. 355—368.

230. LTSG: Latent Topical Skip-Gram for mutually learning topic model and vector representations / J. Law,

H. H. Zhuo, J. He, E. Rong // arXiv preprint arXiv:1702.07117. — 2017.

231. *Blei, D. M.* Latent dirichlet allocation / D. M. Blei, A. Y. Ng, M. I. Jordan // Journal of machine Learning research. — 2003. — Т. 3, № 1. — С. 993—1022.

232. Bag of tricks for efficient text classification / A. Joulin, E. Grave, P. Bojanowski, T. Mikolov // arXiv preprint arXiv:1607.01759. — 2016.

233. *Ianina, A.* Multi-objective topic modeling for exploratory search in tech news / A. Ianina, L. Golitsyn, K. Vorontsov // Conference on Artificial Intelligence and Natural Language. — Springer. 2017. — С. 181—193.

234. BLEU: a method for automatic evaluation of machine translation / K. Papineni, S. Roukos, T. Ward, W.-J. Zhu // Proceedings of the 40th annual meeting on association for computational linguistics. — Association for Computational Linguistics. 2002. — С. 311—318.

235. Open source toolkit for statistical machine translation: Factored translation models and confusion network decoding / P. Koehn, M. Federico, W. Shen, N. Bertoldi, O. Bojar, C. Callison-Burch, B. Cowan, C. Dyer, H. Hoang, R. Zens [и др.] // Final Report of the 2006 JHU Summer Workshop. — 2006.

236. *Green, S.* Phrasal: A toolkit for new directions in statistical machine translation / S. Green, D. Cer,

C. Manning // Proceedings of the Ninth Workshop on Statistical Machine Translation. — 2014. — C. 114—121.

238. *Vorontsov, K.* Additive regularization of topic models for topic selection and sparse factorization / K. Vorontsov, A. Potapenko, A. Plavin // International Symposium on Statistical Learning and Data Sciences. — Springer. 2015. — C. 193—202.

239. *Koltsov, S.* A Full-Cycle Methodology for News Topic Modeling and User Feedback Research / S. Koltsov, S. Pashakhin, S. Dokuka // International Conference on Social Informatics. — Springer. 2018. — C. 308—321.

240. *Seroussi, Y.* Authorship attribution with author-aware topic models / Y. Seroussi, F. Bohnert, I. Zukerman // Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2. — Association for Computational Linguistics. 2012. — C. 264—269.

241. Discovering research topics from library electronic references using latent Dirichlet allocation / D. Fang, H. Yang, B. Gao, X. Li // Library Hi Tech. — 2018. — Февр. — Т. 36, № 3. — С. 400—410. — DOI: 10.1108/LHT-06-2017-0132. — URL: https://app.dimensions.ai/details/publication/pub.1101114990.

242. Understanding LDA in Source Code Analysis / D. Binkley, D. Heinz, D. Lawrie, J. Overfelt // Proceedings of the 22Nd International Conference

on Program Comprehension. — Hyderabad, India : ACM, 2014. — С. 26—36. — (ICPC 2014). — ISBN 978-1-4503-2879-1. — DOI: 10.1145/2597008.2597150. — URL: http://doi.acm.org/10.1145/2597008.2597150.

243. *Agrawal, A.* What is wrong with topic modeling? And how to fix it using search-based software engineering / A. Agrawal, W. Fu, T. Menzies // Information and Software Technology. — 2018. — Янв. — Т. 98. — С. 74—88. — DOI: 10.1016/j.infsof.2018.02.005. — URL: https://doi.org/10.1016/j.infsof.2018.02.005.

244. *Storn, R.* Differential Evolution – A Simple and Efficient Heuristic for Global Optimization over Continuous Spaces / R. Storn, K. Price // J. of Global Optimization. — Hingham, MA, USA, 1997. — Дек. — Т. 11, № 4. — С. 341—359. — DOI: 10.1023/A:1008202821328. — URL: https://doi.org/10.1023/A:1008202821328.

245. On Smoothing and Inference for Topic Models / A. Asuncion, M. Welling, P. Smyth, Y. W. Teh // Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence. — Montreal, Quebec, Canada : AUAI Press, 2009. — С. 27—34. — (UAI '09). — ISBN 978-0-9749039-5-8. — URL: http://dl.acm.org/citation.cfm?id=1795114.1795118.

246. Evaluation Methods for Topic Models / H. M. Wallach, I. Murray, R. Salakhutdinov, D. Mimno // Proceedings

of the 26th Annual International Conference on Machine Learning. — Montreal, Quebec, Canada : ACM, 2009. — C. 1105—1112. — (ICML '09). — ISBN 978-1-60558-516-1. — DOI: 10.1145/1553374.1553515. — URL: http://doi.acm.org/10.1145/1553374.1553515.

247. Reading Tea Leaves: How Humans Interpret Topic Models / J. Chang, J. Boyd-Graber, S. Gerrish, C. Wang, D. M. Blei // Proceedings of the 22Nd International Conference on Neural Information Processing Systems. — Vancouver, British Columbia, Canada : Curran Associates Inc., 2009. — C. 288—296. — (NIPS'09). — ISBN 978-1-61567-911-9. — URL: http://dl.acm.org/citation.cfm?id=2984093.2984126.

248. *Koltcov, S.* Latent Dirichlet Allocation: Stability and Applications to Studies of User-generated Content / S. Koltcov, O. Koltsova, S. Nikolenko // Proceedings of the 2014 ACM Conference on Web Science. — Bloomington, Indiana, USA : ACM, 2014. — C. 161—165. — (WebSci '14). — ISBN 978-1-4503-2622-3. — DOI: 10.1145/2615569.2615680. — URL: http://doi.acm.org/10.1145/2615569.2615680.

249. *Mimno, D.* Bayesian Checking for Topic Models / D. Mimno, D. Blei // Proceedings of the Conference on Empirical Methods in Natural Language Processing. — Edinburgh, United Kingdom : Association for Computational Linguistics, 2011. — C. 227—237. —

(EMNLP '11). — ISBN 978-1-937284-11-4. — URL: http://dl.acm.org/citation.cfm?id=2145432.2145459.

250. Sharing Clusters Among Related Groups: Hierarchical Dirichlet Processes / Y. W. Teh, M. I. Jordan, M. J. Beal, D. M. Blei // Proceedings of the 17th International Conference on Neural Information Processing Systems. — Vancouver, British Columbia, Canada : MIT Press, 2004. — C. 1385—1392. — (NIPS'04). — URL: http://dl.acm.org/citation.cfm?id=2976040.2976214.

251. *Blei, D. M.* The Nested Chinese Restaurant Process and Bayesian Nonparametric Inference of Topic Hierarchies / D. M. Blei, T. L. Griffiths, M. I. Jordan // J. ACM. — New York, NY, USA, 2010. — Февр. — Т. 57, № 2. — 7:1—7:30. — DOI: 10.1145/1667053.1667056. — URL: http://doi.acm.org/10.1145/1667053.1667056.

252. Hierarchical Topic Models and the Nested Chinese Restaurant Process / D. M. Blei, M. I. Jordan, T. L. Griffiths, J. B. Tenenbaum // Proceedings of the 16th International Conference on Neural Information Processing Systems. — Whistler, British Columbia, Canada : MIT Press, 2003. — C. 17—24. — (NIPS'03). — URL: http://dl.acm.org/citation.cfm?id=2981345.2981348.

253. *Bryant, M.* Truly Nonparametric Online Variational Inference for Hierarchical Dirichlet Processes / M. Bryant, E. B. Sudderth // Proceedings of the

25th International Conference on Neural Information Processing Systems - Volume 2. — Lake Tahoe, Nevada : Curran Associates Inc., 2012. — С. 2699—2707. — (NIPS'12). — URL: http://dl.acm.org/citation.cfm?id=2999325.2999436.

254. *Rossetti, M.* Towards Explaining Latent Factors with Topic Models in Collaborative Recommender Systems / M. Rossetti, F. Stella, M. Zanker // 2013 24th International Workshop on Database and Expert Systems Applications. — IEEE, 09.2013. — DOI: 10.1109/DEXA.2013.26. — URL: https://doi.org/10.1109/dexa.2013.26.

255. Automatic Evaluation of Topic Coherence / D. Newman, J. H. Lau, K. Grieser, T. Baldwin // Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. — Los Angeles, California : Association for Computational Linguistics, 2010. — С. 100—108. — (HLT '10). — ISBN 1-932432-65-5. — URL: http://dl.acm.org/citation.cfm?id=1857999.1858011.

256. *Koltcov, S.* Application of Renyi and Tsallis entropies to topic modeling optimization / S. Koltcov // Physica A: Statistical Mechanics and its Applications. — 2018. — Дек. — Т. 512. — С. 1192—1204. — DOI: 10.1016/j.physa.2018.08.050. — URL: https://doi.org/10.1016/j.physa.2018.08.050.

257. *Bing, X.* A fast algorithm with minimax optimal guarantees for topic models with an unknown number of topics / X. Bing, F. Bunea, M. H. Wegkamp // CoRR. — 2018. — Май. — Т. abs/1805.06837.

258. *Lipton, Z. C.* The Mythos of Model Interpretability / Z. C. Lipton // Queue. — New York, NY, USA, 2018. — Июнь. — Т. 16, № 3. — 30:31—30:57. — DOI: 10.1145/3236386.3241340. — URL: http://doi.acm.org/10.1145/3236386.3241340.

259. Progressive Learning of Topic Modeling Parameters: A Visual Analytics Framework / M. El-Assady, R. Sevastjanova, F. Sperrle, D. Keim, C. Collins // IEEE Transactions on Visualization and Computer Graphics. — 2018. — Янв. — Т. 24, № 1. — С. 382—391. — DOI: 10.1109/TVCG.2017.2745080.

260. *Nikolenko, S. I.* Topic modelling for qualitative studies / S. I. Nikolenko, S. Koltcov, O. Koltsova // Journal of Information Science. — 2016. — Июль. — Т. 43, № 1. — С. 88—102. — DOI: 10.1177/0165551515617393. — URL: https://doi.org/10.1177/0165551515617393.

261. Nonparametric Spherical Topic Modeling with Word Embeddings / K. Batmanghelich, A. Saeedi, K. Narasimhan, S. Gershman // Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). — Association for Computational Linguistics, 2016. — DOI: 10.18653/v1/

p16-2087. — URL: https://doi.org/10.18653/v1/p16-2087.

262. LTSG: Latent Topical Skip-Gram for Mutually Improving Topic Model and Vector Representations / J. Law, H. H. Zhuo, J. He, E. Rong // Pattern Recognition and Computer Vision. — Springer International Publishing, 2018. — C. 375—387. — DOI: 10.1007/978-3-030-03338-5_32. — URL: https://doi.org/10.1007/978-3-030-03338-5_32.

263. *Das, R.* Gaussian LDA for Topic Models with Word Embeddings / R. Das, M. Zaheer, C. Dyer // Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). — Beijing, China : Association for Computational Linguistics, 2015. — C. 795—804. — DOI: 10.3115/v1/P15-1077. — URL: http://aclweb.org/anthology/P15-1077.

264. Improving Topic Models with Latent Feature Word Representations / D. Nguyen, R. Billingsley, L. Du, M. Johnson // Transactions of the Association for Computational Linguistics. — 2015. — Т. 3. — С. 299—313. — URL: http://aclweb.org/anthology/Q15-1022.

265. *Mantyla, M. V.* Measuring LDA Topic Stability from Clusters of Replicated Runs / M. V. Mantyla, M. Claes,

U. Farooq // Proceedings of the 12th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement. — Oulu, Finland : ACM, 2018. — 49:1—49:4. — (ESEM '18). — ISBN 978-1-4503-5823-1. — DOI: 10.1145/3239235.3267435. — URL: http://doi.acm.org/10.1145/3239235.3267435.

266. *Mehta, V.* Evaluating topic quality using model clustering / V. Mehta, R. S. Caceres, K. M. Carter // 2014 IEEE Symposium on Computational Intelligence and Data Mining (CIDM). — 12.2014. — С. 178—185. — DOI: 10.1109/cidm.2014.7008665. — URL: https://doi.org/10.1109/cidm.2014.7008665.

267. *Bezdek, J. C.* Cluster Validity with Fuzzy Sets / J. C. Bezdek // Journal of Cybernetics. — 1973. — Т. 3, № 3. — С. 58—73. — DOI: 10.1080/01969727308546047. — URL: https://doi.org/10.1080/01969727308546047.

268. *Dunn, J. C.* Well-Separated Clusters and Optimal Fuzzy Partitions / J. C. Dunn // Journal of Cybernetics. — 1974. — Янв. — Т. 4, № 1. — С. 95—104. — DOI: 10.1080/01969727408546059. — URL: https://doi.org/10.1080/01969727408546059.

269. *Davies, D. L.* A Cluster Separation Measure / D. L. Davies, D. W. Bouldin // IEEE Trans. Pattern Anal. Mach. Intell. — Washington, DC, USA, 1979. — Февр. — Т. 1, № 2. — С. 224—227. — DOI: 10.1109/

TPAMI.1979.4766909. — URL: http://dx.doi.org/10.1109/TPAMI.1979.4766909.

270. *Halkidi, M.* Clustering Validity Checking Methods: Part II / M. Halkidi, Y. Batistakis, M. Vazirgiannis // SIGMOD Rec. — New York, NY, USA, 2002. — Сент. — Т. 31, № 3. — С. 19—27. — DOI: 10.1145/601858.601862. — URL: http://doi.acm.org/10.1145/601858.601862.

271. *Xie, X. L.* A Validity Measure for Fuzzy Clustering / X. L. Xie, G. Beni // IEEE Trans. Pattern Anal. Mach. Intell. — Washington, DC, USA, 1991. — Авг. — Т. 13, № 8. — С. 841—847. — DOI: 10.1109/34.85677. — URL: http://dx.doi.org/10.1109/34.85677.

272. *Rousseeuw, P.* Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis / P. Rousseeuw // J. Comput. Appl. Math. — Amsterdam, The Netherlands, The Netherlands, 1987. — Нояб. — Т. 20, № 1. — С. 53—65. — DOI: 10.1016/0377-0427(87)90125-7. — URL: http://dx.doi.org/10.1016/0377-0427(87)90125-7.

273. *Pennington, J.* Glove: Global Vectors for Word Representation / J. Pennington, R. Socher, C. Manning // Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). — Doha, Qatar : Association for Computational Linguistics, 2014. — С. 1532—1543. —

DOI: 10.3115/v1/D14-1162. — URL: http://aclweb.org/anthology/D14-1162.

274. Enriching Word Vectors with Subword Information / P. Bojanowski, E. Grave, A. Joulin, T. Mikolov // Transactions of the Association for Computational Linguistics. — 2017. — Т. 5. — С. 135—146. — URL: http://aclweb.org/anthology/Q17-1010.

275. StarSpace: Embed All The Things! / L. Y. Wu, A. Fisch, S. Chopra, K. Adams, A. Bordes, J. Weston // AAAI. — 2018.

276. *Vorontsov, K.* Additive Regularization of Topic Models for Topic Selection and Sparse Factorization / K. Vorontsov, A. Potapenko, A. Plavin // Statistical Learning and Data Sciences. — Springer International Publishing, 2015. — С. 193—202. — DOI: 10.1007/978-3-319-17091-6_14. — URL: https://doi.org/10.1007/978-3-319-17091-6_14.

277. Generating Cohesive Semantic Topics from Latent Factors / P. V. Bicalho, T. d. O. Cunha, F. H. J. Mourao, G. L. Pappa, W. Meira // 2014 Brazilian Conference on Intelligent Systems. — IEEE, 10.2014. — С. 271—276. — DOI: 10.1109/bracis.2014.56. — URL: https://doi.org/10.1109/bracis.2014.56.

278. *Kuhn, A.* Semantic clustering: Identifying topics in source code / A. Kuhn, S. Ducasse, T. Girba // Information and Software Technology. — 2007. — Март. — Т. 49, № 3. —

C. 230—243. — DOI: 10.1016/j.infsof.2006.10.017. — URL: https://doi.org/10.1016/j.infsof.2006.10.017.

279. TopicCheck: Interactive Alignment for Assessing Topic Model Stability / J. Chuang, M. E. Roberts, B. M. Stewart, R. Weiss, D. Tingley, J. Grimmer, J. Heer // Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. — Denver, Colorado : Association for Computational Linguistics, 2015. — C. 175—184. — DOI: 10.3115/v1/N15-1018. — URL: http://aclweb.org/anthology/N15-1018.

280. *Greene, D.* How Many Topics? Stability Analysis for Topic Models / D. Greene, D. O'Callaghan, P. Cunningham // Machine Learning and Knowledge Discovery in Databases. — Springer: Berlin, Heidelberg, 2014. — C. 498—513. — DOI: 10.1007/978-3-662-44848-9_32. — URL: https://doi.org/10.1007/978-3-662-44848-9_32.

281. Stable Topic Modeling with Local Density Regularization / S. Koltcov, S. I. Nikolenko, O. Koltsova, V. Filippov, S. Bodrunova // Internet Science. — Springer International Publishing, 2016. — C. 176—188. — DOI: 10.1007/978-3-319-45982-0_16. — URL: https://doi.org/10.1007/978-3-319-45982-0_16.

285. *Oliva, R.* Death spirals and virtuous cycles / R. Oliva, J. D. Sterman // Handbook of Service Science. — Springer, 2010. — С. 321—358.

293. *Ng, H. T.* Feature selection, perceptron learning, and a usability case study for text categorization / H. T. Ng, W. B. Goh, K. L. Low // ACM SIGIR Forum. Т. 31. — ACM. 1997. — С. 67—73.

294. *Lam, S. L.* Feature reduction for neural network based text categorization / S. L. Lam, D. L. Lee // Database Systems for Advanced Applications, 1999. Proceedings., 6th International Conference on. — IEEE. 1999. — С. 195—202.

295. *Lam, W.* Automatic text categorization and its application to text retrieval / W. Lam, M. Ruiz, P. Srinivasan // IEEE Transactions on Knowledge & Data Engineering. — 1999. — № 6. — С. 865—879.

296. *Kim, Y.* Convolutional neural networks for sentence classification / Y. Kim // arXiv preprint arXiv:1408.5882. — 2014.

297. Recurrent neural network based language model / T. Mikolov, M. Karafiat, L. Burget, J. Cernocky, S. Khudanpur // Eleventh Annual Conference of the International Speech Communication Association. — 2010.

298. *Krizhevsky, A.* Imagenet classification with deep convolutional neural networks / A. Krizhevsky, I. Sutskever, G. E. Hinton // Advances in neural information processing systems. — 2012. — С. 1097—1105.

299. *Miller, G. A.* WordNet: a lexical database for English / G. A. Miller // Communications of the ACM. — 1995. — Т. 38, № 11. — С. 39—41.

300. Learning word vectors for sentiment analysis / A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, C. Potts // Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies-volume 1. — Association for Computational Linguistics. 2011. — С. 142—150.

301. Recursive deep models for semantic compositionality over a sentiment treebank / R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Ng, C. Potts // Proceedings of the 2013 conference on empirical methods in natural language processing. — 2013. — С. 1631—1642.

302. *Akkaya, C.* Subjectivity word sense disambiguation / C. Akkaya, J. Wiebe, R. Mihalcea // Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1. — Association for Computational Linguistics. 2009. — С. 190—199.

304. *Krasnov, F.* Seismic Patents (BoW) / F. Krasnov. — 2019. — DOI: http://dx.doi.org/10.17632/x5z5r6br2z.1.

305. Tackling the poor assumptions of naive bayes text classifiers / J. D. Rennie, L. Shih, J. Teevan, D. R. Karger // Proceedings of the 20th international conference on machine learning (ICML-03). — 2003. — С. 616—623.

308. Generating bilingual pragmatic color references / W. Monroe, J. Hu, A. Jong, C. Potts // arXiv preprint arXiv:1803.03917. — 2018.

309. *Borg, I.* Modern Multidimensional Scaling: Theory and Applications / I. Borg, P. Groenen // Journal of Educational Measurement. — 2003. — Сент. — Т. 40, № 3. — С. 277—280. — DOI: 10.1111/j.1745-3984.2003.tb01108.x. — URL: https://doi.org/10.1111/j.1745-3984.2003.tb01108.x.

310. *Calinski, T.* A dendrite method for cluster analysis / T. Calinski, J. Harabasz // Communications in Statistics. — 1974. — Т. 3, № 1. — С. 1—27. — DOI: 10.1080/03610927408827101. — URL: https://www.tandfonline.com/doi/abs/10.1080/03610927408827101.

311. Trend monitoring for linking science and strategy / P. Bakhtin, O. Saritas, A. Chulok, I. Kuzminov, A. Timofeev // Scientometrics. — 2017. — Т. 111, № 3. — С. 2059—2075.

312. *Kuzminov, I.* Global energy challenges and the national economy: stress scenarios for Russia / I. Kuzminov,

A. Bereznoy, P. Bakhtin // foresight. — 2017. — Т. 19, № 2. — С. 174—197.

315. *Blei, D. M.* Latent Dirichlet Allocation / D. M. Blei, A. Y. Ng, M. I. Jordan // J. Mach. Learn. Res. — 2003. — Март. — Т. 3. — С. 993—1022. — URL: http://dl.acm.org/citation.cfm?id=944919.944937.

Публикации автора по теме диссертации

В изданиях из списка ВАК РФ

7. *Краснов, Ф.* ЦИФРОВАЯ ПЛАТФОРМА РАЗВЕДКИ И ДОБЫЧИ УГЛЕВОДОРОДОВ / Ф. Краснов, А. Ершов, А. Маргарит // Открытые системы. СУБД. — 2019. — № 2. — С. 36—39.

14. *Krasnov, F.* Team assembly in R&D: A review of imitating modeling approach for science and technology center in Oil&Gaz industry / F. Krasnov, S. Dokuka, R. Yavorskiy // International Journal of Open Information Technologies. — 2018. — Т. 6, № 1. — С. 17—24.

22. *Краснов, Ф.* Нефтегазоразведка без больших данных / Ф. Краснов // Журнал "Открытые системы. СУБД". — 2015. — № 4.

23. *Krasnov, F.* High Spatial Image Classification Problem: Review of Approaches / F. Krasnov, A. Butorin // International Journal of Open Information Technologies. — 2019. — Т. 7, № 4. — С. 6—10.

26. *Krasnov, F. V.* Reconstruction of medium reflectivity coefficients based on seismic data through machine learning / F. V. Krasnov, A. V. Butorin, A. V. Mikheyenkov // Научно-технические ведомости Санкт-Петербургского государственного политехнического университета. Информатика. Телекоммуникации. Управление. — 2018. — Т. 11, № 1.

29. *Butorin, A.* Modern Approaches to Numerical Modeling of Microseismic Events / A. Butorin, F. Krasnov // International Journal of Open Information Technologies. — 2019. — Т. 7, № 3. — С. 7—16.

42. *Краснов, Ф.* Модель процесса публикаций научно-практических статей по специальности 25.00 «Науки о Земле» / Ф. Краснов // Интернет-журнал «НАУКОВЕДЕНИЕ». — 2017. — Т. 9, № 5.

55. *Кучма, И.* Управление знаниями: открытые цифровые образовательные ресурсы и архивы: материалы международного семинара. Ярославль, 14 октября 2011 г. / И. Кучма, Ф. Краснов. — 2011.

58. *Krasnov, F.* Optimization Methodology for the Selection of Frequencies to Produce an RGB Representation of the Results of Spectral Decomposition / F. Krasnov,

A. Butorin // International Journal of Open Information Technologies. — 2018. — Т. 6, № 11. — С. 21—27.

75. *Краснов, Ф.* Обзор подходов к анализу пространственных изображений высокого разрешения для применения в геофизике / Ф. Краснов, А. Буторин, А. Ситников // Cloud of Science. — 2019. — Т. 6, № 1. — С. 127—143.

98. *Краснов, Ф.* Вероятностная модель скрытых тем на основе архива журнала «Нефтяное Хозяйство» / Ф. Краснов, С. Докука // Вестник Евразийской науки. — 2018. — № 2.

160. *Krasnov, F.* Automatic Detection of Channels in Seismic Images via Deep Convolutional Neural Networks Learning / F. Krasnov, A. Butorin, A. Sitnikov // International Journal of Open Information Technologies. — 2018. — Т. 6, № 3. — С. 20—26.

175. *Krasnov, F.* Spectral Inversion in Estimation of Change in the Dominant Frequency of the Wave Field / F. Krasnov, A. Butorin // International Journal of Open Information Technologies. — 2019. — Т. 7, № 3. — С. 42—49.

178. *Буторин, А.* Возможности использования результатов спектральной инверсии при интерпретации сейсмических данных / А. Буторин, Ф. Краснов // Геофизика. — 2017. — № 4. — С. 2—7.

179. *Буторин, А.* МЕТОДИКА ОЦЕНКИ ИЗМЕНЕНИЯ ДОМИНАНТНОГО ЗНАЧЕНИЯ ЧАСТОТЫ ВОЛНОВОГО ПОЛЯ ВДОЛЬ СЕЙСМИЧЕСКОЙ ТРАССЫ / А. Буторин, Ф. Краснов // Геофизика. — 2018. — № 4. — С. 33—39.

180. Сравнение содержания коллекций научных журналов на основе разработанных тематических моделей и методики Т4С / Ф. Краснов, М. Хасанов, А. Дементов, М. Шварцман // Cloud of science. — 2019. — Т. 6, № 3.

187. *Krasnov, F.* A review of two algorithms for proxy model of enhanced oil recovery / F. Krasnov, N. Glavnov, A. Sitnikov // International Journal of Open Information Technologies. — 2017. — Т. 5, № 10. — С. 18—23.

190. *Khasanov, M.* Transactionality of Digital Transformation within an R&D Organization / M. Khasanov, F. Krasnov // International Journal of Open Information Technologies. — 2019. — Т. 7, № 5. — С. 39—42.

237. *Krasnov, F.* Exploration of Hidden Research Directions in Oil and Gas Industry via Full Text Analysis of OnePetro Digital Library / F. Krasnov, O. Ushmaev // International Journal of Open Information Technologies. — 2018. — Т. 6, № 5. — С. 7—14.

282. *Краснов, Ф.* Командообразования в научной деятельности: анализ подходов на основании имитационной

модели для научно-технического центра в нефтегазовой отрасли / Ф. Краснов, С. Додука, Р. Яворский // International Journal of Open Information Technologies. — 2018. — Т. 6, № 1. — С. 17—24.

283. *Краснов, Ф.* Прогнозирование развития соавторства в написании научных статей научно-технического центра Газпромнефть на основе модели / Ф. Краснов, И. Макаров // Вестник Евразийской науки. — 2018. — № 1.

284. *Краснов, Ф.* Моделирование изменений интеллектуального капитала в условиях повышенной нагрузки на персонал научно-исследовательской организации / Ф. Краснов, Н. Курчакова // Интернет-журнал «НАУКОВЕДЕНИЕ». — 2017. — Т. 9, № 6.

288. *Краснов, Ф.* Моделирование и оценка влияния от применения каркаса SCRUM в процессе написания научных статей / Ф. Краснов, С. Докука // Интернет-журнал «НАУКОВЕДЕНИЕ». — 2017. — Т. 9, № 6.

289. *Краснов, Ф.* Разведка скрытых направлений исследований в нефтегазовой отрасли с помощью анализа библиотеки OnePetro / Ф. Краснов, О. Ушмаев // International Journal of Open Information Technologies. — 2018. — Т. 6, № 5.

290. *Krasnov, F.* Allocation of the scientific directions of development of science and technologies center in oil and gas industry based on the co-authorship network /

F. Krasnov, M. Khasanov // International Journal of Open Information Technologies. — 2018. — Т. 6, № 4. — С. 1—6.

291. *Краснов, Ф.* Анализ методов построения графа соавторства: подход на основе двудольного графа / Ф. Краснов // International Journal of Open Information Technologies. — 2018. — Т. 6, № 2. — С. 31—37.

292. *Краснов, Ф.* Применение машинного обучения по ансамблю решающих правил для вычисления прогноза дополнительного коэффициента извлечения нефти / Ф. Краснов, Н. Главнов, А. Ситников // International Journal of Open Information Technologies. — 2017. — Т. 5, № 10.

303. *Краснов, Ф.* Анализ тональности текста научно-практических статей по нефтегазовой тематике с помощью искусственных нейронных сетей / Ф. Краснов // Вестник Евразийской науки. — 2018. — № 3.

307. *Krasnov, F.* Evaluation of Optimal Number of Topics of Topic Model: An Approach Based on the Quality of Clusters / F. Krasnov // International Journal of Open Information Technologies. — 2019. — Т. 7, № 2. — С. 8—15.

314. *Краснов, Ф.* ТРАНЗАКЦИОННОСТЬ ЦИФРОВОЙ ТРАНСФОРМАЦИИ В НАУЧНОЙ ОРГАНИЗА-

ЦИИ / Ф. Краснов // PROНЕФТЬ. Профессионально о нефти. — 2019. — 1 (11). — С. 64—67.

317. *Krasnov, F.* Digital Twin for R&D organization: approaches and methods / F. Krasnov, M. Khasanov // International Journal of Open Information Technologies. — 2019. — Т. 7, № 6. — С. 62—66.

319. ПРИНЦИПЫ ПОСТРОЕНИЯ ЦИФРОВОЙ ПЛАТФОРМЫ ДЛЯ НАУЧНО-ТЕХНИЧЕСКОГО ЦЕНТРА / Ф. Краснов, М. Хасанов, Р. Галеев, А. Маргарит // Вестник кибернетики. — 2019. — № 4. — С. 66—73.

В изданиях из списка WoS и Scopus

3. *Klyuchnikov, N.* Data-driven model for the identification of the rock type at a drilling bit / N. Klyuchnikov, F. Krasnov, A. Zaytsev // Journal of Petroleum science and Engineering. — 2019.

21. *Krasnov, F.* Measurement of maturity level of a professional community / F. Krasnov, R. Yavorskiy // Business Informatics. — 2013. — Т. 23, № 1. — С. 64—67.

24. *Krasnov, F.* Comparative Analysis of Scientific Papers Collections via Topic Modeling and Co-authorship Networks / F. Krasnov, A. Dimentov, M. Shvartsman // Conference on Artificial Intelligence and Natural Language. — Springer. 2019. — С. 77—98.

25. *Краснов, Ф.* Автоматизированное обнаружение геологических объектов в изображениях сейсмического поля с применением нейронных сетей глубокого обучения / Ф. Краснов, А. Буторин, А. Ситников // Бизнес-информатика. — 2018. — 2 (44).

30. *Krasnov, F.* MODERN APPROACHES TO NUMERICAL MODELING OF MICROSEISMIC EVENTS / F. Krasnov, A. Butorin, A. Sitnikov // GEOPHYSICAL RESEARCH. — 2019. — Т. 20, № 2. — С. 39—55.

35. *Krasnov, F.* High-Resolution Seismic Data Deconvolution by A0 Algorithm / F. Krasnov, A. Butorin // Geosciences. — 2018. — Т. 8, № 12. — С. 497.

36. *Краснов, Ф.* СРАВНИТЕЛЬНЫЙ АНАЛИЗ КОЛЛЕКЦИЙ НАУЧНЫХ ЖУРНАЛОВ / Ф. Краснов, М. Шварцман, А. Дементов // Труды СПИИРАН. — 2019. — Т. 18, № 3. — С. 766—792.

46. *Krasnov, F.* Comparison of online communities on the base of lexical analysis of the news feed / F. Krasnov, D. Ustalov, R. Yavorskiy // Proceedings of 2-nd conference on Analysis of Images, Networks and Texts, Yekaterinburg. — 2013. — С. 254—257.

134. *Krasnov, F.* Unsupervised Co-Authorship Based Algorithm for Clustering of R&D Trends at Science and Technology Centers in Oil and Gas Industry. / F. Krasnov, M. Khasanov // AIST (Supplement), CEUR Workshop Proceedings. — 2018. — С. 1—12.

176. *Krasnov, F.* Indicators of Connectivity for Urban Scientific Communities in Russian Cities / F. Krasnov, R. E. Yavorskiy, E. Vlasova // Analysis of Images, Social Networks and Texts. — Springer, 2014. — С. 111—120.

181. *Krasnov, F.* The Structure of Organization: The Coauthorship Network Case / F. Krasnov, S. Dokuka, R. Yavorskiy // International Conference on Analysis of Images, Social Networks and Texts. — Springer. 2016. — С. 100—107.

182. *Krasnov, F.* Connectivity Analysis of Computer Science Centers based on Scientific Publications Data for Major Russian Cities / F. Krasnov, E. Vlasova, R. Yavorskiy // Procedia Computer Science. — 2014. — Т. 31. — С. 892—899.

184. *Krasnov, F.* Topic Classification Through Topic Modeling with Additive Regularization for Collection of Scientific Papers / F. Krasnov // Proceedings of the 14th Central and Eastern European Software Engineering Conference Russia. — ACM. 2018. — С. 5.

185. *Krasnov, F.* THE NUMBER OF TOPICS OPTIMIZATION: CLUSTERING APPROACH / F. Krasnov, A. Sen // CEUR Workshop Proceedings Ser. "MACSPro 2019 - Proceedings of the Modeling and Analysis of Complex Systems and Processes Workshop 2019". — 2019. — С. 1—15.

186. *Krasnov, F.* A Machine Learning Approach to Enhanced Oil Recovery Prediction / F. Krasnov, N. Glavnov, A. Sitnikov // Lecture Notes in Computer Science. — Springer International Publishing, 2017. — С. 164—171.

189. Моделирование самоорганизующихся команд в научной среде / Ф. Краснов, Т. Вознесенская, Р. Яворский, П. Чеснокова // Бизнес-информатика. — 2019. — Т. 13, № 2.

191. *Krasnov, F.* Clustering of Translation via Topic Modeling / F. Krasnov, V. Lebedev // Journal of Physics: Conference Series. Т. 1405. — IOP Publishing. 2019. — С. 012008.

286. Analysis of Strong and Weak Ties in Oil & Gas Professional Community / F. Krasnov, S. Dokuka, I. Gorshkov, R. Yavorskiy // CEUR Workshop Proceedings Ser. "Proceedings of International Workshop on Formal Concept Analysis for Knowledge Discovery, FCA4KD 2017". — CEUR Workshop Proceedings. 2017. — С. 22—33.

287. *Krasnov, F.* Application of multidimensional interpolation and random forest regression to enhanced oil recovery modeling / F. Krasnov, N. Glavnov, A. Sitnikov // Proceedings of the 13th Central and Eastern European Software Engineering Conference in Russia on - CEE-SECR '17. — ACM Press, 2017.

316. A Thematic Coherence Study of a Bilingual Corpus of Articles on Oil and Gas Research / F. Krasnov, M. Shvartsman, A. Dimentov, A. Sen // Automatic Documentation and Mathematical Linguistics. — 2019. — Т. 53, № 3. — С. 138—142.

В прочих изданиях

27. *Краснов, Ф.* Использование сериализации для хранения геолого-геофизической информации / Ф. Краснов, А. Ситников // Научно-теоретический журнал. — 2015. — С. 25.

28. *Краснов, Ф.* Развитие через общение / Ф. Краснов // Intelligent Enterprise. — 2012. — № 9. — С. 18—21.

34. Numerical Modeling of Microseismic Events on the Surface / A. Butorin, F. Krasnov [и др.] // SPE Russian Petroleum Technology Conference. — Society of Petroleum Engineers. 2017.

43. *Краснов, Ф.* Человек и коммуникации / Ф. Краснов // Директор информационной службы. — 2008. — № 11.

177. Spectral Inversion Methods and its Application for Wave Field Analysis (Russian) / A. Butorin, F. Krasnov [и др.] // SPE Russian Petroleum Technology Conference. — Society of Petroleum Engineers. 2017.

306. *Krasnov, F.* The Number of Topics Optimization: Clustering Approach / F. Krasnov, A. Sen // Machine Learning and Knowledge Extraction. — 2019. — Т. 1, № 1. — С. 416—426.

318. Digital Twin of a Research Organization: Approaches and Methods / M. Khasanov, F. Krasnov [и др.] // SPE Annual Caspian Technical Conference. — Society of Petroleum Engineers. 2019.

Патенты и регистрации программных продуктов

1. Программа для ЭВМ «СПО "НауБот"» : №2017661666 / Ф. Краснов. — Заявл. 2017.
2. КСПО "HiRGB" : №2017661769 / Ф. Краснов, А. Буторин. — Заявл. 2017.
3. Экспертная Система для оценки сейсмических неопределенностей : №2018617693 / Ф. Краснов, А. Буторин. — Заявл. 2018.
4. ПРОГРАММА "SPECTRA CLUSTER"(СПЕКТРА КЛАСТЕР) ДЛЯ РАСЧЁТА КУБА СПЕКТРАЛЬНЫХ КРИВЫХ И ПОСЛЕДУЮЩЕЙ ЕГО КЛАСТЕРИЗАЦИИ : Свидетельство о регистрации программы для ЭВМ RU 2019664988, 15.11.2019. / А. Буторин, Ф. Краснов, Д. Муртазин.

В сборниках трудов конференций

1. Corporate Wikipedia in Upstream: Bimodal IT Case / M. Khasanov, F. Krasnov, B. Belozerov [и др.] // SPE Annual Technical Conference and Exhibition. — Society of Petroleum Engineers. 2016.

31. *Буторин, А. В.* Применение методов спектральной инверсии / А. В. Буторин, Ф. В. Краснов // Сейсмические технологии-2017. — 2017. — С. 192—195.

32. *Butorin, A.* Approaches to the Analysis of Spectral Decomposition for the Purpose of Detailed Geological Interpretation / A. Butorin, F. Krasnov // SPE Russian Petroleum Technology Conference and Exhibition. — Society of Petroleum Engineers. 2016.

33. *Буторин, А.* Сравнительный анализ методов спектральной инверсии волнового поля на примере модельных трасс / А. Буторин, Ф. Краснов // Геофизика. — 2016. — № 4. — С. 68—76.

57. *Яворский, Р. Э.* Сравнение онлайн-сообществ на основе лексического анализа ленты новостей / Р. Э. Яворский, Ф. В. Краснов, Д. Усталов // Доклады всероссийской научной конференции АИСТ'2013. — Национальный открытый университет «ИНТУИТ». 2013. — С. 242—245.

188. *Krasnov, F.* Segmentation of IT customers on internal market / F. Krasnov, A. Sergeev // SEC(R)'09. The 5th Software Engineering Conference (Russia). — TEKAMA. 2009.

313. *Фагерева, В.* ПРАКТИЧЕСКОЕ ПРИМЕНЕНИЕ РЕЗУЛЬТАТОВ СПЕКТРАЛЬНОЙ ИНВЕРСИИ / В. Фагерева, А. Буторин, Ф. Краснов // В сборнике: Новые идеи в науках о Земле Материалы XIV Международной научно-практической конференции: в 7-ми томах. — 2019. — С. 220—223.

Список рисунков

Список таблиц

Приложение А

Листинг программного кода

Листинг программного кода для задачи оптимизации числа тематик в тематической модели.

Листинг А.1 Листинг программного кода

```
# -*- coding: utf-8 -*-
import numpy as np
import artm
from sklearn.metrics.pairwise import cosine_distances
from sklearn.metrics import silhouette_score,
    davies_bouldin_score, calinski_harabaz_score
import os
import sys
import warnings
warnings.filterwarnings("ignore")

wrk = sys.argv[1]

gdim = 100
vecfile = open(wrk+'.model.vec','r')
vtext = vecfile.read().split("\n")
vecfile.close()

vecs = dict()
for l in vtext:
  if len(l) < 1 : continue
  w = l.split()[0]
  v = l.split()[1:]
  vecs[w]=[float(i) for i in v]

text = open(wrk+".vw",'r').read().split("\n")
NR = 10
```

```
batch_vectorizer = artm.BatchVectorizer(data_path=wrk+".vw"
    , \
data_format="vowpal_wabbit", target_folder=wrk, batch_size
    =100)

dictionary = artm.Dictionary()
dictionary.gather(data_path=wrk)

metrics = [] #Все метрики по порядку.
for nr in range(NR):
  txt_ind = np.random.randint(0,len(text)-1,len(text))
  text_new = "\n".join([text[i] for i in txt_ind])
  fh = open(wrk+".vw%d" % nr,'w')
  fh.write(text_new)
  fh.close()

  batch_vectorizer = artm.BatchVectorizer(data_path=wrk+".
    vw%d" % nr, \
    data_format="vowpal_wabbit", target_folder=wrk,
    batch_size=100)

  os.remove(wrk+".vw%d" % nr)

  for T in range(5,20,5) + range(20,40) + range(40,71,5):
    N = 2
    ttopics = ["sbj"+str(i) for i in range(T)]
    ntopics = ["nz"+str(i) for i in range(N)]
    topic_names= ttopics + ntopics
    model_artm = artm.ARTM(num_topics=T+N, topic_names=
    topic_names)
    model_artm.initialize(dictionary)
    #model_artm.cache_theta = True
    #model_artm.cache_phi = True
    model_artm.scores.add(artm.TopicKernelScore(name="
    top_score", topic_names=ttopics,
    probability_mass_threshold=0.5))
    model_artm.fit_offline(batch_vectorizer=
    batch_vectorizer, num_collection_passes=10)
    model_artm.regularizers.add(artm.
    SmoothSparseThetaRegularizer(name='SparseTheta',tau=-1,
    topic_names=ttopics))
```

```
    model_artm.regularizers.add(artm.
    SmoothSparseThetaRegularizer(name='SmoothTheta',tau=1,
    topic_names=ntopics))
    model_artm.regularizers.add(artm.
    SmoothSparsePhiRegularizer(name='SparsePhi',tau=-1,
    topic_names=ttopics))
    model_artm.regularizers.add(artm.
    SmoothSparsePhiRegularizer(name='SmoothPhi',tau=1,
    topic_names=ntopics))
    model_artm.regularizers.add(artm.
    DecorrelatorPhiRegularizer(name='Decorrelation',tau=-1
    e3,topic_names=topic_names))
    model_artm.fit_offline(batch_vectorizer=
    batch_vectorizer, num_collection_passes=10)

    ts = model_artm.score_tracker['top_score']
    tls = ts.last_tokens
    #glove id
    y = []# для силуэта
    X = []
    glove_vec_topic = {}
    for topic_name in  ttopics:
      for w in tls[topic_name]:
        if w.encode('utf8') in vecs:
          gv = vecs[w.encode('utf8')]
          y.append(topic_name)
          X.append(gv)
          if topic_name in glove_vec_topic:
            glove_vec_topic[topic_name].append(gv)
          else:
            glove_vec_topic[topic_name] = [gv]
        else:
          pass #print w

    # centroids
    centroids = np.zeros((T,gdim))
    intra_dists = np.zeros(T)
    inter_topic_dists_list = []
    for tid,topic_name in enumerate(glove_vec_topic):
      centroids[tid] = np.mean(np.array(glove_vec_topic[
    topic_name]),axis=0)
```

```
      #inter topic dists
      glen = len(glove_vec_topic[topic_name])
      intra_dists[tid] = np.average( cosine_distances(
    glove_vec_topic[topic_name], [ centroids[tid] * glen
    ]).ravel())
      #inter_topic_dists += cosine_distances(
    glove_vec_topic[topic_name], [ centroids[tid] * glen
    ]).ravel().tolist() # dist with centroid

  extra_dists = cosine_distances(centroids,centroids)
  #extra_topic_dists = extra_topic_dists[
  extra_topic_dists != 0]

  #inter_topic_dists = np.array(inter_topic_dists_list)
  #inter_topic_dists = inter_topic_dists[
  inter_topic_dists != 0]

  score = (intra_dists[:, None] + intra_dists) /
  extra_dists
  score[score == np.inf] = np.nan
  cdbi = np.mean(np.nanmax(score, axis=1))

  metrics.append ((nr, T,  max(ts.average_contrast), max(
  ts.average_purity), max(ts.average_size),\
    silhouette_score(X,y,metric='cosine',random_state=42)
  , davies_bouldin_score(X,y), calinski_harabaz_score(X,y
  ) , cdbi))

  print metrics[-1]

np.save(wrk+'_metrics.npy',  np.array(metrics))
```

Saint Petersburg State University

Printed as manuscript

Fedor Krasnov

Methodology for Building a Digital Twin of a Scientific and Technical Center in the Oil and Gas industry

(Translation from Russian)

Science of Specialization 05.13.01 —
«System analysis, management and information processing
(technical sciences)»

DISSERTATION

seeking a degree of a Doctor of Science in Informatics

Academic Adviser:
Doctor of Science in Informatics,
Professor of Department of Computer Modelling and
Multiprocessor Systems
Alexander Degtyarev

Saint Petersburg — 2020

Table of content

Page

Page

Page

Preface

A qualitative leap in the structure and dynamics of the development of productive forces is provided by the activities of sectoral scientific and technical centers (STC). The number of STCs in the energy industry is growing from year to year, and as the reserves of easily extracted oil are depleted, the role of the scientific component in its production increases. Therefore, the effectiveness of the STC is a crucial characteristic that needs to be assessed and planned.

The questions concerning the methods of evaluation of STC, considered in this paper, allow us to determine the list of observed characteristics that provide a reliable assessment of scientific and technical centers and allow both to compare them and to build mathematical models for scenario planning of their effectiveness. Traditionally, STC was created according to the patterns of Russian research and design institutes, which estimated oil and gas reserves, put them on the state balance sheet and formed project documents for field development. The tasks of such institutions also included the development and implementation of new technologies and materials, but often manifested their main vulnerability - isolation from the business.

Initially, after obtaining permits from the Central Commission for the development of deposits, the STC stepped aside, and production workers came into action. Modern STC is a

research and design structure that is fully integrated into production. The evaluation of such STCs needs to be reviewed.

Thus, the object of research of this work is the results of scientific activity of STC. The subject of the study is the methods of measurement, evaluation, and planning of performance.

Chapter 1. Introduction

Many recent studies show a strong correlation between the rise in oil prices and the volume of investment in promising research and development of new technologies in the oil industry. The optimal range of oil prices for innovative investments can be recognized in modern conditions the range of 60-70 USD per barrel. At price values in the region of 50-55 and less than USD per barrel oil the industry falls into survival mode with the corresponding tight optimization of all costs. If the price is more than 80 USD per barrel occurs a well-known euphoric effect with a preference for investing profits in other sectors economies with expected rapid returns, particularly in speculative sectors financial instruments and markets. The situation is somewhat different for the Downstream industry, as expensive raw materials stimulate the need for more in-depth processing. However, currently, traditional the refining processes have achieved a certain technological the limit, and the introduction of new methods requires overcoming the known psychological barrier on the part of owners of oil refineries productions'. Sharp fluctuations in oil prices and their potential cartel decisions (e.g., OPEC) create a nervous background in the industry, which is not conducive to innovative financial investment.

Thus, financial investments in the development and development of new technologies are impulsive in time, tied to

oil price fluctuations. At the same time, development, testing, and implementation of new technologies require much longer time than the duration speculative business cycle of the hydrocarbon market. Moreover, many start-up stage technologies or even more Mature will require for its refinement and technical implementation of additional funds.

At not every peak of investment and innovation activity will bring funds to the budget for the development of this particular technology. Technological ideas are still quite a lot, also takes place a competitive struggle of scientific groups and directions for the allocated funds. Investors, for psychological and behavioral reasons, may invest the next tranche of investment in any new projects instead of projects under active development, but not yet demonstrated from management its practical efficiency.

Based on the above, it can be concluded that candidates for survival are technology projects that can be brought to the funds of the first investment tranche at least to stage of feasibility, and better until stage pilot plant.

The situation in the gas industry is somewhat different. Gas is a cheaper raw material, the process of its extraction and transportation is in a certain sense more technological, and the market is more stable due to large constant volumes of demand from the systems of power generation, domestic and industrial heating, production of high-grade process heat and well-known gas chemistry industries. However, these same factors at the same time and limit innovation activity and investment and in-

novation attractiveness in the gas industry. The development of gas chemistry concerning new gas processing technologies is attractive from a theoretical point of view and may in the future be worthy of competition to many traditional areas of petrochemistry. However, in practice, the technology of energy-efficient methane conversion has not yet been developed, and the existing technology through steam or steam-oxygen reforming can compete in costs with refining only at oil prices of 90 USD per barrel. As for the processing of higher hydrocarbons, they are to a certain extent developed, and the raw material for them is the expander selection (separation) of natural gas into fractions. However, the feedstock for the same group of technologies can also be used for the associated refining gases, primarily ethylene, which processes have been implemented in many refineries.

Separately, it should be noted the promising role of coal when used as a fuel, and chemical raw materials. The key in both areas is the industrial introduction of efficient gasification and pyrolysis technologies with a full cycle of conditioning and purification of the resulting product. Despite all the current situation in the present case on the hydrocarbon market, the use of coal remains essential in the long term for such industrialized countries as the USA, Germany, China, South Africa, and the Russian Federation. Ukraine and Kazakhstan. We have here not accidentally attributed the Russian Federation, Ukraine, and Kazakhstan to the industrialized powers, although someone can say that such an assignment is conditional. In-

deed, these states are at an economic crossroads but still, have both a reasonably powerful industrial potential and raw material capabilities. It depends on the weighted investment and innovation-technological policy of these states, and primarily in the fuel and energy sector of the economy, whether they will join the club of leading world economic players or will continue to be subject to disintegration and degradation processes.

In addition to short-term and long-term economic trends, social and political factors significantly influence the mining and processing of hydrocarbon raw materials, and in particular its innovative technological sector. So for the United States, as well as for transnational corporations in the hydrocarbon market, environmental problems are relevant, which can be divided into local (ecological impact in the places of direct extraction and processing of hydrocarbon raw materials) and global (greenhouse effect, pollution of the oceans, pollution of groundwater In particular, when applying new popular technologies for the production of shale oil and shale gas). For the countries of Eastern and Western Europe, the political problem of dependence on gas supplies from the Russian Federation and the search for alternative sources of fuel and chemical raw materials is relevant.

Therefore, the effectiveness of the development, development, and introduction of new technologies in the hydrocarbon industry should be assessed by a multi-criteria approach, which should take into account the conjuncture (determining current investments), economic long-term, technological and socio-polit-

ical components. This situation requires the use of multivariate analysis methods using the latest algorithms from the field of Data Mining, Big Data Analysis, neuroscience, machine learning methods, search, systematization, and analysis of digital artifacts of scientific and technical centers and laboratories, semantic and computer linguistic analysis of texts, etc.

For the Russian Federation, there is a list of specific problems that can be attributed to both the sociopolitical and the technological sphere.

To review these problems, we turn to a brief history of the oil and gas industry in the Russian Federation.

1.1 Why the Russian oil and gas industry?

The beginning of industrial oil production is considered to be the second half of the nineteenth century, but since time immemorial oil has been mined by the open method at its exit to the surface and used by people living in those areas for various purposes who lived in different parts of the world where oil seeped to the surface. According to written sources in Russia, the tribes living in the territory of the Timan-Pechersk district, in particular along the banks of the Ukhta river, collected oil from the surface of water bodies and used it as a lubricant, as well as for medical purposes. Oil from this region was first delivered to Moscow in 1597. The year 1684 is the date of a

report on the discovery of oil by the head of the Irkutsk prison, Leonti Kislyansky. In 1703, the first issue of the newspaper "Vedomosti" published a message about the discovery of oil on the Sok River in the Volga region. Later, there were reports of oil production by residents in the North Caucasus.

Locals extracted oil using buckets from wells 1-2 meters deep. The use of oil was mostly medical. The manifestations of oil and gas on the western coast of the Caspian Sea in the 10th century were reported by Arab travelers and by a historian as early as the tenth century. According to the data of the Italian historian and traveler Marco Polo, people in this region used oil for medical purposes and religious purposes. From the fourteenth century, oil from the Caspian coast was supplied to the countries of the Middle East.

The first attempt to organize the oil refining industry could be attributed to 1745, when a native of Arkhangelsk, Fedor Pryadunov, received permission to extract oil on the Ukhta River in the already mentioned Timan-Pechersk district. Pryadunov also created an oil refinery and some oil refineries supplied to Moscow. However, this technology has not received further development, because, throughout the XVIII century, the practical use of oil and products from it remained extremely narrow. This situation has not changed significantly in the first half of the XIX century. Nevertheless, the commissioning of the Dubinins Brothers' oil refinery, the raw material for which was the oil from the open Voznesensky field near the city of Mozdok, dates back to 1823.

The expansion of the Russian Empire to the Caspian region at the beginning of the XIX century and the accession of the North Caucasus designated these two regions as the main ones regarding oil. The world's first exploration oil well was drilled in the Bibi-Aibat field of the Absheron Peninsula (near Baku) in 1847, which was more than a decade ahead of the start of the US oil industry. However, the first full production well close in its structure to new wells was put into operation in the Kuban on the river. Kudako in 1864.

1849 can rightly be considered a turning point in the global oil industry since Canadian geologist Abraham Gesner received kerosene from oil this year as a stable product with reproducible properties. In 1853, Lviv pharmacists Ivan Lukasevich and Jan Zeh invented a safe kerosene lamp, which marked the beginning of an era of widespread consumption of oil.

A direct oil refinery for the production of kerosene was launched in Baku in 1863 under the supervision of engineer David Melikov. A few years later, he also founded an oil refinery in the city of Grozny.

Meanwhile, the first well was drilled in Pennsylvania in the USA in 1859 and oil production begins. The oil field is developing rapidly, and the oil is transported in standard wooden barrels of 42 gallons or 168 liters, initially intended for carrying salted herring. So there is a measure of the volume of oil 1 barrel, equal to 42 gallons. In 1865, the world's first oil pipeline with a capacity of 2,500 barrels per day was built to transport oil from the oil wells to the Miller Farm Station railway station.

This node also served as a prototype for oil-loading transport terminals and a cluster (flowering) scheme for combining oil flows from several nearby wells before transporting oil through the main oil pipeline.

In 1870, Rockefeller founded the company Standard Oil, whose share in US oil production in less than ten years has grown from 10% to 90%, which led to the introduction of the antimonopoly law for the first time in the world.

It is interesting that in 1871 Ivan Mikhailovich Gubkin (1871-1939) was born in Russia - one of the founders and creators of petroleum geology as a separate section of general geology. Gubkin made an almost invaluable contribution to the development of the Russian oil industry, and today his name is given to the Russian State University of Oil and Gas.

In Russia, in the area of Baku, the first oil pipeline was commissioned in 1878. Unlike the United States, he connected wells with an oil refinery. Moreover, in 1877, Russia for the first time in the world mastered the use of oil tankers (tankers) for the transportation of oil.

Initially, the state in Russia was a monopolist in the oil industry, but by the end of the seventh decade of the 19th century, foreign companies were allowed to oil production. A large concentration of fields with easily recoverable reserves of oil was found on the Absheron Peninsula, but the transportation of oil and refined products to the final consumer was utterly unregulated. One of the critical achievements of the Nobel brothers and the Rothschild family in Russia was precisely the unifica-

tion of oil production, oil refining, and transportation of oil and oil products to end users within the framework of single commercial companies. It was in Russia in 1874 that the first vertically integrated oil company appeared - the Baku Oil Society. During this period, the oil industry of Russia showed significant growth, and by the beginning of the twentieth century, Russia's share in total world oil production was about 30Interestingly, Shell Transport and Trading, which later became part of Royal Dutch-Shell, at the first stage of its operations transported Baku oil from Russia to Western Europe.

The processes of oil production and refining did not remain outside the sphere of interests of the Russian science of that time. Chemist Zelinsky, mathematicians, and mechanics, L.S. Leibenzon, I.P. Moskalkova, I.A. Charny, V.N. Schelkacheva, Ya.I. Hurgin and many other now recognized classics.

The basis of petroleum science are the achievements of organic chemistry, as well as the apparatus of theoretical mechanics, soil and rock mechanics, and hydromechanics. It was developed and achieved high perfection, the apparatus of partial differential equations describing the transfer of fluids in porous media based on phenomenological concepts, such as Darcy's law.

Dmitry Ivanovich Mendeleev played a significant role in the development of the science of oil in Russia. In the early 1990s, the bulk of the scientific interests of the scientist were related to petrochemical and oil refining issues. So Mendeleev proposed a method of continuous crushed distillation of oil, ana-

lytical methods for determining the composition of the products of oil distillation, suggested the use of selective solvents. He tirelessly argued the need to use all fractions of oil, including the heavy ones. They were suggested to use tanning oil instead of kerosene in lighting lamps. He also contributed to the construction in the city of Rybinsk, thanks to which instead of an annual loss of about 100 000 rubles in the prices of that time (costs for the purchase of lubricants) Russia soon acquired several million rubles annually from the export of such lubricants.

Mendeleev opposed the preferential use of oil products in the boilers of steam boilers. "It is possible to drown with banknotes," he wrote in one of his economic articles, justifying the expediency of using oil as chemical raw materials, and coal as fuel.

Back in 1881, Mendeleev proposed to study the possibility of deep thermal processing of oil by passing it through pipes with a temperature of 300–400 degrees Celsius. He assumed that such heavy refining should be subjected to such heavy distillation, in order to obtain from them an additional quantity of suitable products. These ideas were all the more critical because Russian oil was denser than American oil and more heavy oils and other residues remained from its distillation. Mendeleev was the continent of the abiogenic concept of oil formation through the interaction of hot iron and nickel carbides with water in the early geological epochs of the Earth.

Mendeleev paid great attention to the rational organization of the production cycle of oil production and refining. He

proposed to place refineries not only near wells (fields) but also on the banks of the Volga, where at that time there was a large concentration of industrial production. With his participation, one of the oldest Russian refineries in Yaroslavl was founded.

Known controversy with Nobel, who was a supporter of the widespread use of oil as fuel, and also often gave orders to pour out distilled gasoline, because merely there were not yet sufficient applications for it at that time. It demonstrates the opposite of scientific-technological and economic-conjuncture concepts in evaluating production efficiency, which was discussed in the first sections of this chapter.

Mendeleev advocated the construction of the Baku-Batumi oil pipeline and kerosene pipeline. He wrote: "With the pipeline, the demand for crude oil increases, and the prices for it will be settled, because new sales sites will appear, and therefore new drilling rigs will appear in the city Baku itself and in other places in the Caucasus, which should be the case".

The invention in the 90s of the nineteenth century of internal combustion engines, in particular, the diesel engine, and the emergence of the automotive industry further increased the demand for oil and led to the development of technologies for more advanced oil refining. Along with kerosene, such fractions as gasoline and ligroin appeared. The remains of oil refining were used as lubricating oils in machines and mechanisms.

However, the dramatic events in Russia related to the First World War and the 1917 revolution led to a drop in oil production and the loss of Russia's dominant position in the

hydrocarbon market. If in 1913 more than 9 million tons of oil were produced in Russia, then in 1920 this figure decreased by more than 40%. The countries of the Entente tried to separate the oil-bearing regions from the territory of the Soviet Republic but ultimately suffered defeat. As a result, in 1920, the Nobel brothers sold a significant part of their Russian assets to Standard Oil from New Jersey. Later this company became the basis of Exxon. Standard Oil opposed the decisions of the Soviet government on the nationalization of oil fields and refused to cooperate further with the Soviet authorities.

On the contrary, the New York oil companies (later transformed into the Mobile company) continued making investments in the Russian oil industry, so that by 1923 the export of oil and oil products from Russia had reached the pre-revolutionary level again.

Thus, already in the 20s, a partial dependence of the Russian (Soviet) oil industry on Western capital and western technologies was formed. The Soviet government decided, in particular, to intensively train its personnel in the field of petroleum engineering and geological exploration.

Ivan Mikhailovich Gubkin played a considerable role in the implementation of this program— organizer of Soviet petroleum geology, academician of the USSR Academy of Sciences (1929), vice president of the USSR Academy of Sciences (1936), chairman of the Azerbaijan branch of the Academy of Sciences CCCP (1936 — 1939), laureate of V.I. Lenin's awards (1931), Deputy of the Supreme Soviet of the USSR of the 1st

convocation (1937). In contrast to D.I. Mendeleev, I.M. Gubkin was a supporter of the theory of the biogenic production of oil. In particular, he wrote: “We believe that oil formation, starting from the decomposition of fats in biogenic sludge before its burial, continued even after its burial, with the active assistance of anaerobic bacteria during the entire period of diagenetic rock change”. Unfortunately, the theory of oil formation by IM Gubkin remained unknown in the framework of world science, since the works of Gubkin at that time were not translated into foreign languages.

In 1930, under the leadership of I.M. Gubkin, a textbook “The Teaching of Oil” was published, according to Gubkin himself “outlining the main issues of petroleum science”. The basis of the textbook served as a course of lectures of Gubkin himself. However, the materials of other authors were widely used. So A.I.Kosygin was the author of the section “Basic techniques of oil field exploration”, and geophysicist A.I. Zaborovsky wrote the chapter “Elements of geophysical methods of exploration”.

Alexander Ignatievich Zaborovsky doctor of physical and mathematical sciences, a geophysicist. He was one of the founders of the Soviet school of geological exploration geophysics and the developer of a program for training specialists in universities in this area. Zaborovsky - author of the monograph “Geophysical methods of exploration”, which was used in educational institutions of the USSR as a textbook on applied geophysics.

In 1919 - 1926 Zaborovsky magnetometric works on the Kursk Magnetic Anomaly. He worked in the same team with PP Lazarev, A. D. Arkhangelsky, I. M. Gubkin, O. Yu. Schmidt and other prominent Russian scientists of the time. As a result of the activities of this group, significant accumulations of ferruginous quartzites were found on the territory of the Kursk Magnetic Anomaly, and according to the estimates made, the total amount of iron in this field exceeded the total reserves of iron, explored by that time in Europe.

In 1926, Zaborovsky developed many geophysical methods based on seismic data. Since 1929, he taught courses in geological exploration geophysics at Moscow State University, and since 1930 he headed the faculties and department of geological exploration geophysics created by him at Moscow State Geological Prospecting University. In the periods from 1944 to 1949 and from 1954 to 1968, Zaborovsky also headed the department of geophysical methods at the geological department of Moscow State University.

Even in these two examples of figures of Soviet geological science, we see that in the 30s, 40s, and 50s, along with practical achievements and theoretical developments, considerable attention was paid to the training of qualified personnel for the industry.

Up until the beginning of the Second World War, the Caspian region and the North Caucasus remained the main areas of oil production and the oil industry. One of the main strategic tasks of the command of Nazi Germany was the cap-

ture of these oil-bearing areas. It is known that Germany does not have its oil reserves, so Hitler went to war with gasoline produced from acetylene, which in turn was obtained complexly and expensively of electric arc pyrolysis of coal in inert gases. After the war, oil production in the Caspian region increased again and in 1951 reached a record level of 850,000 barrels per day. In addition to the actual oil production, Baku has become an industrial center for the production of equipment for oil production and petrochemistry throughout the USSR. However, the Soviet government began targeted work on finding new deposits, primarily in the Volga-Ural region, in which initial exploration was carried out back in the 30s. The advantages of the deposits of this region were their low geological complexity and proximity to the nodes of the transport infrastructure. From the mid-50s, production from the fields of the Volga-Ural region amounted to 40% of the total oil production in the USSR for that period. The extracted oil was sent for processing to new plants. An interesting fact is that one of the largest in the world for that time, the Omsk Oil Refinery, which was commissioned in 1955, is located in Western Siberia, which is itself an oil-bearing region, initially used raw materials from the fields of the Volga region.

However, the Volga oil was inferior in its properties of Baku and North Caucasus one. This situation stimulated a new round of research in petrochemistry and oil refining.

In the 1930s, oil and gas fields were searched in the Yelshano-Kurdyumskaya gas-bearing area in the Saratov Re-

gion. In 1941, in the area of the village of Elshanka near Saratov, the first gas well was drilled with a daily production of 800 thousand cubic meters of gas. In June 1942, another well was drilled, which, like the first one, turned out to be highly productive, which allowed specialists to conclude the discovery of a field with commercial reserves of natural gas. These dates can be considered the dates of birth of the gas industry of the USSR (Russia). Since 1942, the gas produced from the wells was sent to supply the Saratovskaya TPP, for which in October 1942 the gas pipeline "Elshanka — Saratov" 16 km long was built. Before the start of natural gas production in the USSR, luminous gas, produced by the conversion of hot coal with water vapor, was used at the production as a combustible process gas. Natural gas turned out to be much more technological and less toxic than illuminating gas. The composition of which is carbon monoxide CO. Following the power plant and at other enterprises of Saratov began the use of natural gas to produce process heat and for space heating.

In 1943, another gas field was discovered near the Kurdyum settlement in the Saratov Region with one million cubic meters of gas per day, and in 1944 significant reserves of gas — 6 billion cubic meters — were discovered in the region. At the end of 1944, the USSR State Defense Committee decided to build an 843-kilometer-long gas pipeline "Saratov - Moscow" to provide gas for the industry and population of the capital.

Up to 30 thousand people worked daily on the construction of the facility. Dozens of engineering, instrument-making,

heavy engineering, electrical and other industries produced almost 9 thousand names of various equipment and materials required by the pipeline. The gas pipeline has become an experimental testing ground where new technologies were developed. Here, the flow-rate method of conducting linear work was first applied, construction mechanisms and devices for route operations, gas welding units were tested, butt welding of thin-walled high-pressure pipes with a wall thickness of 6.25 mm was tested in practice.

The development of the gas industry of the USSR (Russia) was further noted by such milestones as the construction and commissioning of the gas transmission system "Central Asia - Center", which connected the gas fields of Turkmenistan, Kazakhstan, Uzbekistan with industrially developed areas of central Russia, the construction of Orenburg gas processing plant. In the late 1970s, the construction of the Urengoy-Pomar-Uzhgorod gas pipeline laid the foundation for the export of Russian gas to Western Europe.

1.2 Why Science and Technology Centers?

Scientific and technical centers were established for the scientific, technical and engineering support of the gas industry. Such of them as VNIIgaz, VNIPIgazpepererabotka is still operating organizations.

Chemical processing of natural gas is mainly associated with the processes of producing methanol and nitric acid. Relevant technologies, including catalysts for all stages of the processes, were developed in particular for Novomoskovsk (the NIAP Institute, now part of Alvigo).

In the 50s and 60s, exploration and commissioning of the oil fields of the European North of the USSR (the Komi Republic, Timan-Pechora basin) continued. Construction of the oil pipeline transportation system has begun. The growth of oil production opened for the USSR the possibility of increasing exports and strengthening its position in the international market. In the 60s, the USSR took the second place among oil exporters in the world, having pressed Venezuela. Already at that time, the negative trend of exports of predominantly crude oil, instead of the value-added products of oil refining, that persisted in modern Russia, was outlined. The dumping oil prices set by the USSR on the world market ultimately led to a conflict between Western oil-producing companies and the governments of the Middle East countries, where the main oil fields used by the West at that time were located. The governments of the Middle East countries have established the Organization of the Countries of the Oil Producers (OPEC) to resolve this range of issues. The 1972 Arab-Israeli conflict further aggravated the situation. The USSR sided with the Arab countries, not least because of the considerations of retaining dominant positions in the oil market. Interruptions in oil supplies to Western coun-

tries led to the start of oil production by Great Britain and Norway on the North Sea shelf.

The flowering of Soviet petroleum science in all three sectors — Upstream, Midstream, Downstream — also belongs to this period. As is known, according to the classification adopted in the west, the full production cycle of oil production and refining is divided into three parts - Upstream, Midstream, Downstream. Upstream includes oil production processes and, more generally, all technological processes associated with the exploitation of fields. By Midstream are the processes of preparing oil for transportation and the actual transportation. Midstream processes encompass the operation of a pipeline transportation system for transporting oil. Downstream processes are associated with the refining of oil at refineries (refineries). The focus of this work is Upstream technologies.

1.3 Upstream technologies

Geophysical models of deposits and oil-bearing formations and gas-hydrodynamic models of oil production processes have been intensively developed in the joint works of the Moscow and Makhachkala schools of mathematical physics. We can mention, for example, the works of Kholodov, AS and Magomedova K.M. with employees in the field of the numerical solution of

multidimensional nonlinear equations of gas dynamics and hyperbolic hydrodynamics. The Siberian Branch of the Academy of Sciences of the USSR, primarily the Institute of Catalysis named after Boreskov, is becoming one of the centers of catalytic chemistry and its applications in oil refining. Branch research centers in Yaroslavl, Sterlitamak and Nizhny Novgorod are beginning intensive work in the field of catalysts based on artificial zeolites. Large-Scale industrial production of zeolites is mastered. General and specialized issues in the industry, including the operation and diagnostics of oil pipelines (and gas pipelines), are dealt with by industry centers in Krasnodar, Saratov, Ufa, VNIIGAZ in Moscow. In the future, it is planned to extend geological exploration to the bottom of the sea shelves. Organizations such as "Yuzhmorgeo", "Soyuzmorgeo" and others are starting exploration on the shelves of the Black, Okhotsk and Japanese seas. Naval expeditions are being organized for reconnaissance in the South China Sea in the framework of cooperation with the Republic of Vietnam.

At that time, the need to organize dedicated scientific and technical centers (STC) in the industry and not just design institutes, was finally clarified. STC was a potential intellectual accumulator, and a bank of intellectual values, which could use for technological development and re-equipment of the industry as a whole. The disadvantages of this period include the fact that the achievements of sectoral science often remained limited not only by industry but even by the territorial administration. The achievements of academic science, in particular

in the field of oil refining and petrochemistry, were not introduced since industry associations did not have the necessary incentives for this.

Meanwhile, the USSR began to develop fields in Western Siberia. The high level of production, determined by the large volumes (reserves) of individual fields and the relatively small costs of production, was one of the critical factors in the emerging decline of the USSR oil industry. In the wake of success, the costs of exploration and development of new fields, as well as the improvement of oil production and refining technologies, were reduced. Guided by the priority of maximizing the volume of oil production in short rather than in the long term, the Soviet planning authorities encouraged production associations to extract as much oil as possible from fields already developed without taking into account the consequences for the state of the fields. An excessive number of wells were drilled in each developed field, and an enormous amount of water was pumped into the oil-bearing formation. As a result, by the mid-70s of the last century, the USSR was confronted with a sharp drop in recoil from operating wells in western Siberia. The government of the USSR managed to halt this negative process through massive investments in geological exploration and the commissioning of new fields, but this gave only a postponement due to the failure in the development and implementation of new technologies throughout the technological cycle. Strangely enough, it was during this period that new promising ideas for

automating the processes of drilling and oil production arose and developed in the oil industry of the USSR, in particular:

- automation of the drilling process, automatic control of parameters of a drilling rig, power consumption, rock resistance;
- prediction and prevention of emergency conditions and breakdowns, optimization of the distribution of labor and material and technical resources during repairs;
- automated diagnostics, power consumption monitoring and emergency mode prediction for deep sucker-rod pumps and oil transfer pumps.

The next drop in production fell in the period from 1982 to 1986 and, thanks to the political crisis and the collapse of the USSR, smoothly flowed into the decline of the oil industry in the 1990s. Disintegration processes caused a sharp drop in the demand for oil in the domestic market; also, domestic oil consumers often could not pay for the consumed raw materials on time. Opportunities for oil exports remained limited; besides, financial flows from export sales passed through the hands of financial monopolists and near-criminal structures, so that mining companies got the minimum amount of proceeds from the sale. The result of all these negative processes was a further drop in oil production, which stopped only in 1997.

This difficult period in the history of the oil industry in Russia is marked by some promising developments in the field of catalytic cracking of oil, including those introduced into production processes at Russian refineries. Such business interest

in technology was caused by the fact that the refineries, being located at the end of the process chain, mostly felt the negative impact of the above destructive economic and political processes.

According to independent expert organizations, oil reserves in Russia are still quite significant. So the remaining reserves in the region of Western Siberia are estimated at more than 150 billion barrels (more than 20 billion tons), while the level of production can be increased two to three times compared with the current. However, oil production is complicated by difficult geological conditions, since the deposits in this region are usually several oil-bearing strata.

All this will require investing both in geological exploration of new and a priori refinement of the profiles of fields already in operation and in improving oil production technologies, including automation, as well as the development and use of sophisticated digital models of production processes with their direct adaptation to the operation of specific fields.

The estimate for the European North of Russia (the Timan-Pechora basin) is nine billion barrels (1.25 billion tons). The indicated reserves relate mainly to hard-to-recover, oil, by its qualitative properties, refers to heavy oil. Also, the development of oil production in this region is complicated by the harsh climatic conditions and the degradation of the transport system of the Soviet era. Nevertheless, the potential of this region, as well as the Volga-Ural region, the estimate for which amounts

to a value comparable to the estimate for the Timan-Pechersk basin, should not be discounted.

Estimated remaining reserves in the region of Eastern Siberia amount to three billion barrels (0.45 billion tons), but the available geological exploration data are not enough for more accurate estimates, with the result that real oil reserves may be several times larger. The development of oil fields in this region is hampered by both geological reasons and the remoteness of fields from sales markets and the reduced level of development of transport infrastructure in the region.

Recently, both the Russian government and Western oil corporations have been showing interest in offshore fields - in the Kara Sea and near Sakhalin Island. The development of these fields is constrained by high capital intensity, but the decisive factors include the possibility of direct transportation of mining products by sea using tankers.

The history of the development of the oil industry in Russia (USSR) broadly follows the trends described in the first sections of this chapter, with the only difference that in a planned economy, the predictor of the market price for oil should be replaced by the profit from oil production (revenues fewer costs). In the most favorable periods, the government did not pay enough attention to the development and introduction of new technologies, and in critical periods the rate was often placed on the import of ready-made technologies from abroad. Dependence on foreign technologies for the Russian oil industry remains critical even now. Thus, the announcement

in the Financial Times on October 30, 2014, about the withdrawal of foreign oil companies from Russian projects plunged Russian officials and company executives into depression and pessimism. Indeed, in 2014, ExxonMobil closed ten joint ventures with Rosneft. Other western companies (both Shell and Total corporations, as well as mid-level companies specializing in equipment maintenance and engineering support) also minimize their activities in Russia. According to experts, these trends create additional obstacles in the first place for the development and development of new deposits.

In part, an interview with "Oil and Gas - Eurasia" magazine with the CEO of "Gazpromneft STC" Mars Magnavievich Khasanov was devoted to answering the challenges of this situation.

LLC "Gazprom Neft Scientific-Technical Center" was established on October 30, 2007. The company was established to improve the efficiency of field development and the development of the mineral resource base of PJSC 'Gazprom Neft". The main activities of the Scientific and Technical Center are the design, analysis, and monitoring of oil field development and geological exploration, geological and hydrodynamic modeling, technological support and operational control of drilling. STC's responsibilities include the creation and maintenance of a corporate base of geological field information, managing the process of extracting oil from the subsoil using permanent geological and technological models, planning and organizing pilot projects to introduce new technologies in oil production.

Also, LLC "Gazpromneft STC" performs a range of works on the development, examination, and protection of project documentation to fulfill license obligations, carries out planning, analysis, and maintenance of exploration, conducts training and retraining of specialists of PJSC "Gazprom Neft".

According to Mr. Khasanov, one of the priorities of LLC "Gazpromneft STC" is interaction with leading Russian universities and the involvement of young specialists in cooperation. So the efforts of LLC "Gazpromneft STC" created a laboratory center at St. Petersburg State Mining University, and at the RSU of Oil and Gas. With the participation of Gazpromneft STC, the department of geology of hydrocarbon systems, jointly organized by the university and Gazpromneft STC, was opened. The "Oil Engineering" specialization is also open at the Moscow Institute of Physics and Technology. The Scientific and Technical Center has established scholarships for postgraduate students and undergraduates who complete the program and participate in research.

At the same time, the general director of LLC "Gazpromneft STC" notes that "today all technologies are available on the market, you can buy any of them". The competitive advantage of the oil company in the modern world is not the availability of its technologies, but the ability to correctly select and apply these technologies, to continually improve their level. Successful companies differ from the rest in that they correctly apply technologies, use their potential by 100% and change them on time. Also, Mr. Khasanov notes: "As for tech-

nologies, STC is often the design office for creating a pipeline for their implementation, the definition of technological calls and their ranking, the introduction of technology into production according to the project principle". Thus, conceptually, the position of the General Director of Gazpromneft STC LLC is in line with the approaches of the Soviet government in the 20s of the previous century - the import and adaptation of technologies and the cultivation of their staff.

At the same time, in the conditions of modern Russia, the introduction of advanced foreign technologies is complicated by a gap in technological structures. The implementation should be comprehensive, targeted, project-oriented and problem-driven to prove the effectiveness. And in this process, the role of scientific and technical centers of the oil and gas industry should not be underestimated.

1.4 Industrial value chains

As noted earlier, scientific and technical centers (STC) are primarily designed to serve as centers of competence, combining responsibility for exploration, reserves assessment, primary identification of parameters of newly developed fields, construction and commissioning of fields, monitoring, control and management of oil production processes in the fields in order to maximize oil recovery, optimization of capital costs and

operating costs, selection of equipment and technologies, the introduction of new technologies and the formation and implementation of testing programs of new technologies with the dissemination of experience in other production units of the company.

The concentration of intellectual values, functionally-oriented knowledge, high-performance computing resources, and qualified personnel within the STC allows servicing a large number of geographically remote fields in almost real time.

At the same time, real-time systems for managing the development, drilling and oil production processes with remote access to equipment and measurement and sensor systems of the fields are already being implemented and used by STC. Such systems will allow for the development and operation of geologically complex fields to quickly attract the full potential of geological, hydrodynamic and 3D-modeling, available to specialists of STC, including in the form of computer information and analytical tools, specialized application software, databases and knowledge, combined with expert systems with elements of artificial intelligence, including neural network technology and machine learning elements.

The typical chain of occurrence of intellectual values in the oil and gas industry demonstrated on Figure 1.1.

In the first model chain, the initiating factor is the problem that arises directly from the development or operation of a particular field. However, the problem should become a typical one, i.e. typical for several fields, or for one colossal oilfield,

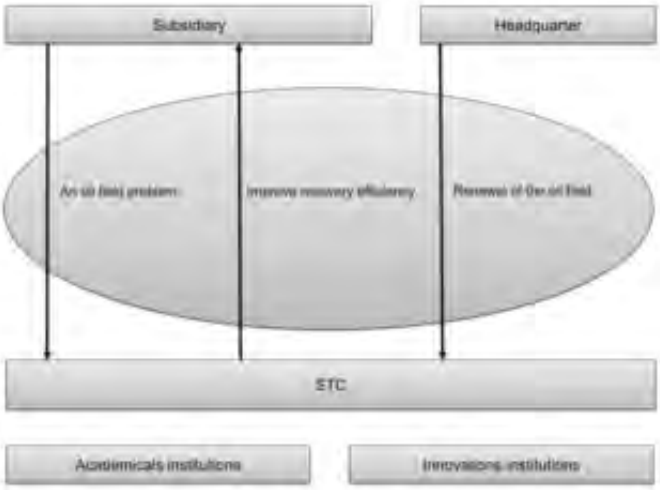


Figure 1.1 — Industrial value chains.

and have a significant impact on the oil production process, so that the company's management decides to order the relevant research in any STC, after which intellectual values arise, ready for subsequent use in other fields and other situations. Another option is the situation when in the course of standard design and service engineering works ordered by the oil company for the development, arrangement, and operation of a particular field. STC specialists make a predictive forecast based on a posteriori models built on the data and knowledge available to the STC and recommend the Customer to take specific preventive measures and proactively carry out the necessary engineering, geological and technical measures to ensure the further sustainable operation of the field with the high index of recovery.

In the second scheme, the company's management is the initiator of research and development and implementation of the technology. As a rule, we are talking about the commissioning of sites recognized as unprofitable in the framework of the use of oil production technologies, for example, low-productivity areas with low permeability, fractured reservoirs, low-debit wells, which require, in particular for their development and operation of the intelligent adaptive control of the production process.

In both of the described flow of intellectual values is of the order of the research and development of STC can be built in the purchase of already existing technology on the external market. However, even in this case, the role of STC is critical regarding technology adaptation, its implementation in a particular field, the collection, and analysis of primary data on the use of technology and the problems arising in this regard.

It is also evident that the transfer of a single technology can be difficult or even impossible due to fundamental differences in the technological structures of the Russian and foreign oil industries. Therefore, it is necessary first of all to provide a purposeful step-by-step transfer of the optimal technological environment, and for this, it is first of all essential to introduce modern concepts of business processes organization and to follow the main current trends steadily. Only within the framework of the updated conceptual understanding, the explication of individual technological processes will not remain isolated inclusions, will not dissolve over time, but will serve as the embryos of a new phase, around which the effective crystal-

lization of a new technological structure will begin. Indeed, the processes of standardization and innovative integrated technological training of personnel play a significant role. It is possible to direct the inevitable influence of the human factor in the introduction of new innovative technologies in the right direction only by nurturing the project and corporate culture of a new level.

Also, the introduction of almost any modern technology should be accompanied by changes and a presentation of new technologies in its sphere, the information environment of the company, system and application software. The conceptual priorities here are well known and clearly defined:

- Comprehensive digitalization (digitalization) of the oil industry from "Digital field" to "Digital electronic oil company";
- Application of artificial intelligence systems with elements of neural network technologies and machine learning algorithms for process control;
- Wide implementation of Big Data concepts and methods, including cloud technologies, analytical tools, and specialized software.

In essence, we are talking about the introduction of Industry 4.0 concept in Upstream. It is necessary to create conditions for the widespread of digital culture, as well as to ensure a direct interest in the successful digital transformation on the part of employees of all levels and specializations, but above all the top management of the company. It is necessary to fully involve

representatives of the “digital generation” in the production process - specialists in Big Data, neural networks, cluster analysis, machine learning methods, as well as existing employees of the company who are loyal to the dynamic digital ecosystem and other employees who can safely work in a vibrant digital ecosystem. After that, measures can be taken to speed the death of old “non-digital” approaches.

As is well known, the smart digital oil and gas field is a system of automatic control of oil and gas production operations, providing continuous optimization of the integrated field model and production management model. Due to the complexity and fuzziness of geological models (as part of the integrated model) to build a fully automatic control of oil production in the foreseeable period is impossible. But it is possible to use this standard for the formation of goals for programs to reduce the human factor in the management of the life cycle of deposits.

Intellectual digital field (IDF)- is a class of asset management systems of oil-producing enterprises, built by a formal, integrated asset model, processed by an automated control system that ensures optimal management at all levels of the enterprise while controlling the goals set by the owners of the asset. The term is based on the concept of intelligent management. An analog of IDF is the Digital oil field (Digital OilField, DoF), integrated operations management at the area. A particular concept of this term is “smart well”.

– formalization of the field information model;
– control apparatus;

- the most accurate feedback interfaces (sensors, communication);
- interfaces to optimize processes, models and criteria.

The necessary conditions for the existence of the intellectual field is:

To ensure the integrity of field management, an integrated asset information model should include and integrate all aspects of existing asset knowledge, including submodels such as:

- geological model;
- geographic model;
- technology model;
- logistic model of supply chains;
- economic model;
- financial model.

The introduction of an intelligent digital oil field is based on open standards ISO 15926, ISA-95, ISA-88.

The intelligent digital field includes several control circuits, first of all:

- The operational loop which provides control over the efficiency of processes for managing field operations (production, monitoring, and control of operating modes and condition of equipment, auxiliary processes, etc.));
- Modeling circuit that provides a dynamic development management model under varying external (context) and internal (content) conditions.

However, the process of digitalization (digitalization) faces some organizational, administrative and behavioral-psychological obstacles.

1.5 Big Data in oil and gas industry

The importance of taking into account the randomness factor is also confirmed by other promising works on accounting for randomly changing dependencies between permeability and porosity of the formation. As predictors for the establishment of such dependences for a particular field, in addition to well logging data, it is proposed to use the division into zones with approximately the same conditions of sedimentation. In the framework of the model used, it is considered that the statistical regularities for porosity and permeability for each of these zones and their parts are the same.

Thus, we see that digitalization (digitalization) is faced in Russian conditions not only with the established administrative-organizational and behavioral aspects of activity in the oil industry but also with physics, the reason for which is the existence of random fields. This means that the physics and models of geophysical environments, although they cannot be entirely attributed to the digital stage and the Big Data era, still have a chance of survival in the transition to Industry 4.0.

Together with physics have a chance of survival, at least at the initial stage, and the Russian STC in the oil industry

For this reason, the analysis of the activities of Russian STC in the oil industry is still of great interest, including in the context of innovative and technological solutions developed by these STC. To be entirely carried out, such an analysis requires the development of integrated criteria for the effectiveness of the STC.

The Big Data approach is characterized by exponential growth in the number of measurement operations and their corresponding data. At the same time, both the data itself and the algorithms for processing them, implemented in the form of information and analytical computer systems using network and cloud technologies, along with human resources, technological know-how, and capital, become one of the main assets of industrial, including oil companies. Data Mining analytical tools are at the same time the essential tools to achieve a competitive advantage in the market. Systems for the collection, primary processing, storage and security of information are also necessary. It is not surprising that many experts and analysts say now: "we have realized that there are terabytes of information around us, and now we need to understand what to do with these terabytes".

It is possible that the exit is a departure from traditional computer architectures in the direction of neuromorphic computational and analytical systems equipped with deep machine learning algorithms.

It is also necessary to use the methods of probabilistic programming based on the Bayesian inference since often a large proportion of sensors that monitor the digital field, characterized by a situation where the random spread of the observed values of the measured values is comparable to the price of the division of the measuring instrument. The reason for this is an attempt to control complex technological processes by controlling an increasing number of degrees of freedom of complex distributed systems. A corresponding increase in the number of signals from measuring devices with the simultaneous requirement to increase the accuracy of measurements leads to a decrease in the useful signal/noise ratio for a large proportion of measured values and parameters. There are cases when a further increase in the accuracy of measurement of a particular technological value (parameter) is impossible due to the achievement of the physical and technical limit for this method of analysis or too costly. At the same time, increasing the useful signal/noise ratio for this type of measurements carried out in real conditions is also either technologically impossible or very costly. As a result, an increase in the total number of measured parameters of the system or process does not lead, starting with a specific limit value, to an improvement in the accuracy of monitoring data and the quality of control and management. The output is seen in the development and use of hierarchical adaptive controllers based on fuzzy logic, as well as neural network systems, including algorithms for deep multi-layer learning with elements of formation of abstract clusters

of data within the neural network (deep learning technology and neuromorphic computing). A neuromorphic computer network, equipped with a specialized algorithm of configuration and training, is able in the long term to absorb the flow of Big Data generated by sensors and sensors of the digital field and subject this flow to multi-criteria analysis, separating essential data from non-essential. In the same vein, it should be considered, and the prospects for the machine to machine communication, when armed with neural processors of the device along the oil transportation line from the well to pumps share data with the purpose of optimization of technological modes and predictive forecasting of adverse, unplanned and potentially dangerous situations. This approach is entirely consistent with the concepts of *Internet of things* and *Industry 4.0*.

Another approach to the utilization of Big Data generated by a digital (digitalized) field is to apply a universal code based on the Shannon information entropy theory and the Laplace and Krichevsky predictors to the archiving data flow based on the identified comprehensive system for this flow.

This kind of ideas refers us to the works of the late Soviet period (late 70's-early 80's of the last century) in the field of automation of the oil and gas industry. Then, for example, the approaches of constant diagnostics of drilling equipment or deep sucker rod pumps based on continuous wattmeter were proposed, as well as algorithms for energy-efficient control of facilities due to adaptive speed control of electric drives of various types of equipment. Of course, in modern oil production and

drilling equipment, many of these principles have already been implemented in different versions. Thus, to ensure high-precision drilling of horizontal shafts with multi-stage hydraulic fracturing, Gazprom Neft has created a center for geological support of drilling, whose specialists manage drilling in on-line mode under remote access using rotor-controlled systems. In particular, similarly, the latest digital technology is found and combined in a real field with electro-mechanistic technology.

We cannot ignore supercomputer computing technologies and their application in the oil and gas industry. Over the past 20 years and to the present time, foreign countries, together with oil and service companies, have made significant efforts to stimulate research and applied for work on the long-term development and effective implementation of high-performance information and computing technologies to solve computational problems in the search, exploration, and development of hydrocarbon deposits. As a result of this activity, foreign oil and service companies gained competitive advantages and were able to oust Russian companies in the market of oilfield services significantly. Including the production, sale and maintenance of software, production and use of supercomputers (high-performance computing systems), which led to the technological dependence of Russian organizations, a high level of maintenance costs, lagging in scientific and technical development, the growth of the threat to information security and, ultimately, the risk of complete loss of a promising high-tech market of produc-

tion, sale and maintenance of complex scientific and technical products and information and computing services.

At the same time, in Russia, in the last 20 years, the development and production of domestic software and hardware systems and software aimed at solving the problems of prospecting, exploration, and development of deposits have significantly decreased. The lag in the development of scientific research, the creation of software products, the quality of training of specialists from the level achieved by foreign countries was manifested. In some areas, there is an almost complete replacement of domestic equipment and technologies with imported products. In the domestic market, more than 80% of high-level computer technology for solving geological and geophysical problems is imported. When using external information and computing technologies in the field of Geophysics and Geology, there are inevitably prerequisites for the leakage of valuable information about the national subsoil and strategically essential resources. Such a situation, under certain foreign policy circumstances, can have a very detrimental effect on Russia's energy security. This situation does not mean that the state should limit access to the domestic market of high-performance computing technologies in this area from abroad using organizational or economic levers. Instead, it is necessary to stimulate and initiate the creation and implementation of high-quality domestic software products that can effectively compete with similar foreign developments.

There are quite good economic reasons for this. For example, standard software packages of leading foreign companies (Petrel, Eclipse, Roxar) used for seismic and geological-hydrodynamic modeling cost about 4.5 million rubles for one workplace plus 1 million rubles annually is paid for support. At the same time, in many software packages, there is a limit on the use of the number of nodes of the cluster computing system. For the possibility of using each next computational node, you have to pay about 12 thousand dollars. Thus, for an oil and gas company of average size, which involves about 100 geological jobs of about 200 potential users in geophysical works, and half less for hydrodynamic modeling, if necessary, the full use of the computational cluster of 400 nodes, the initial cost of costs is not less than 1350 million rubles plus the cost of annual support (expert evaluation). At the same time, the domestic software package *tNavigator* for the calculation of hydrodynamic models of production of the company "Rock Flow Dynamics" is supplied without limitation on the number of involved computing nodes, are several times cheaper and considers several times faster.

The tasks of development of scientific research, creation, and implementation of the most useful information technologies and the obstacles that stand in this way were noted in the Energy strategy of Russia until 2030. Awareness of the importance of the development of supercomputer technologies and algorithms and software for high-performance computing for modernization and innovative development of various sec-

tors of the economy is recognized in Russia at the state level. This position reflected in the decisions of the Commission under the President of the Russian Federation on modernization and technological development, decisions of the Security Council, reflected In the strategy of development of the geological industry until 2030 and the Energy strategy of Russia until 2030, the State program "Information society for 2011 – 2020", the Project "Creation of a system of training of highly qualified personnel in the field of supercomputer technologies and specialized software" Of the Commission of the President of the Russian Federation on modernization and technological development of the Russian economy, speeches of specialists, developers of supercomputers, software for the oil and gas industry, representatives of oil and gas and service companies concentrated in the solution of the First conference "Supercomputers in the oil and gas industry".

1.6 Criteria for assessing the scientific effectiveness of the STC

Let us turn to the criteria for assessing the effectiveness of the STC in the oil and gas sector of the Russian Federation. The definition and practical use of such criteria are one of the objectives of this work. First of all, let us consider what types

of STC are represented in Russia for activities in the oil and gas sector. They can be divided into the following groups:

- scientific and technical centers for large Russian oil and gas companies such as Gazprom, Rosneft, LUKOIL, TNK-BP, Gazprom Neft, Surgutneftegas;
- state scientific and technical centers;
- independent Russian scientific and technical centers;
- Russian scientific and technical units of foreign service companies, for example, DCS Department in Schlumberger.

Let's consider the activities of STC in the oil and gas industry by short-term and long-term perspectives. Both of these aspects have both economic and technological components. Let us first consider the short-term aspect. As previously noted, the economic component in assessing the efficiency of STC activities in the short term is determined by the dynamics of the company's profit from the sale of oil, gas and their products. At the same time, from the STC as a commercial organization, its internal structure should be optimized for the effective execution of orders, and the experience baggage and competencies should meet the current needs of the market. PR-technologies also play an essential role here.

Indeed, the analysis of the issues of the journal "Oil and Gas industry", conducted with the participation of the author of this work, shows that in recent years in the articles appeared a large number of "digital" terms (topics) such common memes as "data", "method", "system", "study", "sensor", "stan-

dard", "scheme", in contrast to those present in the earlier issues "pipeline", "pipe", "specialist", "Geology", "field", "technology", "territory", "well", "refinery". Is it true that in the Russian oil and gas industry is growing interest in the use of information technology and the use of intelligent data analysis methods are becoming increasingly popular in the oil sector of the economy?

Undoubtedly, however, the articles in a trade magazine, Pestryaev such buzzwords also reflect a splash. Whether this wave will turn into a long-term technological trend will show the development of the situation in time. Thus, the economic situation determines the current advertising campaign in the activities of the STC, implemented through publications in open editions. Of course, not all publications in industry scientific journals are speculative and market-advertising.

Let us consider the assessment of the STC's activities from the long-term perspective. The economic component here can be characterized by using standard integrated indicators of economic analysis of the enterprise, such as integrated financial indicators for the operating period of activity or specific financial indicators per employee. The results of a study [1] by Deloitte, which evaluated the activities of 33 STC operating in the oil sector of the Russian economy, including all of the above types, are shown in Figure.1.2.

[1] https://www2.deloitte.com/content/dam/Deloitte/ru/Documents/energy-resources/Russian/key-trends-of-market-research-in-oilgas-industry-in-russia.pdf

Figure 1.2 — The structure of supply in the market of research works in 2009.

So, we see that despite a specific actual orientation of the industry to import technologies, the share of STC of foreign oil service companies is low, while the percentage of STC in large Russian vertically integrated oil and gas companies exceeds the total part for STC of all other types. Such deviations do not mean that the industry uses domestic technologies, but only that significant industry players prefer to develop and adapt imported techniques on their own.

Given the current structure of the technology development market in the oil and gas sector of Russia, close attention should be paid to the activities of independent STC, taking into account the time of their operation in the market. So it is possible to reveal the hidden technological trends and ac-

tual production and technological requests in the oil industry of Russia. Of course, if the STC is young enough, for example, operates in the market for no more than 5 years, this does not mean that such a company can not demonstrate high efficiency, but it is still evident that the activities of quite young companies in the market require additional analysis in terms of assessing the effectiveness and identifying promising trends.

Let us turn to the description of the criteria for the effectiveness of the activity of the Scientific and Technical Center in the long term, related to the actual production of the Scientific and Technological Center in the form of new technologies. To do this, we consider an approach based on the analysis of digital artifacts of STC activities, primarily various types of documents in electronic format, reflecting the results of STC activities and available for review from open sources. These types of digital artifacts can include:

- industrial e-libraries documents;
- patents;
- scientific papers in industrial journals;
- open access documents about oil and gas industry.

All of these documents may be subjected to computer analysis, primarily a highly promising method of the topic model, the essence of which is the use of bi-clustering, that is, simultaneous clustering of words and documents on their semantic proximity. In this case, as a rule, Dirichlet's hidden placement is used, which, although convenient for algorithmic

computer calculations when conducting thematic modeling, is not entirely justified from a linguistic point of view.

The results of the topic modeling conducted by the author of this work for the articles in all issues of the journal "Oil and Gas industry" for the period from 2008 to 2016 (which was mentioned in this Chapter earlier) showed that contrary to the initial assumption about the smooth evolution of the topics identified in the framework of thematic modeling from issue to issue, in various volumes of the journal were fixed fundamentally different subjects. Does this mean that the journal focuses on the latest technological developments in each new issue and does not refer to outdated and unused technical approaches? Yes and no. The tactical component in the selection of publications by the Editorial Board to increase the attractiveness of the magazine in a wide range of oil and gas and related fields is visible. This tactic is the essence of specialized publishing. At the same time, the fate of the technologies mentioned once cannot be judged by such single publications. Have they been rejected in practice, or have they been tried and proven to be untenable? Or, perhaps, have you become a part of the oil industry's tools during the analyzed period of 8 years? Such conclusions cannot be drawn from the information analyzed. What is the way out of this situation?

It consists in the analysis of a broader range of documents, from patents to abstracts at industry conferences. These documents can be subjected to cluster analysis using different algorithms. At the same time, if we are talking about using a

classification with a training template (“with a teacher”), then expert descriptions of the most modern technological trends and innovative topics in the oil and gas industry can be chosen as a training template. The same sets of analyzed documents that will not be included in the classes (clusters) defined in this way should not be considered a priori as ‘noise", but should be subjected to additional analysis for the fact that they represent evidence (digital artifacts) of latent innovation and technological trends.

On the basis of this performed clustering (classification) of documents (digital artifacts of innovation and technological development of the oil and gas industry) can be defined multicriteria integral numerical performance indicators of specific STC, calculated on the basis of the distribution shares of digital artifacts produced by employees of the STC, clusters (classes). Changes in such distributions over time can serve as a basis for posterior predictive modeling of the performance of specific STC in future periods of time.

A separate topic, although related to the technical side of this study, is the provision of information exchange, access to documents (digital artifacts) and their pre-processing, including uniform electronic formats and stemming.

All these issues will be discussed in the following chapters of this work.

Chapter 2. Related works

2.1 Organizational efficiency

Scientific management guru Michael Porter in his book "Competitive advantage of nations: creating and sustaining superior performance" [1] highlights the effectiveness of research work as one of the processes of competition between countries.

There are many approaches to the interpretation of the concept of performance in general and in research, in particular. But it is essential to understand that efficiency is not a number [2, 3, 4].

The main components of the research process include the formation of researchers collaboration, the process of creating a scientific article and its publication. The published scientific article is one of the embodiments of the results of scientific research. There are many methods of conducting research. Most of them use the structuring of research activities into stages to simplify its understanding. For example, in the book [5] the following seven steps are highlighted:

1. Selection of the research topic;
2. Study of world experience on the chosen topic through scientific sources;
3. Preparation of research work plan;

4. accumulation of material to test the validity of the proposed hypothesis;
5. Data processing, model building;
6. Analysis of research results and conclusions;
7. Documentation of research work.

Thus, the creation of a scientific article, as a result of scientific research, can be presented in the form of a formalized process implemented by the participants of the research group. This process belongs to the category of everyday social interaction. And its study is one of our tasks in this research. Therefore, the author set the task of considering the process of joint research activities and writing a scientific article with subsequent publication. Also, the author of this study tried to take into account the processes of collective thinking and communication, noted in [6].

In scientific practice, researchers should share the results of their research with colleagues. Publication of an article in a scientific journal is a form of communication between a researcher and the scientific community [7]. In addition to the publication of the article, communication can be carried out in the form of publication of monographs, abstracts of conferences or patents, as well as personal presentations at conferences and seminars. Therefore, scientific research cannot be considered in the in isolation from the publication process. Thus, the Editorial Board of scientific journals and committees of scientific conferences should be included in the broad collaboration of the research team.

In the simplified view of the Editorial Board and the committees of the conferences are grouped not by formal categories like Code of State Categories Scientific and Technical Information but in specific mental codes [8], hidden behind the descriptions of the format and editorial policies. An example of such mental code would be "we only accept articles from members of the SPE (Society of Petroleum Engineers)" or "Authors must have a degree in CS". The concept of mental code is widely used in the analysis of grouping [9, 10]. The mental code can be made up of individual fragments, like a DNA molecule. It is essential to understand that a new member is accepted into the community by the coincidence of the mental codes. That, in our case, means acceptance by the Editorial Board or Committee of the conference of scientific work for publication. Sometimes a part of the mental code can be declared. But this does not mean that a significant portion of it, by which the decision will be made, does not remain the internal property of the Editorial Board or the Program Committee. In this case, the author will be puzzled by the fact that he was "unmotivated" refused, as a significant part of the mental code of the Editorial Board or the Program Committee of the conference is not available to him.

The process of publication of a scientific article also has formal stages, which, however, does not reflect the network process of work on the result:

1. Announcement of the date and topic of the conference;
2. Call for papers;
3. Peer review;

4. Creating of bilingual text of the paper;
5. Creating of a presentation;
6. Oral performance of the presentation;
7. Preparing text for publication;

Thus, it is possible to speculate about the fundamental process contains the logic of the extension of the small co-author's group to broader groups. Broader co-authors groups include the representatives of the Editorial Boards, committees of conferences, guest authors, translators, experts, presentation designers. The consideration of such collaborations is necessary to understand the process of publishing scientific articles and then estimate the contributions of individual participants.

The division of labor [11] characterizes the maturity of production processes. For the scientific writing process, it means mean that specialized pools of resources are created to maintain certain stages without personification. For example, from the Soviet Union past, we know the slang term "cooperative for recording formulas" concerning Ph.D. theses. Despite the marginality of this phenomenon, which was publicly condemned and flourished due to the demand in a narrow specialization, the author sees in it the first prerequisites for the division of labor in the production of scientific research and publication on its basis. Currently, due to the acceleration of research production, new forms of division of labor (and new requirements for the effectiveness of research personnel) have appeared, which need to be studied.

The question of collective knowledge creation and writing research, in particular, has many aspects related to the ethics of the researcher. Should the author fully comply with all stages of work on the study? If there are two co-authors in work, then what division of labor does not violate the ethical norms of the researcher? What roles are ethical among co-authors? In the well-known "Course of Theoretical Physics" by Landau and Lifshitz [12], what role did L.D.Landau play, and which E.M.Lifshitz?

After the unification by the mental code, the development of relations within the framework of collaborations takes place in a full (with external participants) and narrow (within the research group) sense. Strengthening co-authorships as a result of writing several papers creates more sustainable working groups. There are examples of ongoing co-authorship over decades. On the other hand, there are examples when, having written a single research paper, the authors no longer collaborate. What are the reasons for sustainable associations in co-authors?

The author believes that many scientific and methodological sources focus on the technology of writing a scientific article and its design, but not studying the process of creating scientific articles. Therefore they consider this work practically useful for administering and planning research regarding scientific management according to F. Taylor [11].

The problem of an objective assessment of the effectiveness of R&D has been in the center of attention of researchers for a long time, and this, first of all, is related to the issues of

financing both budget and within grants. In the framework of the traditional approach, the following indicators for evaluating the effectiveness are highlighted [13]:

- Financial
- Human resources
- Innovational
- Bibliometrical

Actually, within the bibliometrics, the following parameters are taken into account:

- The number of publications in international journals characterizes the quality of articles;
- Citation indicator and Hirsch index show the degree of significance of the research and recognition of scientific schools by the world community;
- "publication load" of scientists shows the productivity of scientists;
- availability of patents;
- co-authorship with foreign scientists is an indicator of international cooperation.

As many researchers have noted [14, 15, 16], this set of parameters is far from perfect, because it does not provide a completely objective picture of the research of the selected scientist or team. For example, the Hirsch index depends on the discipline, and it also does not fall if a person has not published new works for ten years or more. The citation bases WoS and Scopus, firstly, reflect inadequately research in Russian, and secondly, unequal shares are assigned to different disciplines. This

study tests the hypothesis that improving the quality of the evaluation of the effectiveness of research and development is possible through the consideration of additional factors, which will be discussed later.

Organization effectiveness is a very complex and multi-faceted concept. It is influenced by various factors. One of the most important precursors of the market success of a research and development company is a well-developed communication and cooperation between employees.

Many theoretical and practical studies demonstrate the relationship between the productivity of an organization and the communication structure of its employees, for example, see [17, 18]. The study of the social structure of organizations and professional communities is becoming one of the main areas of applied analysis of social networks. In the field of public relations and management, communication patterns within organizations are studied in depth. This research began in 1956 with the work of C.H.Cooley "Social Organization" [19].

Information about employee interactions can be obtained in various ways, for example, through corporate databases, public surveys, and personal reports. However, the data obtained in such ways must be interpreted with some reservations, since they do not reflect the whole mechanism of professional interaction in integrity. According to Wasserman and Faust [20], about half of what people report about their interactions is wrong for one reason or another. Thus, people are not very

good at communicating well with their relationships, so ways to collect data should avoid such subjectivity.

The source of such information may be Google Scholar, arXiv and other online libraries. Consideration of open scientific communities is as impressive as the narrowing of the sample to one country, industry, and organization.

One of the more objective ways of analyzing human interactions is a formal conceptual analysis (FCA). Formal concept analysis (FCA) is a way to analyze a collection of objects and their properties.

A *formal context* is a triple $K = (G, M, I)$, where G is a set of objects, M is a set of attributes, and $I \subset G \times M$ is a binary relation that expresses which objects have which attributes.

In FCA implication $A \to B$ for subsets A, B of the set of attributes M (A,$B \subseteq M$) holds if $A' \subseteq B'$, i.e. every object possessing each attribute from A also has each attribute from B.

An *association rule* is an implication expression of the form $X \to Y$, where X and Y are disjoint sets, i. e. $X \cap Y = \varnothing$. The strength of an association rule can be measured in terms of its *support* and *confidence*. Support determines how often a rule is applicable to a given data set, while confidence determines how frequently items in Y appear in transactions that contain X. See [21] for a detailed introduction to the subject.

In this paper we utilize the FCA framework for studying the author - keyword relationship. For us

- G denotes the set of keywords.
- M stands for the set of all co-authors of the papers.

– $I \subset G \times M$ is a binary relation. One has $(g,m) \in I$ if m co-authors a paper for which g is among the keywords.

Then the association rules are interpreted as indicators of connectivity between different research fields, and also used to recognize weak ties between authors of different papers.

The idea to apply FCA in the context of social network analysis is not new. In [22] it was used for collective network analysis. In [23] a combination of Formal Concept Analysis and well-known matrix factorization methods were used to address computational complexity of social networks analysis and the clarity of their visualization. Bi-clustering and tri-clustering were used in [24] to analyze data collected from the Russian online social network Vkontakte for extracting groups of users with similar interests, finding communities of users which belong to similar groups, and revealing users' interests. FCA was extensively used for analyzing social networks based on co-references, see [25], and detecting criminal networks [26]. For other applications of FCA in social network analysis see [27]. Another rather detailed overview of FCA-based applications for Social Networks Analysis could be found in [30, 28, 29].

One of the particular cases of communication is cooperation, which can in the case of research work go in collaboration with the creation of scientific publications.

The publication of scientific research is the primary object by which the effectiveness of research work is assessed. Therefore, it is essential to follow the process of this process, starting with the birth of the research idea, conducting the experiment

and ending with the publication of the work. It is necessary to analyze what conditions contribute to the successful publication of the article. As part of this study, the ratio of publications of individual scientists and research teams was studied. It has been shown that over the past decade there is a clearly expressed tendency of scientists to unite in groups of co-authors to publish articles. From this, we can conclude that one of the factors that positively affect the publication of works is the unification of people into teams.

In turn, team building is also victorious and unsuccessful, it is also amenable to study, as a result of which it is possible to single out the conditions for successful team building. The task of finding the optimal parameters of the team of co-authors for the most productive writer of scientific articles belongs to the class of optimization problems. Traditionally, researchers pay attention to the following parameters that are important for productive scientific creativity:

- Team size
- Community mental codes
- Emploee competences
- Week connection

Unlike the size of the team, which is a visible, not a hidden sign, and also easily formalized, the sign of the mental models of the community is much more difficult to identify and fix. Many researchers have noted the importance of changes over time in mental models in addition to the structure of the team [32, 31].

The concept of a mental model is the development of the concepts of [33], knowledge structure [34], knowledge schemes [35], and the implicit theory [36].

The author of this study interprets the concept of the mental model as a strategic consistency of team competencies. For example, the mental model of "Agile geoscience" [37] of the largest community of geophysicists is based on the competencies of "flexible techniques" and "geology".

Researchers agree that the coincidence of the mental models of team members has a positive effect on the performance of [39, 38]. This fact addresses the connection between the mental model of the team and the full team code, which is described in more detail below.

The formation of the primary system of internal interaction within the team according to the study [40] occurs when the participants meet the principle of complementarity. Nevertheless, one cannot completely deny the value of homophilic competencies. In many works, the dynamic structure of hemophilia is noted [43, 42, 41], during which two processes take place in parallel. On the one hand, individuals similar to each other form social ties (social selection). On the other hand, people already connected adopt the behavior of each other (social influence). The combination of these factors results in a homogeneous social system, in which between individuals with similar behavior and characteristics there is a connection, while the nature of the relationship can be both formal and informal.

Although connections between individuals with similar characteristics are more likely than those between unique ones, the level of similarity is also essential. In the [44], it was shown that cultural similarity in more than one indicator leads to the fact that people are less likely to form relationships with each other. The author explains this effect by the fact that people who are too similar in many respects, as a rule, cannot bring something new and constructive into mutual relations or a team. Productive cooperation requires not only the similarity of interests is necessary, but also a variety of professional and life experience that allows us to offer multi-dimensional approaches to its solution.

2.2 Scientific text

Text analysis is sometimes called *Text mining*. The essence of this process is the transformation of data (text) into high-quality information capable of bringing knowledge. The critical point is that in obtaining knowledge of human costs should be minimal.

The knowledge obtained from the text becomes the basis for making management decisions in the organizational environment. A separate process is considered the receipt of text, sometimes called the creation of the body of texts.

The real world is reflected in the texts with the help of the authors, and the process of analyzing the text does the opposite: from the texts, it compiles information about the real nature of things.

The multimode approach to text analysis is the process of taking into account the information accompanying the main text. For example, the address of a letter, the number of a newspaper issue with news, or the names of the co-authors of a scientific article.

Formally, text analysis is performed in the following sequence:

1. text language analysis;
2. text content analysis;
3. getting information about the author of the text;
4. deduction of certain variables characterizing the nature of things in the text.

Let us consider in more detail the methods of working with texts of scientific articles.

2.2.1 Text preprocessing

Text processing tasks were arranged in the 60-70 years of the 20th century in the processing of natural language [45, 46]. It was necessary to bring the text to a more convenient form for further analysis. This procedure is commonly called

text normalization. For the normalization of text using regular expressions, the concept of which was developed by S. C. Wedge [47]. One of the first people who used regular expressions in the test was K. Thompson [48].

At present, the task of normalizing the text has expanded considerably. It is necessary not only to highlight words but also to take into account special symbols denoting emotions (Emoji), such as 8-)[49], highlight hashtags [50], highlight hyperlinks [51] and process citations [52].

The task of lexical analysis is to divide the text into parts: sentences, words, letters. Sometimes lexical analysis is called tokenization from the English word *tokenizing* [53].

Another task of text normalization is to define words with a single basis and is called lemmatization. The base of the word does not necessarily coincide with the morphological root of the word. Lemmatization for Russian language differs from lemmatization for English language [54, 55, 56]. Therefore, for English use lemmatization procedure based on frequency algorithms [57, 58], also called stemming from the English word *stemming*. But for other languages, lemmatization uses even more complex algorithms. For example, there is a stemming for the ancient Greek language [59].

Therefore, text normalization consists of three steps:

1. Select words from the text
2. Reduction of words to more common forms
3. allocation of the sentences

Libraries in the Python programming language are used to automate text normalization tasks. For example, the library NLTK [60], containing a vast number of different text processing algorithms.

2.2.2 Text models

Models that assign probabilities to words in word sequences are called probabilistic text models. Mathematically, this definition can be written as an equation. Suppose we have a probability of a sequence of n words $P(w_1, \ldots, w_n)$, such that the probability of the third word $P(w_3)$ is $P(w_3|w_1, w_2)$. Then the following expression (2.1) defines the probabilistic model of the text.

$$P(w) = P(w_1, w_2, \ldots, w_n) = \prod_i^n P(w_i|w_1, w_2, \ldots, w_{i-1}) \tag{2.1}$$

Since the computation of $P(w)$ represents the complexity of O^n, modern text studies use the representation of $P(W)$ as a homogeneous Markov Chain and construct approximate models [61]:

1. unigram model: $P(w_1, w_2, \ldots, w_n) \approx \prod_i P(w_i)$
2. bigram model: $P(w_i|w_1, w_2, \ldots, w_{i-1}) \approx \prod_i P(w_i|w_{i-1})$

One can also consider n-gram models for greater context coverage, as in the works of [62, 63]. In relation to speech

recognition problems, this is done in the works of researchers from IBM [64, 65, 66].

Tomas Mikolov in the works [68, 67] shows that these simplified models have too big computational complexity, so his lab has developed a vector representation of words not using prior distributions as in the study of [69], but by the embedded vectors.

From the equation (2.1) follows a method for checking the quality of the probabilistic model of the text. Most common metric is *Perplexity*:

$$\mathcal{P} = \sqrt[n]{\frac{1}{P(w)}} \tag{2.2}$$

In the studies [71, 70] is shown that the following relationship exists between the metric *Perplexity* and the relative entropy per word $H(W)$:

$$H(W) = \log_2(\mathcal{P}) \tag{2.3}$$

Claude Shannon [72] estimated the entropy of the English language as 0.6 - 1.3 bits per letter, asking people to predict the next letter of the word. In the study [73] the estimation of lower bounds for the entropy of English text as 1.25, and in the research [74] using a trigram model of words given a rating of 1.75 bits per word. But there are other approaches to assessing the quality of text models. For example, in the researches [75, 76] used a metric based not on entropy, but on a pairwise comparison (pair ranking approach). In addition

to the approach developed by T.Mikolov, there are other ways of the vector representation of words. It should be noted the work of researchers and the University of Standford, called the GloVe [77]. Vector representation of the words Glove requires significantly fewer calculations, as it uses only the frequency of the use of words, not the probabilities.

2.2.3 Text classification

The most common example of the need for text classification is probably the problem of classifying emails as spam. And the baseline for this task is the Naive Bayes (NB) based classifier. Naive Bayesian text classification was proposed by M.Maron in [78] to assign the category of ownership of the text of the report to a particular journal. His model presented most of the features used and currently used for the classification of texts.

Bayesian methods [79] were also applied to the task of text classification according to authorship in the pioneering research of F. Mosteller and D. Wallace [80]. Firstly Naive Bayes classifier was applied for detection of spam in the study [81].

In the studies [82, 83, 84] was shown that the use of binary features with multinomial distribution gives better results than word counters.

Binary Bayes with Multinomial distribution is often confused with another option naive Bayesian algorithms that also use a binary representation of whether a word occurs in a document: Multivariate Naive Bayes (MNB) using the Bernoulli distribution. The NB variant with Bernoulli distribution estimates the probability that the word is not included in the document.

The study [85] it is shown that the NMB is not always generalize well to new text. The problem of determining the emotionality of the text refers to the issues of classification and is successfully solved with the help of algorithms of NB. There are many good reviews of the application of emotionality analysis of texts among which are the works of [86, 87, 88]. It is also a good overview of the various text classifiers was made by C.Manning with co-authors of [89].

Currently, the vector representation of the parts of the text (embedding) has become very popular. Word2Vec methods are widely used [68], GloVe [77], StarSparse [90], Fasttext [91], Sent2vec [92]. Therefore, it is worth mentioning and a rational view of the benefits from the use of vector representations

The main demonstration of the advantages of using vector representations of words became the formula: $king - men = queen$. The meaning of this formula is that vector representations of words (king, men, queen) can be subjected to arithmetic operations.

But not all words sum up. Some obvious human analogies in vector representation are not close vectors of [93]. The

search for smears in vector representations is undertaken in [95, 94]. The AdaGram algorithm is proposed in [96] to find vector representations for ambiguous words. A study to clarify the understanding of vector representations depending on the context is undertaken in [97, 98]. Significant advantages in the classification of texts were obtained from the use of recurrent neural networks. From the whole mass of works in this direction it is necessary to note numerous studies of text models based on neural networks performed by the staff of the natural language laboratory at Stanford University [99, 100, 101, 102, 103].

2.3 Social Network Analysis

In the book [104] it is noted that the basis for the analysis of social networks is a theory of sociometry founded by J.L.Moreno [105]. Sociometry studies the relative positions of social atoms in groups. A Moreno sociogram is a graphical representation of the social choice of members of a social group. Social choice can be the choice of a leader, friendship, casual tasks, etc. The sociogram is a graph consisting of the vertices and the edges.

In the book [106] M. Tsvetovat raises the question: Which of the participants of the organization represented in the form of a graph is more important? Thus making a logical connection between graphs and organizational theory, as noted in [107].

An excellent example of the dimension of the innovativeness of the organization by analyzing social ties serves research [108]. Direction in-depth study of graphs (Social media mining) communities of developed in [110, 109]. Thus, translating various metrics of graphs into properties for classification problems of constituent graphs of vertices and edges. For example, in the study [111] metrics of the co-authorship graph is used to predict the new collaborations. And in [112] the prediction of co-authorship is used to improve the efficiency of the scientific organization.

It is necessary to consider how graphs are created to solve the problems of predicting vertices and edges of graphs. Small World [113] and Preferential Attachment Model [114] are models of random graphs consisting of subgraphs.

In the research[115] is shown that the allocation of sub--graphs makes it possible to identify social sub-groups united by a common theme. The definition of such communities is not possible without the use of the mathematical apparatus of graph theory [116].

One of the techniques to identify communities is to build a vector node space (node embedding) [117]. As with the vector space of text parts described in 2.2.3, the construction of the vector space of nodes allows the authors of the study [118] to introduce new properties of graphs based on the proximity of nodes in the vector space.

The co-authorship graph is a particular case of a social network. One of the first studies of the co-authorship graph

is the work of [119], done in 1973. Since that time, research activities with the help of graphs of co-authorship did not stop and gained the status of a proven analysis tool. For example, in a recent study [120] take an attempt to predict future scientific studies based on a count of co-authorship. And in the research [121] was constructed a global graph of co-authorship graph from Google Scholar, which contains over 400 thousand vertices. Both studies were conducted in 2017.

The construction of a co-authorship graph is performed in such a way that if two authors have done a joint research work, then each of the authors is considered the vertex of the graph, and the fact of co-authorship is considered the edge of the graph. We will call this method of creating a graph of co-authorship traditional. The traditionally obtained graph shown in Figure 2.1.

Figure 2.1 — An example of co-authorship graph.

Chapter 3. Object and methods

The emergence of digital ecosystems is the result of the natural development of scientific cooperation and information technologies. The purpose of digital ecosystems is to increase the efficiency of communication between internal and external agents to support business. There are two broad definitions of the concept of digital ecosystems in the literature. The first comes from a structural and functional perspective that sees the digital ecosystem as an open network environment for effective interaction. The second, on the contrary, view the digital ecosystem as an open cluster of loosely coupled components, in which each agent is proactive for its benefit (Figure 3.1).

The concept of "digital artifacts" came into use together with the idea of the digital ecosystem. In a broad sense, digital artifacts are synonymous with any information output from the digital ecosystem. By their informational nature, digital artifacts can be preserved or destroyed. Both the preservation and the destruction of the original digital artifact are modified. In the historical perspective, digital objects can be studied, as well as any other products of human activity.

Digital ecosystems can be considered at the macro level (country, industry) and the micro level (Corporation, a group of companies, individual enterprise, Department). Digital artifacts can also exist at different levels.

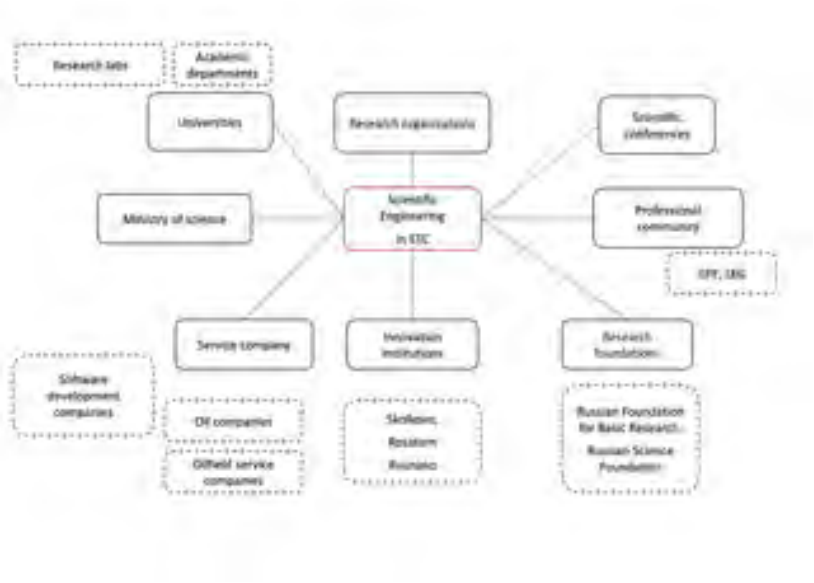


Figure 3.1 — The ecosystem science engineering.

One example of digital artifacts at the micro level is knowledge distribution systems (KDS) in oil and gas companies. A knowledge distribution system is a framework that helps coordinate the processes of management and exchange of knowledge in the field of oil exploration and production within the Gazprom Neft group to solve technological and production problems in decision making.

KDS is designed to set up the processes of collecting, processing and propagating knowledge to maximize the benefits of the company's practices and technologies. KDS is implemented in the form of an information system with several subsystems

that help to obtain the necessary information on various aspects of work at the field.

KDS systematically provides information on the best practices applied in Gazprom Neft in the field of exploration and production. The system allows the user to carry out a comparative analysis and selection of optimal technical solutions by the necessary criteria. It also stores data on all tests of new equipment conducted within the company, which allows the most effective implementation of new equipment and technologies at any field within the company.

The most significant contribution to the development of KDS is made by experts of the Scientific and Technical Center. They form a big structured knowledge database in various fields of Geology, exploration, and production, to which all Gazprom Neft employees have access. KDS is one of the tools to create an innovative climate within the company, necessary for the development of new, more efficient technologies for oil exploration and production

3.1 Modeling of socio-technical objects

The modern paradigm of scientific research is that real objects are replaced by their simplified representations, abstractions, selected so that they reflect the essence of the phenomenon, those properties source objects that are essential

to solving the problem, which was staged. An object that is built as a result of simplification, called a model.

Models can be classified according to different characteristics: dynamic and static, discrete and continuous, stochastic and deterministic, simulation and analytical.

Statistical models operate on characteristics and objects that do not change over time. In dynamic models, the change in model parameters over time is significant. Statistical models are dealing with the equations of balance type, steady-state processes, with marginal characteristics. Simulation of dynamic systems is to simulate the rules of transition from a particular state to another over time.

Models whose state changes continuously over time are called continuous models. Models in which transitions from one state of the system to another occur instantly at separated moments of time are called discrete.

Stochastic models, unlike deterministic ones, take into account the probabilistic nature of the system parameters.

In analytical modeling, the processes of the functioning of the system under study are reflected as algebraic, integral, differential equations and logical relations, and in some instances, the analysis of such associations can be performed using analytical transformations.

In simulation modeling, the structure of the simulated system – its connections and subsystems – is directly represented by the structure of the model, and the process of functioning

of subsystems in the form of equations and rules that bind the variables are simulated on the computer.

Computer systems for predictive modeling (also called engineering decision support systems) with computer-aided design systems have long been used to automate the work of the design engineer and improve the quality of decisions that are made. But until the beginning of the XXI century in predictive modeling used exclusively mathematical models based on the principles of physics, describing the physical phenomena and processes that occur in the operation of the object, complex partial differential equations with boundary conditions. In meaningful situations for such equations neither theorems on uniqueness and existence of the solution, nor the nature of the dependence of the solution on boundary conditions and parameters are unknown. Numerical methods for solving these equations have significant computational complexity and the calculations themselves, and the preparation of initial data and computational grids. Because of this, the possibility of using these models in the design of complex objects is significantly reduced, especially at the stage of conceptual or preliminary design, when a significant number of different solutions are considered, and the price of a solution that is chosen incorrectly is extraordinarily high. An essential part of predictive modeling is simulation modeling, which is used to study complex information and telecommunication systems.

3.1.1 A posteriori and a priori approach to research

Considering the possibilities of a posteriori and a priori approach to research, the author tends to give priority to the experimental study of this phenomenon, and then find out which of the theories can form the basis for further deepening in the study of the phenomenon of co-authorship.

Modern possibilities of direct simulation modeling have become so convenient for computational experiments that for the initial approach to the study of complex social phenomena can quickly give the researcher a significant understanding of their nature. The formalism of the mathematical model, in this case, does not abstract into the world of Greek letters but brings closer to understanding the generic features of the object under study.

The model seeks to describe the system for which it is created. But note that the creation of a model for a complete description of the social system is not a correct statement of the problem. The full model of the social system will be as complex as the social system itself. Let us formulate the following definition of the social system model (3.1.1):

Definition 3.1.1. The $\mathbb{M}_\Omega$ model of the social system Ω can be used to determine the characteristics Z with given accuracy δ.

Thus, the model aims to obtain answers to a set of questions. These questions are implicitly present in the analysis process, and therefore they guide and guide the creation of the model. This indicates that the model will have to answer these questions with a given degree of accuracy. If the model does not answer all questions or its answers are not accurate enough, it is said that the model has not achieved its goal.

Agent-based modeling involves simulating the performance of the system by configuring the behavior of individual agents. Based on the results of the behavior of individual individuals, a complex picture of interactions is formed. The agent modeling method is used in addition to the system dynamics method, which simulates the behavior of the whole system.

Agent modeling software algorithms are developed in several information systems, in particular, Anylogic and NetLogo. These information systems are used to solve practical problems, in Social Sciences, including Economics and Sociology. An essential task of agent modeling is to include information about the interactions of agents with each other since in some social systems it is the complex structure of the communications of individual agents that lead to more complex macro-states. Agent-based modeling is used to study the dynamics of social networks and the mutual influence of exogenous and structural characteristics on each other.

Model for the study of the interaction of agents in the process of creating scientific articles has been implemented by the author in the software environment of agent-based modeling

AnyLogic is based on Java language. In AnyLogic environment, certain rules of behavior are prescribed for each agent – heuristics, individual strategies. After all the rules of behavior are prescribed for each of the agents, a series of simulations is started. Agent modeling software environments are used to predict collective behavior, mass events, educational process, and many other social processes.

The method of simulation based on internal states and actions was used to model the processes in this study. The main advantage of this approach is the ability to conduct a computer experiment to understand the behavior of the system as a whole by adjusting the state graphs and actions of decentralized individual agents. Thus, the result was a database of agent behavior for the study of processes.

Within the framework of the above methodology, the following research questions were formulated:

1. To what extent does the scientific paper reflect the conducted research? Is it possible to judge the quality of research on published research?
2. What are the social mechanisms for bringing together researchers to conduct research? What types of competencies and to what extent do they influence such integration?
3. How does the time of research depend on the number of researchers involved? Are there natural limitations on the number and composition of research teams?

4. What are the heuristic algorithms of handling researchers with the publishers and programme committees of conferences? Are there basic behavioral strategies? If it is possible to identify and simulate basic strategies?
5. Are time management approaches applicable to R&D? How effective is the consideration of research as a project activity?
6. What is the maturity model of the research organization concerning research? To what extent is it possible to determine the degree of maturity of a research organization based on the analysis of scientific articles published by it?
7. What is the structure of the processes that make up the research activity? How appropriate process approach to the study of research activities? There are indicators of research activities, reflecting the characteristic structure of its constituent processes?

3.1.2 Theory of simulation

Simulation modeling is a method of research in which the system under study is replaced by a model that accurately describes the real system with which experiments are conducted to obtain information about this system.

The purpose of the simulation is to obtain approximate knowledge about a specific parameter of the object, without direct measurement of its values. This is necessary when the direct measurement is not possible or is more expensive than the simulation. At the same time, to study this parameter, it is possible to use other known settings of the object and the model of its design. Assuming that the design model accurately describes the object, the authors suggest that the statistical distribution of values of the parameter modeling of the object, obtained during simulation, will to some extent coincide with the statistical distribution of values of the parameter of the real object.

Areas of application of simulation modeling are:

- Agent-based modelling
- System dynamics
- Discrete event simulation
- Dynamic systems

Next, we consider in more detail System Dynamics.

3.1.3 System Dynamics

This approach was developed and proposed by Jay Forrester in the late 1950s as a study of information feedbacks in industrial activity to show how organizational structure, gains (in policies) and delays (in actions and decision-making) interact, influencing the success of the enterprise.

Applications of system dynamics also include urban, social, environmental systems. Processes that occur in reality are represented in System Dynamics regarding drives (stocks, for example, material objects, people, knowledge, money), flows between these drives and information that determines the value of such flows. System Dynamics is abstracted from certain events and objects and assumes an aggregate view of processes. It focuses on the politicians who manage these processes. By modeling in the style of System Dynamics, you represent the structure and behavior of the system as a set of interacting negative and positive feedbacks and delays.

3.1.4 Model building principles

Set of system-dynamic models can describe a socio-economic system. The choice of factors to be included in the model depends on the questions to be answered. But in broader case, the base of the model cannot be limited to any narrow scientific discipline. It is necessary to include in the model economic, organizational, legal, technical, labor, psychological, historical and monetary factors. All of them should find their place in determining the interaction of the elements of the system. Any factor can have a decisive influence on the behavior of the system.

Typically, 30 to 3,000 variables are included in the most important models that meet management requests. The lower limit is close to the minimum, reflecting the main types of system behavior that interest decision makers. The upper limit is limited by our ability to perceive the system and all its interactions.

Particular attention should be paid to such aspects of the system under study as:

- time dependencies,
- backward dependencies,
- distortion of information.

When building a model, its variables must correspond to the variables of the system being modeled and be measured in the same units. For example, the flow of goods should not be measured in monetary units, but in physical units. Cash flows are considered separately. Cash and commodity indices connected with prices. Goods cannot be presented as corresponding monetary amounts, otherwise, the value of prices and the fact that the movement of money is not synchronous to the flow of goods will not be taken into account. Orders for products are not products, shipped products are not equal to invoices, and the invoices are not equivalent to cash.

The economic system model should use actual prices rather than indexed or quoted prices. Actual prices and their fluctuations lead to critical psychological consequences, for example, when determining the number of wages.

The system-dynamic model does not have to be stable. Among the existing socio-economic systems, certain are unstable in mathematical understanding. They do not tend to the equilibrium state even in the absence of external disturbances. Social systems are highly non-linear and most of the time counteract the restrictions that are associated with a lack of labor, overcoming inflation, reduction of monetary resources, the decline in business activity, lack of means of production.

3.1.5 Stages of computer simulation

In addition to the principles, there are typical stages of computer simulation. Typically, it includes the following steps:

- Understanding the system: understanding what is happening in a system that subjects to the analysis: what is its structure, what are the processes leak in it.
- Formulation of the purpose of system modeling: a list of tasks that it is supposed to be solved utilizing the future model. List of weekends and the input parameters of the model, the data source listing, the criteria completeness of future research.
- Development of the conceptual structure of the model: the structure of the model, the composition essential processes to be displayed in the model, fixed level of

abstraction for each subsystem of the model (list of assumptions), description of control logic for subsystems.
- Implementation of the model in the modeling environment: implemented subsystems, their behavior, their parameters, performed the logic of subsystems communication.
- Implementation of the animated representation of the model: model view, user interface.
- Validation of model implementation: the belief that the model correctly reflects the processes of the real system that are required to analyze.
- Calibration of the model: fixation of parameter values, the equations coefficients, and distributions of random variables that reflect situations for which the model will be used.
- Planning and implementing a computer-based experiment: results simulations – tables, graphs, databases, models that correspond to put the question.

In addition to the stages of modeling, it is necessary to consider the principles of data collection required for the experiment. This will be discussed in the next subsection.

3.1.6 Data acquisition

Simulation is a statistical experiment. Its results should be based on appropriate statistical tests: confidence intervals and methods for testing hypotheses. To perform this task, the obtained observations and simulation experiment must meet the following requirements:

1. **Observations have stationary distributions, that is, distributions do not change during the experiment.** The results of observations on the model are dependent on the duration of the simulation period. The initial period of unstable behavior of a model is usually referred to as transitional. When the results of the simulation experiment stabilize, the system goes into a steady state. The longer the run time of the model, the higher the chance of achieving a steady state.
2. **Observations are subject to normal distribution.** This requirement can be fulfilled if we involve the central limit theorem that states that the distribution of the average sample is asymptotically normal, regardless of the distribution of the general the aggregate from which the sample was taken.
3. **Observations are independent.** The nature of a simulation experiment does not guarantee independence between consistent observations of the model.

But the use of samples averages for the presentation of individual observations gives the opportunity to decrease a problem that is associated with a lack of independence.

There are three most common methods for collecting information during simulation modeling:

1. **Subinterval method.** Let us consider a simulation of n observations with a duration T. The information relating to the transient state is cut off according to this method, and the rest of the simulation results is divided into n equal parts. The average value of the desired value within each subinterval is used as the only observation. The advantage of this method is that the influence of non-stationary conditions is reduced. The disadvantage is that successive groups with a common border are correlated, which leads to a failure to fulfill the assumption of independence.
2. **Repeat method.** In this method, each observation is represented as an independent model run, in which the transition period is not taken into account. The calculation of the average sample values for each group is carried out in the same way as in the subinterval method. In this case, the standard formula for dispersion is applicable, since the groups are not correlated with each other. The advantage of this method is that each simulation run of the model is determined by its sequence of random numbers from the interval, due to

which the statistical independence of the obtained observations is provided. The disadvantage is that the initial transient conditions can strongly influence all observations.

3. **Cycle method.** This method can be considered as an extended version of the subinterval method. In this method, we tried to reduce the effect of autocorrelation by selecting groups to provide the same initial conditions for each of them. The length of the queue can be considered as a variable. Then each group should start at the moment when the queue length is zero. In contrast to the subinterval method, in the method of cycles, the length of the intervals of each group may be different. The disadvantages of the method include a smaller, in comparison with the subinterval method, the number of observations obtained at a given run length.

Simulation is a reasonably flexible research tool that can be effectively used in the analysis of complex systems. Its disadvantage is that any result obtained by simulation modeling is subject to experimental errors and must be verified by statistical tests. The task of obtaining observations using simulation modeling, which is both representative and independent in stationary conditions, is somewhat tricky. The use of specialized data collection techniques can mitigate these difficulties.

3.1.7 The use of simulation models in historical research.

Theoretical and methodological problems of application of simulation models have not yet been developed. There are different opinions about the possible use of simulation models in history, but there is a great interest in their application. The existing experience of their practical construction makes it possible to identify three types of tasks that can be solved on their basis:

- modeling alternative, that is subjectively and objectively possible, but unrealized in practice historical situations in order to characterize the real course of development more deeply;
- building models of counterfactual (really non-existent) historical situations that are constructed by the historian in order to use these models as a benchmark for assessing real historical reality;
- imitation of historical phenomena and processes, for the collective characteristics and reflective-measuring modeling of which there is no necessary concrete historical data.

In recent years, significant success has been achieved in the field of creating models of social history. The models currently available can be divided into three groups:

1. model-concepts based on identifying and analyzing general historical patterns, representing them in the form of cognitive schemes describing the logical connections between various factors influencing historical processes (J. Goldstein). These models have a high degree of generalization, but have not a mathematical, but a purely logical, conceptual character;
2. particular mathematical models of the imitation type, which are devoted to the description of specific historical events or phenomena (D. Meadows, J. Forrester). In such models, the focus is on careful consideration of the description of the factors of the processes that influence the phenomena under consideration. The applicability of these models is usually limited to a rather narrow space-time interval; they are " tied " to a specific historical event, they cannot be extrapolated for long periods of time;
3. mathematical models that are intermediate between these two types. These models describe a specific class of social processes without a claim for a detailed description of the features of each specific historical case. Their task is to identify the underlying patterns that characterize the processes of the type in question. By this, these mathematical models are called primary.

3.2 Model of the process of publishing scientific articles

All researchers were faced with the fact that publishing the results of a study is almost as tricky as performing the study itself. Consider the process of publishing research results in detail and analyze the possibilities of its acceleration and simplification for the authors. The starting point for our analysis will be a sharp text describing, from researchers, the result of their research work. Traditionally, this text is called a manuscript.

In the modern world, the speed of publication of manuscripts is a critical factor for the growth of the country's scientific contribution to international science. Publication of articles requires a wide range of administration and communication skills from researchers, which are not always characteristic of scientists. The need for these individual authors to acquire these skills creates the risk of losing focus on research questions and takes time away from scientists, which can be usefully spent on science. On the other hand, by co-sponsoring people, for example, to translate articles into English or lobbying business trips to a conference, the authors blur the research profile of the organization and create so-called "guest-authors".

Historically, the task of a scientist is to make the result of a study accessible to the broadest range of stakeholders; this is the essence of the process of publishing research results. The pri-

mary goal of this thesis is to explore the process of publishing a manuscript, understand bottlenecks, identify opportunities for their elimination, and suggest improvements. Below is the research framework of the study in the form of a diagram (Fig. 3.2):

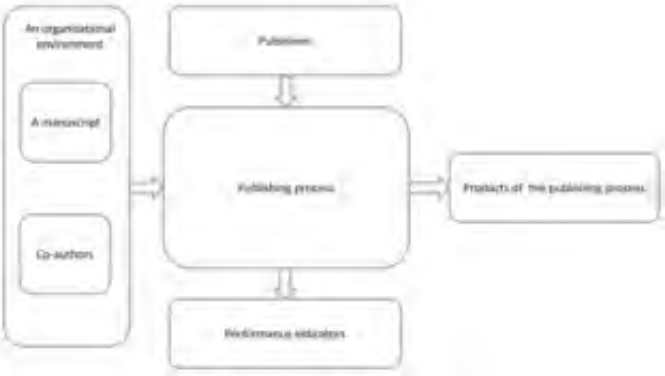


Figure 3.2 — The research framework for publishing processes.

As can be seen from the figure (3.2), the logical framework of the study includes the organizational environment (co-authors and their manuscript), the process of publication, publishers, performance indicators and the results of publication. In the next subsection, each of the components of the logical framework will be discussed in more detail.

3.2.1 Manuscript

As mentioned earlier, a form manuscript is a text. Methodologically, manuscripts are divided into the following main types:

- Monography, dissertation
- Paper
- Report theses

A scientific article is a work of a small volume, usually from five to twenty pages. The content of scientific articles is divided into three types: theoretical articles, practical articles, and methodical articles.

Practical articles are devoted to logical experiments and real experience. Further, it will be considered this type of manuscripts.

3.2.2 Co-authors

The most study is done by research teams, not by single authors. That is why manuscripts are also written as a result of collective work. According to the study [122], in the oil and gas industry, the distribution of the number of co-authors has the form shown in the Figure 3.3.

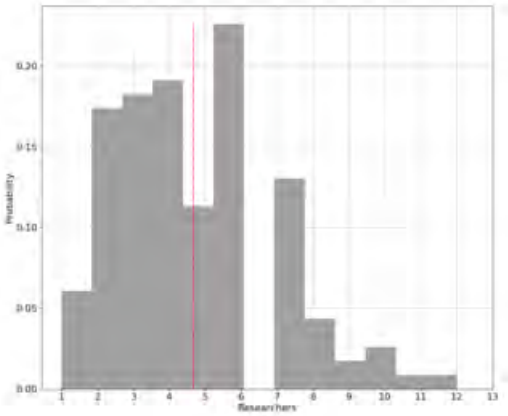


Figure 3.3 — Distribution of the number of co-authors of scientific articles in the oil and gas industry.

The figure 3.3 shows the normal distribution of the number of co-authors. The red line indicates the average value: 4.67. The standard deviation of the distribution is 2.28.

3.2.3 Organizational environment

Research is carried out by employees of research departments. In the oil and gas industry, such units may belong to specialized institutes, scientific and technical centers, service organizations and other participants of the ecosystem. Thus, the co-authors work in an organizational environment. The organizational environment largely determines the communication between co-authors, which is vital in our research.

3.2.4 The publishing process

The publishing process consists of two types of actions:

- Interaction of co-authors with the publisher;
- Interaction between co-authors;

The object of both actions is the manuscript and additional related materials: questionnaires, presentations, letters, reviews. The main task of interaction with the publisher is to meet the conditions for the publication of the article in this edition. Usually, the requirements for authors are indicated on publishers' websites and may differ. Co-authors' interactions during the publication process include the following:

- Creating a list of possible publishers,
- Study of specific topics required by publishers,
- Defining the time limits for the submission of the manuscript,
- Preparation of the manuscript revision plan to meet the requirements of publishers,
- Collection of related documents according to the requirements of publishers,
- Preparation of presentation for the report (required for conferences)
- Make an oral presentation (business trip)
- Confirmation of authorship in the science community and citation indexes.

The most significant are the publishers recommended for publication by the Higher Attestation Commission (HAC). The "list of peer-reviewed scientific publications" of the HAC as of 20.09.2017 contains 2172 magazines. We will choose publishers by one, the closest to the oil and gas industry specialty-25.00 "Earth Sciences". There are 147 such magazines. Then the author of the study proposed the following sequence for data collection (Fig. 3.4).

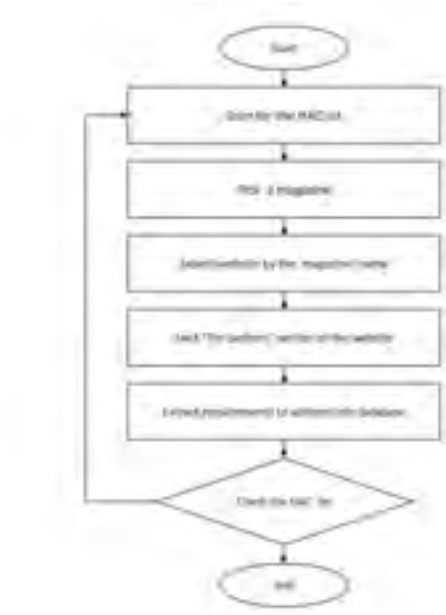


Figure 3.4 — Algorithm for author' requirements collection.

Let's use the results of research [123] to select publishers with the highest publication activity and impact factor on in-

ternational abstract databases. The list contains 16 journals. All 16 journals have standard rules for authors developed by the publisher MAIK "Nauka/Interperiodika". Each edition has its allowable volume of publications, due to the number of articles in one issue and the number of issues per year. The more manuscripts the publisher receives, the higher the competition for the right to be published.

3.2.5 The results of the publication.

The result of the publication is a real contribution to science. The problem of maximizing the availability of research results in the Internet era can be solved by using Internet resources. Here are just some ways to increase the audience:

- The international scientometric databases (Scopus, WoS),
- Digital libraries (for example, eLibrary.ru, OnePetro.org),
- Assignment of the digital object identifier (DOI) to a scientific article,
- Linking a scientific article to an author in online communities of scientists (for example, ResearchGate),
- Publication of the material in open libraries (e.g., aRxiv.org),
- Binding an article to a scientist identification number (e.g., ORCID, SPIN),

– Citation indexes (e.g., Russian Science Citation Index).

Citation index, for example, the Russian Science Citation Index (RSCI), is one of the most common scientometric indicators in Russia and is used for formal evaluation in the scientific community. Alternatives to the citation index are expert evaluation and evaluation of the impact factor of scientific journals. In-depth methods of bibliometric analysis provide an opportunity to consider the contribution of the author from different points of view. Much attention is paid in particular to the analysis of publications using co-authorship graphs, discussed further in 3.7. An example of the co-authorship graph is shown in the figure (Fig. 3.5).

Figure 3.5 — An example of co-author graph for keyword *Oil rims*

Graphs of co-authors allow visually identify the most important scientists on the consideration subject. For example, in the figure 3.6 we can see such a cluster.

Figure 3.6 — A fragment of the graph of co-authorship of the keyword Oil rims.

On the figure 3.6 only nodes belonging to Professor Rahim Masoudi are depicted.

The principle of construction of co-authorship graphs is to classify the number of publications on the selected keyword to the vertices (authors), and the facts of co-authorship to the edges of the graph. This principle of graph construction allows to analyze it using social network analysis (SNA) methods.

3.2.6 Performance indicators

Performance indicators of the publication process need to give an integrated description of the process and allow to compare various implementations process. The table 1 shows the most valuable performance indicators.

Table 1 — Performance indicators of the publishing process.

Performance indicator	Description
Efficiency of publications	The ratio of the number of published manuscripts to the total number of written manuscripts
The share of published manuscripts per author	The ratio of the number of published manuscripts to the number of authors
The proportion rejected by the publishers of manuscripts per author	The ratio of rejected manuscripts to the number of authors in the organization

It is assumed that the process is more productive when *Publication Efficiency* tends to one, *Share of published manuscripts per author* increases, and *Share of rejected papers per author* tends to zero. Strategies for managing the publication process through productivity indicators are given in the table 2.

Table 2 — Performance management strategies for the publication process through productivity indicators.

Performance indicator	**Maximum productivity**	**Minimum productivity**
Publication Efficiency	tends to one	tends to zero
Share of published manuscripts per author	Increases	Decreases
Share of rejected papers per author	tends to zero	Increases

Note that these productivity indicators do not characterize the quality of the scientific article. In this study, the author does not set the task of assessing the quality of scientific work.

By the above-stated methodical principles, the model of the process using system dynamics can be constructed.

3.3 Theory of surrogate modeling

The surrogate model is the basis of a new direction of modeling in engineering. It is a mathematical method of draw-

ing up a model based on the results of tests and computational experiments conducted with various objects of the same class. In some cases, surrogate modeling is the only way to solve the engineering problem.

The task of surrogate modeling is to optimize the original complex function in such a way as to minimize the calculation area and minimize it. Creation of the surrogate model of the objective function subsequently replaces the function itself and simplify many engineering tasks.

The concept of creating surrogate models consists of the following stages:

- Characteristics of the object Z, which determines the properties of an object under certain conditions, can be described in the form of the functional dependence of $Z = \Phi(X, Y)$, where the variable X represents the object, and the variable Y specifies the conditions of operation.
- The function Φ is unknown, and computational experiments are carried out to calculate it.
- There are a number of measurements $\Xi = \{X_i, Y_i, Z_i = \Phi_i(X_i, Y_i), i \in \mathbb{R}\}$, where $Z_i = \Phi_i(X_i, Y_i)$ of the characteristic of Z obtained by M_i for object with description X_i, in the framework of the Y_i.
- Over a well-known Ξ with the help of certain mathematical methods for the analysis and processing of data is a function of $\Phi^s(X, Y)$ whose value is taken as the ap-

proximate value of the characteristics Z of the object with the description of X in terms of operation Y.

– If all values in the set Ξ are obtained using the same model M and $\Phi^s(X,Y) = \Phi^m(X,Y)$, then the constructed function Φ^s can be considered as a "substitute" (surrogate) for the function Φ^m.

Surrogate modeling has been successfully applied in such areas as electrical engineering, oil, water management, military, mechanical engineering, and chemical industry.

The use of surrogate models is also indispensable in construction to optimize aerodynamic contour to identify the optimal shape of unique civil structures, such as high-rise buildings and long-span bridges, which are surrounded by turbulent flow.

The following tasks of the oil and gas industry can also be solved using surrogate models:

– Surrogate reservoir model,
– Optimizing the location of wells,
– Uncertainty analysis of oil production forecast,
– Automatic adaptation of the basin model according to the data.

Some computational experiments use the meta-algorithm described above to solve problems in the oil and gas industry. For example, the hydrodynamic simulator first calculates the function values for specific nodal values of X_i parameters based on the physical laws of fluid motion in a porous medium M_i. And then the function Φ specified in this numerical way is used

to obtain the values of the function Y_i either on a more detailed set of parameter values or for parameter values beyond the nodal values X_i.

One of the main reasons for the meta-algorithm described above is the construction of a surrogate model is the limitations on the speed of hydrodynamic modeling. In the future, when at any time any specialist of the organization will be able to vary the values of parameters in a wide range and in near real time to get the desired values of the function, the need for surrogate models is likely to disappear. In the meantime, modeling is performed on expensive high-performance computational clusters, with the help of specialists for the times measured in hours and sometimes days for one set of parameters, there is a need for foresight data preparation that may be needed in the future. Since the need to change the parameters can occur several times a day and demand a variety of specialists from different departments of the organization, the use of surrogate modeling is an urgent need. The resulting surrogate model Φ^s, sometimes called a proxy model, exceeds the original model Φ^m in computational power many times, that is, does not require a large number of computing resources and operates in near real time.

3.4 Nonparametric models

To understand non-parametric models let us consider a parametric model. A parametric model p for values of y dependent variables X and parameters θ would be $p(y|\theta)$. Finding the θ parameters using the a posteriori probability maximization methods $p(\theta|y,X) \longrightarrow \max_\theta$.

Optimization methods are used to find the optimal parameters of the mathematical model. Numerical optimization methods are:

- Gradient and non-gradient,
- Robust (for optimization problems under uncertainty),
- Surrogate-based.

Let's consider Bayesian optimization methods, which are most often used in surrogate and simulation modeling. In this case, the data and the model are a "black box".

Let the function $f(x)$ be given and we need to find x at which it reaches a maximum of $f(x) \longrightarrow \max_x$. Let's add a condition under which the calculation of each value $f(x)$ is a resource-intensive task. This condition may occur in the following cases:

- x are the geographic coordinates of the well, and $f(x)$ is the amount of oil that can be extracted by drilling the well at x coordinates. In this case, one value of $f(x)$ is worth millions of rubles;

- x are hyperparameters of artificial neural deep learning network, $f(x)$ is a target metric of prediction accuracy. In this case, one value of $f(x)$will take months of work;
- x is the medicine formula, and $f(x)$ is the efficacy of the medicine against the disease. In this case, one $f(x)$ will cost the life of one test animal.

Thus, the formulation of the problem is to optimize the target function in the minimum number of attempts. At the same time, the use of surrogate models of the objective function allows making each optimization step less resource-intensive. Let us introduce the function values of the detection of $\mu(x)$ that characterizes the benefit received from the optimization of $f(x)$ when using a surrogate model $\hat{f}$. Value function detection is a quantitative evaluation function to minimize the number of attempts. Consider the following $\mu(x)$:

- Maximum probability of improvement (MPI): $\mu(x) = P(\hat{(}f(x)) \geqslant f^* + \varepsilon = \Phi\left(\frac{\mathbb{E}f\hat{(}x)-f^*-\varepsilon}{Var[f\hat{(}x)]}\right)$, where f^* - current best value.
- Upper confidence bound (UCB): $\mu(x) = \mathbb{E}f\hat{(}x) + \eta Var[f\hat{(}x)]$
- Expected improvement (EI): $\mu(x) = \mathbb{E}\max(f(x) - f^*,0) = Var[\hat{(}f(x)] \cdot [z\Phi(z) + \varphi(z)]$, where $z = \frac{\mathbb{E}f\hat{(}x)-m(x)}{Var[f\hat{(}x)]}$

3.5 Bayesian Methods of STC' Parameters Estimation

Let us consider the results of the STC activity as observations x. Then in the broadest sense as a problem we set to find the distribution of the random variable θ, leading to the available observations x.

According to Bayes theorem we have equation 3.1.

$$p(\theta|x) = \frac{p(x|\theta)\ p(\theta)}{\sum_i p(x|\theta_i)\ p(\theta_i)} \tag{3.1}$$

To calculate the a posteriori distribution $p(\theta|x)$ based on the likelihood function $p(x|\theta)$, a priori distribution with probability density $p(\theta_i)$ and full probability $p(x) = \sum_i p(x|\theta_i)\, p(\theta_i)$.

Calculating the total probability $p(x)$ is a complex problem, so we use the principle of maximizing the a posteriori probability $p(\theta|x)$. Find the parameters θ_{MAP} for which the expression $p(\theta|x)$ is maximal. The principle of maximizing a posteriori probability (Maximum a Posteriori, MAP) can be written as 3.2:

$$\theta_{MAP} = \arg\max_{\theta} p(\theta|x) \tag{3.2}$$

$$= \arg\max_{\theta} \frac{p(\theta|x)\ p(\theta)}{p(x)} \tag{3.3}$$

Since the total probability $p(x)$ does not depend on θ, we can remove the denominator and obtain a formulation for the optimization problem in the form 3.4.

$$\theta_{MAP} = \arg\max_{\theta} p(\theta|x)\ p(x) \tag{3.4}$$

The equation 3.4 does not contain $p(x)$ and can be solved by numerical methods. But this approach suffers from the following problems:

- There is no invariance with respect to distribution parameters θ_{MAP};
- θ_{MAP} not applicable as a priori distribution;
- There is no possibility to evaluate Bayesian credible interval.

Let us consider the particular case of θ_{MAP} when the probabilities of all θ are uniformly distributed. Then the problem of finding θ is to find the maximum value for $p(\theta|x)$. This approach is called the method of maximum likelihood estimation (MLE). Now we may write the expression of the optimization problem for the maximum likelihood estimation method (3.5).

$$\theta_{MLE} = \arg\max_{\theta} p(x|\theta) = \arg\max_{\theta} \prod_{i} p(x_i|\theta) \tag{3.5}$$

Without losing generality, we can maximize the logarithm from the right side of the expression 3.5 and get next expression 3.6.

$$\theta_{MLE} = \arg\max_{\theta} \log p\left(x|\theta\right) \tag{3.6}$$

$$\begin{aligned} \theta_{MLE} &= \arg\max_{\theta} \log \prod_{i} p\left(x_i|\theta\right) && (3.7)\\ &= \arg\max_{\theta} \sum_{i} \log p\left(x_i|\theta\right) && (3.8) \end{aligned}$$

Let us show in more detail how the MAP is converted to MLE for the case of uniform distribution θ:

$$\begin{aligned} \theta_{MAP} &= \arg\max_{\theta} \sum_{i} \log p\left(x_i|\theta\right) p\left(\theta\right) && (3.9)\\ &= \arg\max_{\theta} \sum_{i} \log p\left(x_i|\theta\right) \; const && (3.10)\\ &= \arg\max_{\theta} \sum_{i} \log p\left(x_i|\theta\right) && (3.11)\\ &= \theta_{MLE} && (3.12) \end{aligned}$$

Another approach to estimating θ is the conjugate a priori distribution method. In Bayes theorem 3.1, only the term $p\left(x\right)$ is mutable, since the likelihood function $p\left(x|\theta\right)$ is defined by the model, and $p\left(x\right)$ by the data.

The distribution of a priori probability is called conjugate to the distribution of a posteriori probability if they belong to the same family of distributions.

Let us explain the above with an example. Let $p\left(x|\theta\right)$ and $p\left(x\right)$ be normal distributions. For the distribution $p\left(\theta\right) = N(x|\mu_0,\sigma_0^2)$ with expectation μ_0 and variance σ_0^2, the expression 3.1 can be written as 3.13.

$$p\left(\theta|x\right) = \frac{\mathbb{N}\left(x|\theta\right)\ \mathbb{N}(\theta|\mu_0,\sigma_0^2)}{p\left(x\right)} \tag{3.13}$$

The product of two normal distributions will also be a Normal distribution, and the following formulas can calculate the parameters of the posterior distribution 3.14.

$$\mu = \frac{\left(\frac{\mu_0}{\sigma_0^2} + \frac{\sum_{i=1}^{n} x_i}{\sigma^2}\right)}{\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}} \tag{3.14}$$

$$\sigma = \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right)^{-1} \tag{3.15}$$

Thus, the use of conjugate families of distributions avoids complex calculations of the total probability.

3.5.1 Latent variables of the model

Speaking about the characteristics of STC as an object of research, we cannot measure such parameters as intellectual capital (IC). Although the IC affects the performance of the STC, which we can measure, for example, the number of publications and the number of authors. We will call such parameters as IC hidden or latent parameters.

A machine learning approach can be considered to define IC. For example, based on an artificial neural network. Then to us for training artificial neural networks will need a dataset

containing the values of IC for different companies with different parameters: number of publications, number of employees, and others. It is known from the theory that datasets with hundreds of thousands of samples and hundreds of parameters are needed to train artificial neural networks. There is no such dataset for STC. However, even have to imagine that such a dataset is there, it will contain a lot of missing values, conflicting data, and other problematic data.

On the other hand, Bayesian statistics can work with small datasets. That fact brings us to the consideration of the probabilistic approach to the assessment of hidden parameters. The first step for constructing a probabilistic model is to build the dependence of the observed parameters (Fig.3.7). Moreover, at first glance, all the parameters will depend on each other. For example, the more authors, the more publications, the more employees with academic degrees, the more publications in journals from the list of HAC.

One solution may be the introduction of hidden parameters such as IC, which reduce the number of links. Suppose that the STC has an IC on which the number of publications and the number of authors depends. Thus, the number of combinations for probabilistic estimation is significantly reduced.

Figure 3.7 — The fragment of Bayesian network for STC.

3.5.2 The Expectation-Maximization algorithm

Consider the probabilistic formulation of Jensen's inequality. Let $(\Omega,\mathcal{F},\mathbb{P})$ be a probability space, and $x\colon \Omega \to \mathbb{R}$ be a random variable defined on it. Let also $\varphi\colon \mathbb{R} \to \mathbb{R}$ be a convex (down) function. If $X, \varphi(X) \in L^1(\Omega,\mathcal{F},\mathbb{P})$, then $\varphi\big(\mathrm{E}[X]\big) \leqslant \mathbb{E}[\varphi(X)]$, where $\mathbb{E}[\cdot]$ means expectation. In other

words, for the convex function f and the probability distribution t, we obtain the following expression 3.16:

$$f\left(\mathbb{E}_{p(t)}\, t\right) \geqslant \mathbb{E}_{p(t)} f\left(t\right) \tag{3.16}$$

For further consideration, let us present the following definition of Kullback-Leibler divergence 3.17.

$$\mathcal{KL}\left(q||p\right) = \int_x q\left(x\right)\, \log \frac{q\left(x\right)}{p\left(x\right)}\, dx \tag{3.17}$$

Note that it is more accurate to call the Kullback–Leibler divergence an asymmetric measure of the difference between two distributions $q\left(x\right)$ and $p\left(x\right)$. Since by definition this measure does not have symmetry 3.18.

$$\mathcal{KL}\left(q||p\right) \neq \mathcal{KL}\left(p||q\right) \tag{3.18}$$

Another useful property of the Kullback–Leibler divergence is shown as inequation 3.19.

$$\mathcal{KL}\left(q||p\right) \geqslant 0 \tag{3.19}$$

To proof it let us make the following computations 3.20.

$$\begin{aligned}
\mathcal{KL}(q||p) &= E_q\left(-\log \frac{q}{p}\right) && (3.20)\\
&= E_q\left(-\log \frac{p}{q}\right) \leqslant \log\left(E_q \frac{q}{p}\right) && (3.21)\\
&= \log \int_x q\left(x\right)\, \frac{q\left(x\right)}{p\left(x\right)}\, dx && (3.22)\\
&= 0 && (3.23)
\end{aligned}$$

Consider use of EM-algorithm for finding the hidden settings STC. Assume that we have an STC model with latent parameters. Let's denote the latent parameters as t_i and the observed parameters as x_i. Then the likelihood function can be expressed as 3.24.

$$p\left(x_i|\theta\right) = \sum_c p\left(x_i|t_i = c\right) p\left(t_i = c|\theta\right) \tag{3.24}$$

Where $p\left(t_i = c|\theta\right)$ is the a priori probability that t takes the value t. The problem is to maximize the probability of the likelihood function by θ. Since the logarithm is a convex continuously increasing function, we will look for the maximum logarithm of $p\left(x_i|\theta\right)$. Suppose also that all N estimations of x_i were made independently. Then the probability $X = \prod_i^N p\left(x_|\theta\right)$.

$$\log p\left(X|\theta\right) = \sum_i^N \log p\left(x_i|\theta\right) = \sum_i^N \log \sum_c p\left(x_i|t_i = c|\theta\right) \tag{3.25}$$

It is worth noting that we can search for the maximum expression 3.25 using gradient methods. For example, using the stochastic gradient descent method. However, the author purposely used a different algorithm and showed it's advantages in the next paragraph.

Apply the Jensen inequality 3.16 to the expression 3.25 and get $\log p\left(x|\theta\right) \geqslant \mathfrak{L}\left(\theta, q\right)$. Next, select the function $\mathfrak{L}\left(\theta, q\right)$ so that it is easy to maximize it (3.26).

$$\mathfrak{L}(\theta, q) = \sum_{i}^{N} \sum_{c} q(t_i = c) \log \frac{p(x_i, t_i = c|\theta)}{q(t_i = c)} \tag{3.26}$$

So as a result, for the parameter θ and the distribution q we obtain an inequality 3.27.

$$\log p(X|\theta) \geqslant \mathfrak{L}(\theta, q) \tag{3.27}$$

Now, to find the maximum of $mathfrakL(\theta, q)$, we apply the following two-step iterative algorithm (3.5.2) for each iteration of k.

– Fix θ^k and choose q^k so that $\mathfrak{L}\left(\theta^k, q^k\right)$ will maximal;

– Get $q^{k+1} = \arg\max_q \mathfrak{L}\left(\theta^k, q\right)$

The first step is called E-step, and the second M-step. Together, they represent an EM-algorithm whose result is θ for the hidden variable t.

3.5.3 The E-step

Let us consider the E-step in more detail. Maximizing the lower bound function $\mathfrak{L}\left(\theta^k, q^k\right)$ means that the distance between $\mathfrak{L}\left(\theta^k, q^k\right)$ and the maximum likelihood function $\log p\left(X|\theta^k\right)$. Lets write this equation for the k-th iteration (3.28) and show that this distance can be expressed concerning Kullback–Leibler divergence.

$$
\begin{aligned}
DIST &= \log p\left(X|\theta\right) - mathfrakL\left(\theta, q\right) && (3.28)\\
&= \sum_{i}^{N} \log p\left(x_i,\theta\right) - \sum_{i}^{N}\sum_{c} q(t_i = c) \log \frac{p\left(x_i,t_i = c|\theta\right)}{q(t_i = c)} && (3.29)\\
&= \sum_{i}^{N} \Big\{ \log p\left(x_i|\theta\right) \sum_{c} q(t_i = c) - \sum_{c} q(t_i = c) \log \frac{p\left(x_i,t_i = c|\theta\right)}{q(t_i = c)} \Big\} && (3.30)\\
&= \sum_{i}^{N}\sum_{c} q(t_i = c) \Big\{ \log p\left(x_i|\theta\right) - \log \frac{p\left(x_i,t_i = c|\theta\right)}{q(t_i = c)} \Big\} && (3.31)\\
&= \sum_{i}^{N}\sum_{c} q(t_i = c) \Big\{ \log p\left(x_i|\theta\right) - \log \frac{p\left(x_i,t_i = c|\theta\right)}{q(t_i = c)} \Big\} && (3.32)\\
&= \sum_{i}^{N}\sum_{c} q(t_i = c) \log \frac{p\left(x_i|\theta\right) q(t_i = c)}{p\left(x_i,t_i = c|\theta\right)} && (3.33)\\
&= \sum_{i}^{N}\sum_{c} q(t_i = c) \log \frac{p\left(x_i|\theta\right) q(t_i = c)}{p\left(t_i|x_i,\theta\right) p\left(x_i|\theta\right)} && (3.34)\\
&= \sum_{i}^{N}\sum_{c} q(t_i = c) \log \frac{q(t_i = c)}{p\left(t_i|x_i,\theta\right)} && (3.35)\\
&= \sum_{i}^{N} \mathcal{KL}(q(t_i)||p\left(t_i|x_i,\theta\right)) && (3.36)\\
& && (3.37)
\end{aligned}
$$

Thus, maximizing the lower bound function $\mathfrak{L}\left(\theta^k, q^k\right)$ is equivalent to minimizing the sum of Kullback–Leibler divergences for $q(t)$ and $p\left(t|x,\theta\right)$. Since the Kullback–Leibler divergences are nonnegative by definition, we can equate them to zero to find the global minimum (3.38).

$$0 = \sum_{i}^{N} \mathcal{KL}\left(q\left(t_i\right) || p\left(t_i | x_i, \theta\right)\right) \tag{3.38}$$

It is also known from the definition of the f Kullback–Leibler divergences that it is zero only if both distributions match (3.39).

$$q\left(t_i\right) = p\left(t_i | x_i, \theta\right) \tag{3.39}$$

The equation 3.39 means that in order to find the optimal distribution $q\left(t\right)$ we must choose it equal to the posteriori distribution $p\left(t|x,\theta\right)$.

3.5.4 The M-step

At the M-step, the likelihood function 3.26 is maximized at a fixed $q\left(t\right)$ by θ.

$$\begin{aligned}
\mathfrak{L}\left(\theta, q\right) &= \sum_{i}^{N} \sum_{c} q\left(t_i = c\right) \log \frac{p\left(x_i, t_i = c | \theta\right)}{q\left(t_i = c\right)} \\
&= \sum_{i}^{N} \sum_{c} q\left(t_i = c\right) \log p\left(x_i, t_i = c | \theta\right) - \sum_{i}^{N} \sum_{c} q\left(t_i = c\right) \log
\end{aligned}$$

Note that since the expression $\sum_{i}^{N} \sum_{c} q(t_i = c) \log q\left(t_i = c\right)$ does not depend on θ, it will be zeroed when

differentiating. Thus, the expression 3.40 can be transformed as follows (3.42).

$$\mathfrak{L}(\theta, q) = \mathbb{E}_q \log p(X,T|\theta) + const \qquad (3.42)$$

Recall that in the expression 3.42 x is all data, and T is all values of latent variables. $\mathbb{E}_q$ denotes the expected distribution of q. Since we choose distributions for x and T, we can ensure that $p(X,T|\theta)$ is smooth and continuous. This choice will significantly simplify the finding of the extremum by θ.

3.5.5 Convergence of the EM-algorithm

The EM-algorithm is designed to find the local extrema of the maximum likelihood function. To do this, we use the lower bound function $\mathfrak{L}(\theta^k, q^k)$, which does not decrease in the optimization process 3.43.

$$\log p\left(X|\theta^{k+1}\right) \geqslant \log p\left(X|\theta^k\right) \qquad (3.43)$$

Use of EM-algorithm to identify hidden topics in the text. The scientific text is one of the indications of NTC activity. The identification of text topics can be made using the Dirichlet distribution. Bayesian model for the posterior distribution of hidden topics in the text can be written in the following form (3.44).

$$p\left(W,Z,\Theta\right) = \prod_{d=1}^{D} p\left(\theta_d\right) \prod_{n=1}^{N_d} p\left(z_{dn}|\theta_d\right) p\left(w_{dn}|z_{dn}\right) \tag{3.44}$$

$$p\left(\theta_d\right) \sim Dir\left(\alpha\right) \tag{3.45}$$

$$p\left(z_{dn}|\theta_d\right) = \theta_{dz_{dn}} \tag{3.46}$$

$$p\left(w_{dn}|z_{dn}\right) = \Phi_{z_{dn}w_{dn}} \tag{3.47}$$

$$\sum_{w} \Phi_{tw} = 1 \tag{3.48}$$

$$\Phi_{tw} \geqslant 0 \tag{3.49}$$

Thus, W - is text data (scientific articles, documents), Φ - word distribution in each subject, Z - distribution of topics for each word, Θ - distribution of topics in the document. The optimization task for searching for topics is as follows (3.50).

$$P\left(W|\Phi\right) \to max_{\Phi} \tag{3.50}$$

To use the EM-algorithm, we write explicit equations for the E-step and M-step (3.51, 3.52).

E-step:

$$\mathcal{KL}\left(q\left(\Theta\right)\, q\left(Z\right)||p\left(\Theta,Z|W\right)\right) \to \underset{q(\Theta)\, q(Z)}{\text{minimize}} \tag{3.51}$$

M-step:

$$\mathbb{E}_{q(\Theta)\, q(Z)} \log p\left(\Theta,Z,W\right) \to \underset{\Phi}{\text{maximize}} \tag{3.52}$$

The resulting expressions for the E-step (3.51) and M-step (3.52) allow to obtain the latent topics from text.

3.6 Modeling of self-organizing teams in the scientific environment

Self-organization of working groups is of great interest for scientific and technical organizations looking for new forms of practical organization of employees' work. Consideration of the phenomenon of self-organization as an alternative to the formation of working groups led to the construction of a model of the process of self-organization. Consideration of the life cycle of the working group concerning the goal allowed to introduce formal criteria for evaluating the effectiveness of the work of the group and predict its productivity.

An important factor affecting the productivity of the working group is the nature of the problem being solved. The author proposed his classification of the creative tasks of the oil and gas industry from the competencies necessary to solve them.

In the framework of the developed methodology for the life cycle of the working group and the classification of tasks, a mathematical model of self-organization of working groups was built to solve creative problems. Calibration of the mathematical model of the appearance of working groups was carried out on the data of Gazpromneft STC.

In the created system of performance indicators, a digital simulation experiment was conducted to identify the main characteristics of the self-organization of working groups.

As a result, the clustering of creative tasks on the effectiveness of solutions by various working groups was made, recommendations for creating organizational measures to increase the likelihood of self-organization of working groups were drawn up, criteria for assigning creative tasks to various types of working groups were highlighted, the main criteria for forming active working groups were identified.

Modeling group actions of individuals depend on potential participants in the group and goals. For example, fans of a football club can easily be combined into a group, but they do not have a specific goal. On the other hand, scientists can join a research group for a specific purpose, for example, writing a scientific article. Are the joining principles the same in these cases?

There are two ways of the appearance of groups:

1. Self-organization - the process of organizing employees due to internal factors, without external specific impact,
2. Formation and promotion of members of the group from the outside.

Example of forming a group in an organizational environment:

The head of the Department decides to form a working group to create an oilfield improvement system of two field developers.

An example of self-organization of groups in an organizational environment:

Several field developers have decided that together they need to improve the methods of processing data on the unique modes of operation of wells using machine learning algorithms and selflessly work together on the weekends on this task.

The decision to create a group inside leads to self-organization, the decision to create a group from outside forms a group. Note that in practice the process of the emergence of working groups is a superposition of self-organization and formation. However, for research purposes in this paper, the authors intend to consider the formation and self-organization separately to identify the common characteristics of these phenomena.

A group is created for a specific time and a specific task. In this sense, the signs of the project activity are evident - the uniqueness of the result and the limited resources to achieve the goal. Thus, it seems reasonable to apply the project methodology for assessing the effectiveness of the group as a project team.

The group structure is determined by the nature of the tasks to be solved. For tasks of mass service, for example, in working groups in call centers unite specialists with a precise, identical profile of competences. There is almost no separation of duties in such a group: typical maintenance tasks require standardized actions by group members. The workload is evenly distributed across the group members.

The group created to solve the creative problem is not homogeneous. Specialists with different competencies are needed to solve the creative problem. Figuratively, we can imagine how the task is decomposed into the competence of the participants.

Moreover, this is not a uniform distribution. The participants get different amounts of work within their competencies. For the above example (3.6), the task requires competences in the technical modes of operation of wells and methods of machine learning. What will happen if the competencies needed to achieve the goal are needed more than the group has? Each of the group members will perform work within their competencies, but the goal will not be achieved, as there will be outstanding work. This circumstance is a common situation, as with the wrong planning of groups (formation), and with self-organizing groups. The result of work in such a situation turns out to be negative, but the attitude to this result is different in the case of self-organization and formation.

The main criteria for the effectiveness of the working groups are the result of their activities and the timing for achieving this result - these are generally accepted organizational performance indicators. The study of the phenomenon of self-organization of working groups in the dynamics is a complex organizational task. Therefore, the author used in this work a mathematical model of the phenomenon of self-organization of working groups. The mathematical model makes it possible to study the most characteristic aspects of the phenomenon of self-organization, but it has a certain degree of approximation, inaccuracy.

A computational experiment by the created model of self-organization of working groups, which is presented below, evaluaтшпп the results of the work of various groups on various

tasks. In connection with this formulation of the experiment, the following research questions arise:

In connection with this formulation of the experiment, the following research questions arise:

1. How do groups self-organize? Which employees can self-organize and which ones not? How does self-organization is influenced by competences, experience, social factors?
2. What organizational conditions are necessary for the self-organization of groups in a scientific and technical environment?
3. What type of tasks do self-organized groups cope with more efficiently than formed ones?
4. What are the principles of forming groups for the most effective solution to creative problems?

Numerous publications are devoted to the evaluation of the effectiveness of research and development projects, as well as to the study of factors affecting the effectiveness of scientific activities, see, for example, [124, 125, 126, 127]. As a rule, in these works the research team is considered as a "black box", producing scientific results, and evaluation of its effectiveness is made only by the results, the internal structure of the research team is usually not taken into account. Self-organizing teams are studied in detail in [128]. Separately examines the motivating factors [127] and the factors influencing the performance of [126].

At the same time, the topic of modeling and analysis of teamwork is also well developed and actively studied since the middle of the XX century, see [131, 129, 130]. A formal description of the competence profile is the subject of numerous studies and publications, see, for example, [132, 130].

The first approximation may be a model limited to the existence of a fixed set of specific skills. In this case, the competence profile of each employee can be described as a vector of values, in which each coordinate describes the level of his knowledge of the relevant skill.

The vector describing the competence profile of the team is the result of the simple addition of the competence profiles of the participants. Such a model naturally occurs if we measure the level of competence by performance when performing the appropriate type of tasks. Then it is natural to assume that when working together in a team, the performance of the participants develops.

A similar vector can also describe the profile of the problem. A certain level of performance is required for each type of task in order to prepare and conduct a scientific study, taking into account the time limit.

In this study, the author consider small self-organizing teams, in which the initiative of creation comes from employees. This assumption corresponds to the real situation in most research teams, where the administration can motivate employees in various ways to apply for participation in a scientific confer-

ence or recommend to prepare an article for a particular journal, but the final decision, as a rule, remains for the researcher.

In this study, it is assumed that the list of competencies and the level of experience are the criteria from which the employee decides to join the team.

A set of topics that correspond to the sequence of incoming invitations from conferences and journals to which applications are open is considered as input to the model. One or more topics are known for each event or publication. Preparation of an article on a given topic requires a specific set of competencies.

The space of scientific activity determines competencies. In the oilfield services industry, the set of competencies differs from the set of competencies in the wood processing industry.

Experience describes a vector of a certain length and direction in the competence space. The projection of the vector on the axis experience competencies demonstrate experience in or the necessary skills.

A task, for example, the topic of a scientific article, also represents a vector in the space of competences. Topics may require competencies that authors do not possess individually. Each co-author closes only a part of the competencies required for solving the problem (writing the article).

3.6.1 Starting the team building process

The process of formation of the team starts with taking the first participant of the decision on the establishment of the team for the preparation of an application for a conference or article in the collection. Usually, this happens as follows. An unoccupied employee reviews the list of invitations and evaluates his / her competencies regarding the announced topics. If at least one of his competencies meets or exceeds the requirements of the goal, he decides to create a team and becomes its first participant. At the initial moment, the competence profile of the team coincides with the profile of the first participant. The following participants will join this team taking into account the requirements corresponding to the chosen topic, as well as the competence profiles of other team members.

3.6.2 Joining new members to the team

The second (subsequent) participant will learn from one of the team members about the purpose and assessment of the current competencies of the team. This information is shared between staff members who are quite familiar with each other. In the model, this is represented by a communication graph. Each participant evaluates his competences for the needs of the

team to achieve the goal and make a decision about joining the team. The solution is positive if at least one of the competencies of this participant when adding to the profile of the team brings it closer to the goal.

3.6.3 Finalizing the team

Because of the limited time to solve the problem, the time to form teams cannot be unlimited. If during the allotted period the team with the required set of competencies could not be formed, the process stops, the participants are released from their obligations and switch to the search for another task. If the team is successfully formed, we believe that its members are busy for some time and the result of this work is the publication.

3.6.4 Formal competency model

Let N denote the number of key skills required to work in a given subject area, W denotes the number of employees in an organization. Then the competence profile for the employee is called vector $\vec{\kappa}(w)$ (3.53).

$$\vec{\kappa}(w) = (\kappa_1, \ldots, \kappa_N), \text{ where } w \in W, \kappa_i \in \mathbf{R}^+ \tag{3.53}$$

The competence profile of a team T consisting of M a person is a vector of the same dimension N, which is defined as the sum of all team members (3.54).

$$\vec{\kappa}(T) = \sum_{i=1}^{M} \vec{\kappa}(w_i), \text{ where } T = \{w_1, \ldots, w_M : w_i \in W\} \tag{3.54}$$

Informally, the i-th component of the vector corresponds to the performance of a person and a team when performing tasks of a particular type. The theme profile p has the same type. Namely, it is an N - dimensional vector 3.55.

$$\vec{\kappa}(p) = (\kappa_1, \ldots, \kappa_N) \tag{3.55}$$

In equation 3.55 i-th component of the vector corresponds to the minimum performance of the team, in which all tasks of the corresponding type will be performed on time and with proper quality.

3.6.5 Key decision making model

The key functions are those that simulate the logic of decision-making at different stages of team formation implementing the process of team erection:

- $\alpha(w,p)$ describes the goal selection by the first team member, namely $\alpha(w,p)=1$ if the employee w considering the goal p makes a positive decision about team creation and $\alpha(w,p)=0$ otherwise;
- $\beta(w,T,p)$ formalizes the decision to join the team by the second and subsequent participants;
- $\gamma(T,p,t)$ models self-timer solutions at time t based on the comparison of the created team profile and the task profile.

This study assumes that α, β, and γ are deterministic Boolean functions that depend only on the competence profile of the individual, team, and task, respectively:

$$\alpha(w,p)=\alpha'(\vec{\kappa}(w),\vec{\kappa}(p)), \tag{3.56}$$

$$\beta(w,T,p)=\beta'(\vec{\kappa}(w),\vec{\kappa}(T),\vec{\kappa}(p)), \tag{3.57}$$

$$\gamma(T,p,t)=\gamma'(\vec{\kappa}(T),\vec{\kappa}(p),t). \tag{3.58}$$

Let K denote the whole space of possible values of the competence vector. Then the fact that in our model the algorithm of team building depends only on the participant's competence profiles, team and goal set the type of functions α', β' and γ':

$$\alpha':K^2\to\{0,1\},\quad \beta':K^3\to\{0,1\},\quad \gamma':K^2\to\{0,1\} \tag{3.59}$$

These functions can be described as the following logical formulas:

$$\alpha'(x,y) = 1 \iff \exists i(x_i \geqslant y_i) \tag{3.60}$$

$$\beta'(x,y,z) = 1 \iff \exists i[(x_i > y_i) \wedge (y_i < z_i)] \tag{3.61}$$

$$\gamma'(x,y,t) = 1 \iff \exists i(x_i < y_i) \wedge (t > \tau_{\max}) \tag{3.62}$$

3.6.6 Team building process

At the time of team building, the list of open problems P is fixed, and for each specific problem, $p \in P$ its profile $\kappa(p)$ is set. Also fixed set of employees W and for each employee, $w \in W$ the profile of his competences $\kappa(w)$is known. Also, the graph of communications between employees $G \subseteq W \times W$ is given. Another parameter is the time $\tau_{\max}$ during which the team should be generated.

At each step, the following occurs sequentially.

1. Each employee w_0 who is not included in any of the teams and has not received an invitation to join the team considers the list of goals P. If there is p_0 for which $\alpha(w_0,p_0) = 1$, the employee decides to create a new T_0 team and sends invitations to join the team to all neighbors in the communication graph G.
2. If the co-worker w_1 was not included in the team and received an invitation to enter the T_1 team created to solve the p_1 problem, he accepts the invitation if

$\beta(w_1, T_1, p_1) = 1$ and sends invitations to all his neighbors in the Gcolumn. Otherwise, the invitation is declined.

3. If the $\gamma(T_2, p_2) = 1$ condition is met for some T_2 team created to solve the p_2 problem, the team starts and all prompts are canceled.
4. If for some team T_3, created to solve the problem p_3, after a specified time $\tau_{\max}$ condition $\gamma(T_3,p_3) = 0$, this team is disbanded and all invitations are canceled.

Even though α, β and γ are deterministic, the algorithm admits a large degree of uncertainty, which is associated with the non-deterministic nature of the interaction of objects within the system. In particular, the result is significantly affected by the following parameters, which are implemented probabilistically:

- the order of consideration of the task list by a free employee:
- the order of the consideration the employee received invitations;
- the order in which employees are selected to apply the next step of the algorithm.

The constructed model is the basis for further research of the process of formation and functioning of project teams in the scientific environment. In particular, on its basis, it is planned to develop a methodology for assessing the effectiveness of research activities. Also interesting is the refinement and expansion of the model, in particular:

- Competency models can be refined using fuzzy logic.
- When modeling long-term periods, there is a need to take into account the professional and career development of employees and the associated changes in their competence profiles.
- Functions α, β and γ, describing the process of making key decisions, can be refined by taking into account other individual and team characteristics, as well as the specifics of the tasks.
- The team building algorithm can have a more complex iterative logic that takes into account different approaches to flexible project management.
- A separate study deserves the situation with the unsuccessful completion of the project. Regarding scientific activity, this means that the written publication has not been accepted for publication, but the results are a good start for further work. In the current work, the author made the assumption that employees do not write *into the desk*, and each co-authorship leads to publication.

3.7 Methodology of the co-authorship graph

The current practice of constructing graphs of co-authorship involves the use of the mathematical apparatus of graph

theory. Traditionally, undirected graphs are used to construct co-authorship graphs. The co-authorship graph provides a visual visualization of the chosen scientific community and allows analysis using such common graph metrics as: Betweenness centrality [133, 134, 135] and Closeness centrality [136, 138, 137]. These metrics, as well as the Degree metric, are intended for formal selection of important vertices of the graph.

3.7.1 Bipartite graphs

An essential aspect for the construction of the graph of co-authors is the selection of data for analysis. Usually, researchers use public bibliographic information containing a list of co-authors. The source of such information may be Google Scholar, ArXiv and other online libraries. The consideration of open scientific communities is as interesting as the narrowing of the sample to one country [139], industry [140] and even organization [122]. Adding fields related to the author's affiliation into the graph allows us to research the relationship of organizations. As an example, in [140] the authors analyze the links between research institutes and industrial research centers in the oil and gas industry. This approach to sampling allows us to analyze the topology of relationships between organizations, based on the authors ' affiliation with the organization.

Note that all the above studies do not take into account the content of research articles. This feature will be important in the future. The average number of co-authors may vary depending on the industry, but overall the number of co-authors is growing. We note this fact as a structural feature of the study area.

In the above study, the co-authorship graph is built on an undirected graph. The authors are equivalent in co-authorship, although it is not. In the work of the author [141] the structure of the team of co-authors is analyzed, and possible roles in the research process are formulated.

Besides, in the traditional construction of the graph of co-authorship, information on all joint research work is contained in the edges of the graph. Often edges are drawn with different thickness or color depending on the number of collaborations, but this characteristic of edges is not considered in the context of graph metrics, as it does not reflect the communication meaning of re-authorship. Taking into account these limitations, we formulate the following research questions:

- Are there other ways to construct a co-authorship graph?
- What are the advantages and disadvantages of different ways to build a graph of co-authors?
- What are the quantitative, comparable characteristics of co-authorship graphs?

In the above studies, the graph of co-authorship is constructed as an undirected graph: articles become equivalent

edges connecting authors. The author of this study believes that the construction of the co-authorship graph as a bipartite graph will be more informative. Such an approach makes it possible to include information on scientific articles in the co-authorship graph. The Figure 3.8 shows the basic principle for constructing a graph of co-authors by a directed bipartite graph.

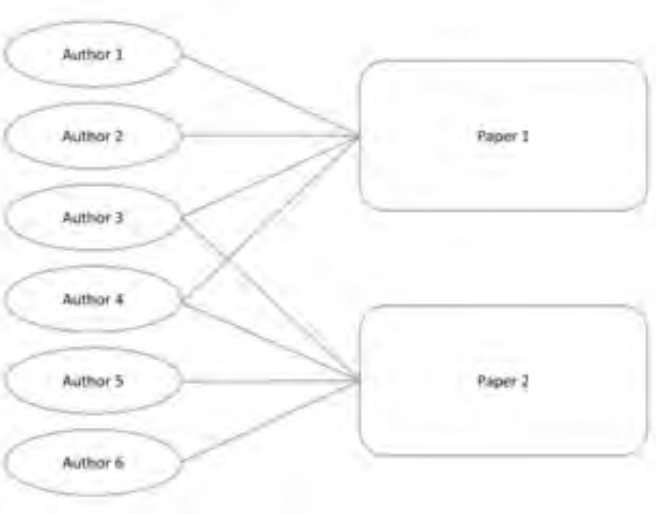


Figure 3.8 — Bipartite graph of co-authorship.

The advantages of this approach are that in the graph of co-authors it becomes possible to save for further analysis the bibliographic information about the article:

1. the title of the article
2. publication year
3. the publisher
4. keywords

Note that the traditional representation of a co-authorship graph as an undirected graph is a projection of a bipartite graph onto the set of vertices of the authors. Let us explain this in more detail. An oriented graph $G = (V,E)$ is called bipartite if the set of its vertices can be divided into two parts $a \cup P = V$, such that

- no vertex in a is connected to vertices in P
- no vertex in P is connected to vertices in a.

In this case, A is a set of authors, P is a set of articles. A and P are parts of the graph G. Note that the graph G can be either complete or incomplete depending on whether the authors have connections to all the articles. Shown in Figure 3.8 bipartite graph is incomplete. Let's denote G_A projection of the graph G on the set of vertices a. The graph G_A is a traditional representation of the graph of co-authors and is shown in figure 3.9.

From the figure 3.9 we can see that when constructing the projection, only cumulative characteristics of co-authorship can become the attributes of edges of the graph G_A. For example, the number of co-authorships of two authors.

3.7.2 Modeling of the co-authorship graphs

The following stochastic approaches are widely used to model co-authorship graphs as a social network:

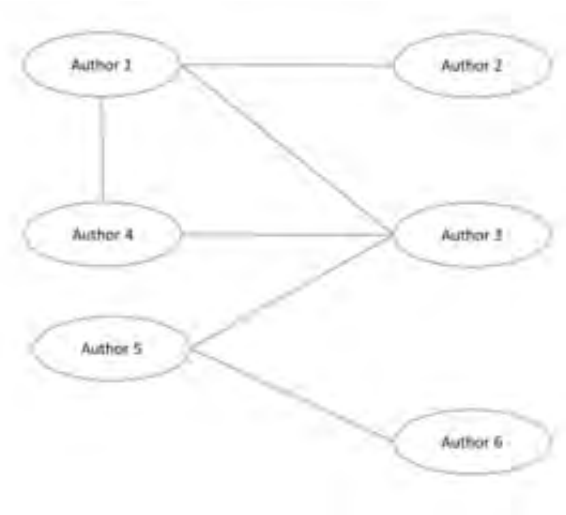


Figure 3.9 — Undirected co-authorship graph.

1. Random graphs,
2. Small-world model,
3. Preferential attachment model.

One of the essential limitations of stochastic models is the fixed number of vertices and their constant growth. In practice, the number of potential co-authors changes in the organization. It is also important to understand that stochastic models aim to model a graph with specific parameters. Such as clustering and density.

On the other hand, the formation of small groups, which include a group of co-authors of the scientific article, is modeled by the principle of additional competencies related to the class of deterministic methods of creating graphs of co-authorship.

A combined machine learning-based approach is used to predict new vertices of the co-authorship graph. There are a preliminary selection of the authors' features, statistical indicators of activity for the last few time intervals, as well as structural indices of influence and local metrics in the co-authorship network. The results obtained in this study allow the author to conclude the applicability of communication prediction methods for the analysis of collaborative behavior patterns in a large organization with a dynamic team structure, as well as changing external and internal factors affecting individual and collective publication activity.

The basis for predicting changes in the graph of co-authors for the scientific and technical center are the following components:

- Current structure of the co-authorship graph,
- External impacts,
- Internal changes.

Consider each of the components in more detail. The current structure of the co-authorship graph represents a set of metrics describing a given co-authorship graph. These metrics include the following:

- For the edges
    - Common Neighbours (CN)
    - Salton Index (SI)
    - Jaccard Index (JI)
    - Hub Promoted Index (HPI)
    - Hub Depressed Index (HDI)

    - Leicht-Holme-Newman Index (LHN1)
    - Preferential Attachment Index (PA)
    - Adamic-Adar Index (AA)
    - Resource Allocation Index (RA)
- For the vertices
    - Degree centrality
    - Betweenness centrality
    - Closeness centrality
    - Harmonic centrality
    - Clustering

Each of these metrics represents a specific set of features of the co-authorship graph, affecting the forecast of its changes. External influences to the scientific and technical center compose of the publication policy of editorial offices publishing scientific articles. In the simplest case, the lack of opportunity to publish an article due to limitations on the volume of the issue of the journal results in unsuccessful co-authorship. The main dependencies of the publication activity of the scientific and technical center on the editorial staff are discussed in the [141]. Changes in the staff cause internal changes in the scientific and technical center. New employees come to the organization, some employees leave. In the process of mentoring and training, employees acquire new competencies. As a result of research, new research and scientific groundwork are born. Often, changes in the internal requirements for the quality of publications can also cause structural changes, confirming the principle of "publish or perish", and affecting both the structure

of the team and the activity parameters of individual employees and research teams. Let us consider in more detail what the forecast of the development of the graph of co-authors for the scientific and technical center is. By development, we mean the emergence of new peaks and edges. The graph of co-authorships can be considered as the cumulative total for the period, as well as incremental changes over the years. Next, we will consider the fact of authorship as a sign of the top of the graph of co-authorship. In other words, an employee represented by the top of a graph of co-authorship can either write or not write an article in the next period. The forecasting process, in this case, will solve the problem of binary classification. For each employee, the probability of creating an article on a specific topic will be determined. The article is a collective effort of the work of co-authors with a specific set of competencies that have found their application in the purpose of the study. This is the basic idea of the principle of complementarity of competencies. Authors with the same competencies do not have a rational justification for combining to conduct scientific research. Let us consider competences as attributes of graph vertices. To identify the competencies necessary for writing an article, we will use keywords, and in their absence, the method of thematic modeling of the text of the work.

3.8 Modern processes of labor organization based on agile methods

Agile methods of software development are widely used in various industries. Writing code is the process of creating a logically structured text as well as writing a scientific article. Teamwork in writing scientific articles requires a division of labor to improve productivity, just as writing code requires the allocation of specialists for testing and documentation.

The use of the role model of agile methods seems to be a promising cross-industrial experience for application but needs theoretical verification. One of the variants of testing hypotheses, which proved to be in conditions when the formulation of a real experiment seems to be highly expensive, is the method of simulation. The author sees additional benefits from the institutionalization of the process of writing scientific articles and the use of proven industrial performance indicators for its evaluation.

The proclamation of the basic principles of agile methods in the form of the Manifesto [142] indicated the urgent need to move to more effective methods of software development. The determination of this step has repeatedly proved itself in practice and later found theoretical justification [143]. The essence of agile methods can be described in different ways, but for this study, we have chosen the following phrasing:

1. Priority of team interactions

2. Priority of a working program code
3. Priority of the reactions under a plan

Modern methods of writing scientific articles remain on the positions of consistent, “waterfall” approach. This approach was appropriate in Isaac Newton’s time when one unique mind worked on the commitment of his life. In the conditions of the current speed of exchange of scientific information, singles remain out of work. Research teams replace them. It is intuitively clear that the coordinated work of the research team of co-authors depends on their productivity: the optimal ratio of quality and speed of publication of research results in the form of scientific articles available to the broadest range of stakeholders. In agile software development techniques, team education is based on the principles of self-organization [144, 128]. Self-organizing teams in [145] are divided into three types:

1. “Pilots of the aircraft”,
2. “Computer teams”: creation of new software products,
3. “Brainstorm teams”: solving single complex problems.

For further research, we will be more interested in the type of “Computer teams”.

3.8.1 Team size

Team sizes play an essential role. Agile software development practices consider small (5-7), large (10-50), and extra-large (100-200) [146] teams.

It is important to note that the above estimates converge with those obtained in [122] for teams of co-authors: current creative teams of co-authors on average consist of 3 participants. In what follows, we shall mean that the number of team members with an average of 3 co-authors.

3.8.2 Team assembling

Agile methods [142] mean by self-organization of the team only limited control from outside. The author study how teams are assembling in detail.

As we said in section 3.6, it is necessary to understand that the team is assembled for a specific purpose. The authors of the [147] study propose an empirical probabilistic algorithm for joining a new participant to an already formed group.

On the figure 3.10 depicted the probability-based team formation algorithm. p is the probability for new members, q -for group members [147]. The authors of [147] assess the impact of the internal structure of the team on its expansion.

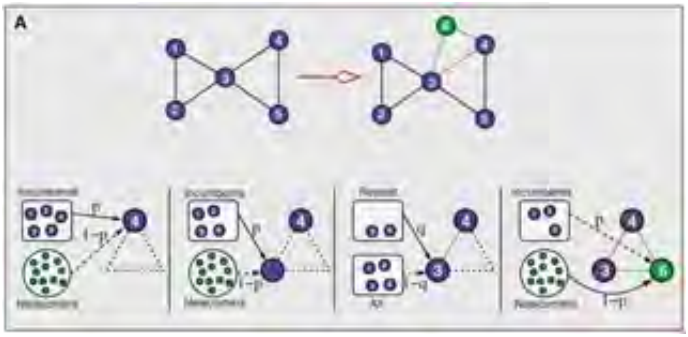


Figure 3.10 — Probability-based team formation algorithm [147].

In [145] it is noted that the main factor for self-organization of teams is individual competencies. The competencies of each participant are evaluated regarding usefulness for achieving the goal. In studies [149, 148] it is argued that such an assessment leads to the emergence of a system of statuses of team members, which is expressed in the hierarchy of communications. For the present study, it is sufficient that:

1. the goal is forming the basis for the team;
2. the goal demand for the competencies of team members;
3. participants assess each other's competencies to achieve a goal.

The fundamental algorithm of team formation for two participants can be represented as the following time sequence (3):

Table 3 — Team flow.

Step	Action	Result
One	The competences and experience necessary for the achievement of the goal are defined.	The goal
Two	The first member of the team self-evaluate hist competences and experience to the goal.	The first member
Three	The first member decided to create a team for the goal.	The team with one participant created
Four	The second participant learns from the first participant about the scope of the goal.	The goal has not covered by the competences yet.
Five	The second participant evaluate the rest of the required competences.	The competences of the second participant are in demand for the goal.
Six	The second participant decide to join the team.	The team has two members.

Shown in table 3 sequence describes the main action of assembling the team. We can say that after joining the team members have their profile of competencies to achieve a given goal. The team competencies are a superposition of competencies of the participants. The experience of the participants covers some of the competencies required to achieve the goal, and some are not.

The second participant joins the team with the set of competencies different from the first participant. For the convenience of the further notation let us formulate the following statements:

- Employees team up to achieve the goal;
- Team competencies are a function of the competencies of the participants;
- The unification of the first participant with the team to achieve the goal takes place on the same principles as the union team of n participants with $n+1$ participant.

Let us consider the phenomena of the uniting of the first and second participants in more detail. The organizational environment defines the dimension N_{comp} of the competency space. Each participant of the organizational environment a has a vector of competencies c_i such that $i \in N_{comp}$. The experience e_i characterizes each competence c_i. The participant's experience is a natural number, $e_i \in \mathbb{N}$. As a result, we can say that the participant has a vector of experience in the competence space. Note that the competence space of the organizational environment has a significantly larger dimension than the vector of competencies of the participant. Initially, the team t_0 does not contain participants and does not have its competencies (Fig. 3.11).

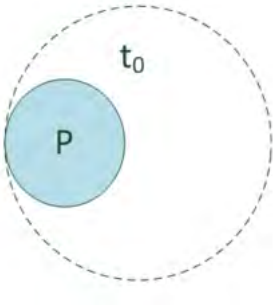


Figure 3.11 — The scheme of the team without participants.

Let P denote the goal to unite the team, c_j denote the vector of competences, and e_j denote the experience for each competence required to achieve P. As a result of successful

uniting, the t_1 team will be formed for the achievement of the goal (Fig. 3.12).

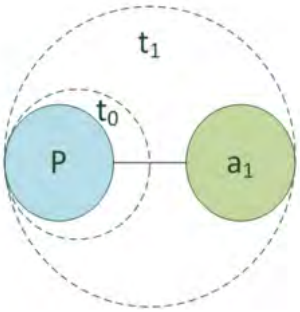


Figure 3.12 — The scheme of the team with one participant.

The team t_1 has a new vector of competencies. Since t_1 has one participant a_1, then the vector competence t_1 is the same as the vector of competences of a_1. The team t_2 will be formed when participant a_2 joining team t_1 as shown in figure 3.13.

Since the participant joins all elements of the team, it is possible to bring the scheme (Fig. 3.13) to the form of the team graph (Fig. 3.14).

The goal P is an attribute of an edge linking a_1 and a_2. We can convert the team graph with two members to an equivalent graph shown in the figure 3.15.

For the case of writing scientific articles, the graph of the team t_2, shown in figure 3.15 denotes $g(t_2)$ and is called a co-authorship graph, where goal P implies a scientific paper. There is

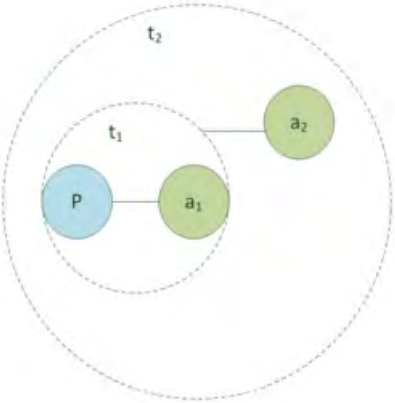


Figure 3.13 — The scheme of the team with two participants.

no information about the history of the team assembling in this notation. An example of a fragment of a co-authorship graph shown in the figure 3.16. The vertices of the graph are the researchers, and the edges are the joint scientific publication. The co-authorship graph is an undirected network.

Note that often for clarity the size of vertex reflects number of scientific articles written by the participant.

Team code

Let us introduce the concepts of *Full Team Code (FTC)* and *Residual Team Code (RTC)*. These concepts play a crucial role in the formation of the team. The ingredients of the team

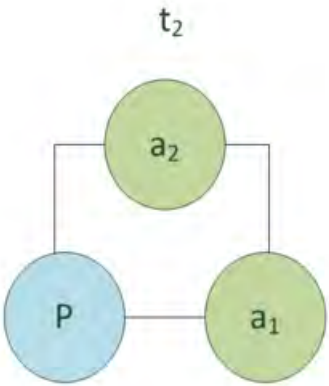


Figure 3.14 — The graph of the team with redundant connections.

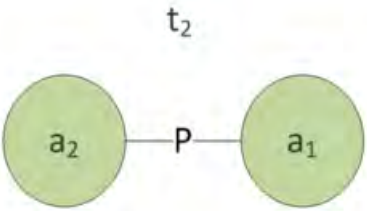


Figure 3.15 — The graph of the team with two participants.

code are competencies. By its type *Full* and *Residual Team Code* are vectors in the space N_{comp} .

Consider how a participant evaluates his competences for the needs in competencies for achieving the goal.

Figure 3.16 — Fragment of a co-authorship graph.

The characteristics of the goal P are the basis for the formation of the team. That is, the first team member a and the goal P should be combined based on the concept of competencies. In other words, the necessary conditions for achieving the goal should be the possession of a a specific set of competencies and experience. Concerning sets of employee competence a and goals P should be in the same space and have intersections. The presence of the intersections will lead to a team.

Let us introduce the evaluation function as $\Phi(P,t,a), \Phi \in [0,1]$. The result Φ be the probability of connection of participant a to the team t for the goal P. Then the function Φ for n-th participant can be written as $\Phi_n(P,t_n,a_n)$.

As the participant joins, the team's competence vector will change. It will include the competencies of new participants, and experience on the same competencies will develop (3.63).

$$ut_{n-1} = \prod_{a_j}^{a_{n-1}} \sum_{i}^{N_{comp}} \left\{ c_j * e_i \right\} \tag{3.63}$$

The value of ut_{n-1} will be called *Full Team Code (FTC)*. The *FTC* characterizes the team's potential to achieve goals.

Expression 3.64 represents the function Φ following the above algorithm 3.

$$\Phi = P \cdot \prod_{a_j}^{a_{n-1}} \sum_{i}^{N_{comp}} \left\{ c_j * e_i \right\} \cdot a_n \tag{3.64}$$

An important semantic part in the expression 3.64 carry out the component rt_n^P, which the author calls *Residual Team Code – RTC*. The expression for *RTC* is 3.65.

$$rt_n^P = P \cdot \prod_{a_j}^{a_{n-1}} \sum_{i}^{N_{comp}} \left\{ c_j * e_i \right\} \tag{3.65}$$

The *RTC* rt_n^P characterizes uncovered t_n team competence for the goal P. The zero vector as *RTS* characterizes the complete staffing of the team's competencies to achieve the goal. Now we can convert the expression 3.64 to a more appropriate view 3.66.

$$\Phi = rt_n^P \cdot a_n \tag{3.66}$$

The expression (3.66) has an intuitive meaning:

In order to assess the possibility of joining the team, the new participant must find out whether he/she has the necessary experience in the required competencies to fulfill the goal, taking

into account that the existing team has already closed some of the necessary competencies with its experience.

In the works of [150, 151] this principle of team formation is called complementary.

Team homogeneity

We have considered the formation of teams by complementarity (complementarity) of competencies. The second driving force for the formation of teams is homogeneity.

The homogeneity of groups in social networks, or the propensity of people with similar characteristics to form links between themselves, also called hemophilia, is an essential factor in the formation and evolution of social networks [43]. Many studies observed dynamic structure hemophilia [42, 41], which occur in parallel two processes. On the one hand-similar individuals form social connections (social selection). On the other hand, people who are already connected adopt each other's behavior (social influence). The combination of these factors results in a homogeneous social system in which there is a connection between individuals with similar behavior and characteristics, and the nature of the connection can be both formal and informal.

Although connections between individuals with similar characteristics are more likely than connections between dif-

ferent ones, the level of similarity is also essential. In work [44] it was shown that social similarity by more than one indicator leads to the fact that people are less likely to form relationships with each other. The author explains this effect by the fact that people who are too similar in many characteristics, as a rule, cannot bring something new and constructive into a mutual relationship or a team. Productive cooperation requires not only the similarity of interests but also different professional and life experience, which allows us to offer multidimensional approaches to its solution.

The main unifying factor in the team is the competence of the participants that affect the achievement of the goal. Based on the concept of *RTS* introduced earlier, we can consider the residual competencies of the participant, that is, competencies that are not required to unite the team to achieve the goal. The influence of this part of competencies on the team can both strengthen it and weaken it during work.

Team work

The team members define the beginning of work to achieve the goal, and it does not depend on the process of team formation. Only a running team may have performance indicators. For example, an important for the scientific domain

term "scientific capacity" means nothing more than the work performed by a team that has not empty *RTS*.

Formation of the central system of internal interaction within the team according to the study [40] occurs when meeting participants. Thus, for this study, the author will ignore the time of establishing stable operation of communication channels.

In agile methods of software development, the most attention is paid to communications within the team [152] and with external agents [153], which are mostly also a team, but in a broader sense.

The author formulates the following statement:

Characteristics of communication channels correspond to the nature of the team.

Thus, measuring the work of communication channels can make conclusions about the nature of the team. Note a necessary consequence: this type of measurement of team performance does not create an additional burden on employees, in contrast to the methods of evaluation based surveys.

The question of measuring the contribution of individual participants or the result of the team is considered in many works [154, 155]. All researchers emphasize the fact that we need to measure both, team performance, and individual performance. The study [156] shows the following measurement scheme (Fig. 3.17).

For example, the measurement of Individual Performance through surveys is explored in [157] by introducing Creative So-

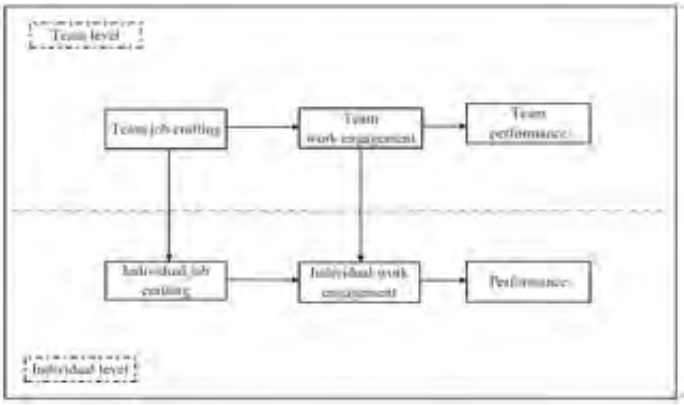


Figure 3.17 — Team and participant performance measurement levels [156].

lution Diagnosis Scale (CSDS) – a scale of creativity. Measure the individual performance of an employee on such a scale the authors of [157] suggest using the Consensual Assessment Technique, which requires additional effort from employees. The strengths and weaknesses of the survey method for measuring Individual Performance are outlined in the fundamental work of [158].

The question of the method of measurement of Individual Performance finds an interesting statement in the modern concept of "sensible organization" [159]. The authors of the [159] study, in addition to measuring traditional digital communication channels, put bracelets on employees that track movements and other body parameters.

The problems of the dependence of the team performance on the team structure are considered in the study [122].

Scrum Methodology

One of the most common agile practices of teamwork is Scrum [160]. Scrum is designed to produce the best possible results for team development of complex, intelligent products. There are three fundamental roles in classic Scrum:

- Product owner - responsible for compliance with the objectives;
- Scrum master - responsible for the effective interaction in the team;
- Development team.

The recommended size of the Scrum team – 5-7 people corresponds to the limitations of this study. According to the ideology of Scrum [160], larger team require significant resources to communications, while the smaller size teams reduce the size of work that a team can complete per unit of time.

The basis of Scrum is Sprint, during which team creates the product. Sprints have the same duration throughout the product creation processes, one week is recommended. The task of Sprint is to materialize the product in its current form. In this study, the product is a scientific article.

The Scrum methodology declares the need for certain activities that are not related to research and writing text, which lead to better results. Also, Scrum sets a specific beat for these additional activities.

Let us introduce indicators that are affected by the application of Scrum to the process of writing scientific articles:

1. accelerate the exchange of messages in the communication channels;
2. duplication in the absence of timely communication on research progress;
3. non-compliance of the written article with the publication rules.

From the formalism of the graph of co-authorship application of Scrum would lead to the selection of vertices of the graph provides the functions *Product owner (PO)* and *Scrum master (SM)* (Fig.3.18).

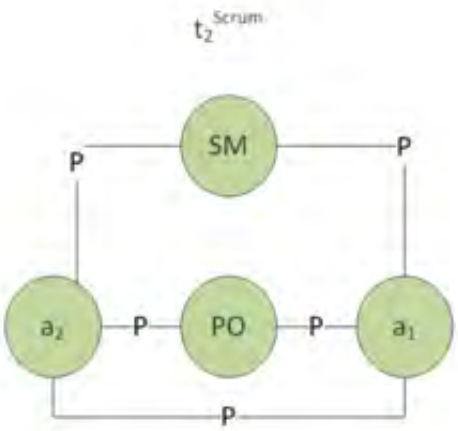


Figure 3.18 — The co-authorship graph with Scrum roles.

With the graph $g(t_2^{Scrum})$ shown in Fig.3.18 we can perform the conversion similar to the one made above with the

graph $g(t_2)$. As we can see, the Scrum roles of PO and SM connect the vertices a_1 and a_2. It follows that PO and SM are the characteristics of the edges of the graph connecting a_1 and a_2. The transformed graph of co-authorship using Scrum roles is shown in Fig. 3.19.

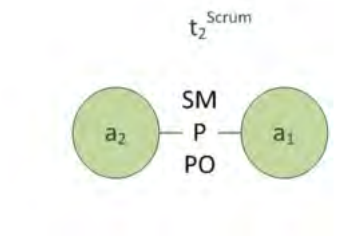


Figure 3.19 — The co-authorship graph with Scrum attributes.

The role of Scrum according to [142] should not interfere in the substantive work of the team, but only to speed up information exchange and to eliminate information barriers. We formulate this as a hypothesis, the formal proof of which is postponed for further research:

The introduction of Scrum roles in the co-authorship process does not change the appearance of the co-authorship graph.

Now consider the performance indicators of the teams.

Team performance metrics

In modern work [161], considered the development of indicators that measure the rate of transition of the product from the research phase to the development phase. Authors of [161] proposed an integral model for such indicators. They consider knowledge to be the object of measurement, and indicators are based on the Knowledge Management process. They did not offer any specific KPIs, but describe the spatial axes of their model such as processes, tools, and people.

The link between the team's ability to assemble and its productivity was investigated in [162]. It is worth noting that in [162] the assembling of a team implies the formation, not self-organization.

The composition of the team in time is not constant and say that the assembling of the team at one time or another time is not completed correctly. Participants can leave a team and participate in several teams at the same time. An important milestone in the work of the team is determined by zero *RTS* when the team members present all the competencies necessary to achieve the goal. Let us formulate this as statements:

- **A team is complete if and only if its RTS is equal to the zero vector in the space N_{comp}.**
- **The minimum time in which the RTS became equal to the zero vector is called the Assemblance Time (T_c).**

Note that T_c can be longer than the time allotted by the publishing house or the Program Committee of the scientific conference for preparation. Thus, the article will not have the required qualities in time and will not be accepted for publication.

Indicators that most accurately reflect the dynamics of the work will be based on changes in the dynamics of all parameters of the team. Let us introduce the function of the application of experience by the team in the competencies defined by the goal: $E(P,t)$. Factors that negatively affect E will be the complexity of communications $\Xi(g)$ within the team and the need to engage in activities not aimed at creating scientific articles: $\Gamma(t)$.

Both $\Xi(g)$ and $\Gamma(t)$ will increase the time required to write a scientific paper. Thus, the team may not achieve the goal in a particular time.

Let's formulate two considered reasons for not achieving the goal by the team:

Abandoned Scientific Article (AAA) will be recognized as an article that does not meet the time frame of the publication process with the required quality.

The ratio of the number of abandoned articles ($Frac_{notpub}$) to the number of published articles is an indicator of the performance of the process of writing scientific articles.

Another more obvious performance indicator is the time taken to publish a scientific paper T_{pub}.

3.9 Text Analysis

Let us consider text mining and analytics (TM&A). The mission of TM&A is to turn text data into high-quality information or actionable knowledge. Concerning this mission there are two important conditions would be mentioned.

1. Minimization of human effort (on consuming text data);
2. Supplying knowledge for optimal decision making.

Related to text retrieval, which is an essential component in any text mining system we must note that:

1. Text retrieval can be a preprocessor for text mining;
2. Text retrieval is needed for knowledge provenance.

The broad picture of TM&A is shown in figure 3.20.

3.9.1 Topic Detection

In recent years, the methods of topic modeling are rapidly developing. Recent studies have led to the development of several main areas: probabilistic [1], based on SVD [2] and generative [4]. Topic modeling defines each topic as the distribution of a certain number of words with specific probabilities. Most modern topic models are based on the Dirichlet distribution (LDA, Latent Dirichlet Allocation) [3]. It is hard to

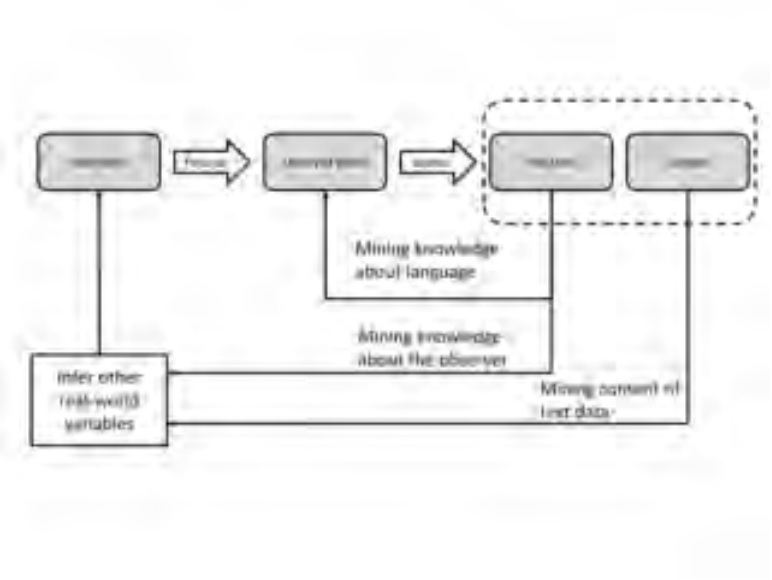


Figure 3.20 — Landscape of Text Mining and Analytics.

imagine that such a universal distribution as LDA would work adequately for any text. We need to fine-tune the algorithm for a specific problem domain. Therefore, the author focused on the primary world source for scientific and practical articles of the oil and gas industry - The OnePetro library. It is important to note that the OnePetro library covers a wide range of engineering disciplines and contains texts in English dedicated to the practical aspects of the application of new technologies in the oil and gas industry. The authors of the articles in the OnePetro are employees of oil companies from all over the world.

The precise formulation of the problem of topic modeling is as follows. Let fixed dictionary of terms W, from which elements are formed documents of given collection D containing

documents $d \in D$. For each document d its length n_d and the number of $n_d w$ uses of each term $w$are known. Let $\Phi = (\varphi_{wt})$ be a matrix of term distributions w in topics t, and $\Theta = (\theta_{td})$ be a matrix of topic distributions t in documents d. Then the problem of thematic modeling is to find such matrices Φ and Θ that the equation is became valid (3.67).

$$p(w|d) = \sum_{t \in T} \varphi_{wt} \theta_{td} \tag{3.67}$$

In equation 3.67 φ_{wt} is the probabilities of the terms w in each topic t, θ_{td} is the probabilities of the t in each document d, and $p(w|d)$ is the probability of the term w appearing in the document d.

The equation 3.67 can be represented as $\Phi \cdot \Theta$. It is easy to show that this problem has many solutions (3.68).

$$\Phi \cdot \Theta = \Phi \cdot \Lambda \cdot \Lambda^{-1} \cdot \Theta = \hat{\Phi} \cdot \hat{\Theta} \tag{3.68}$$

In equation 3.67 $\hat{\Phi} = \Phi \cdot \Lambda$, and $\hat{\Theta} = \Lambda^{-1} \cdot \Theta$.

It follows from the equation 3.68 that the matrices $\hat{\Phi}$ and $\hat{\Theta}$ will also be solutions to the equation (3.67). But not all matrices Φ and Θ contain well-interpreted topics. Thus, the problem (3.67) must enter the conditions which would produce adequate and exciting topics. Figuratively we can say that it is necessary to digitize the specificity of the subject area of the text for embedding in the algorithm of the search of optimal matrices Φ and Θ. Note that when using LDA to create a thematic model, this setting is not made in the subject area.

For solving subtask configure the topic model for the subject domain, the author has used the mechanism of regularizers.

Let $p(t)$ be the distribution of topics in the document collection:

$$p(t) = \sum_{d} p(d)\,\theta_{td} \tag{3.69}$$

Then it may be helpful to ad regularization based on the Kullback–Leibler divergence ($\mathcal{KL}$) shown in 3.70.

$$\mathcal{KL}(\Theta) = -\tau \sum_{t \in T} \ln \left(\sum_{d \in D} p(d)\,\theta_{td} \right) \to max \tag{3.70}$$

Where τ is a regularization parameter to be chosen depending on the subject area of the document collection. The requirement of maximizing $\mathcal{KL}(\Theta)$ lead to the zero probability of the appearance of documents and the greater sparsity of the matrix Θ. The second mechanism for regularization may be the opposite - increasing the probabilities for topics that are present in many documents. Such subjects are called noise. To obtain the seals of the rows of the matrix Θ with noise subjects can apply regularization (3.70) with the opposite sign. Thus, the matrix Θ after the regularization is to introduce an alternation of zones of sparsity for the major subjects and seals for noise topics.

The resulting topic model should be formally checked for quality. Quality metrics of the topic model must be embedded

in the learning process. Moreover, after reaching the formal convergence criteria of the model the history of metric evaluation needs to be visualized. The primary metric for detecting convergence of the topic model is the metric Perplexity calculated by the formula (3.71).

$$\mathcal{P}(D, \Phi, \Theta) = \exp\left(\frac{-1}{n_d} \sum_{d \in D} \sum_{w \in d} n_{dw} \ln\left(\sum_{t \in T} \varphi_{wt} \theta_{td}\right)\right) \quad (3.71)$$

The Perplexity metric is not normalized and therefore cannot be used to compare the convergence of different models. The prevailing logic is that the smaller the Perplexity, the better the model. Therefore, to decide on sufficient convergence, the models are guided by the fact that Perplexity ceases to decrease significantly with the growth of the number of training iterations.

The resulting model of topics can be considered as a soft clustering. In this case, visualization tools used for clusters can be applied to the obtained topics. For example, manifold-based learning methods such as t-distributed Stochastic Neighbor Embedding (TSNE) and multi-dimensional scaling (MDS) can be applied. The results of the TSNE algorithm depend on the chosen metric of the distance between the vectors. So we should find a suitable one. When the dimension of the vector space in a few hundred use the following metrics:

- Cosine distance: $\frac{v_1 \cdot v_2}{\|v_1\|_2 * \|v_2\|_2}$
- Euclidean distance: $\|v_1 - v_2\|_2$

To effectively use the visualization of the topic model with the help of learning methods based on manifold-based methods, it is necessary to present the words that make up the subject in Vector Space Model (VSP). This procedure is called word embedding. For it often uses the method GloVe described in the study [77]. An alternative method of word embedding is FastText [163], so the author of this study decided to make a qualitative comparison of both methods of word embedding on the selected collection. Both methods learn vector representations of words based on how often words are used together. The difference between them is that FastText can be called "predictive", and GloVe is based only on the frequencies of words. In this light, GloVe is much simpler, and the author of this study believes that simplicity in business is the key to efficiency.

Library BigARTM [164] allows you to build several regularizers and control groups of subjects consistently. This tool is unique at the time of writing this study. Widely used methods of constructing topic models based on LDA do not provide such opportunities.

3.9.2 Sentiment Analysis

Sentiment analysis of text is intended to detect in the texts of the emotionally charged vocabulary. Sometimes researchers distinguish the term "Opinion mining", thus empha-

sizing the task of searching in the texts of value judgments. In addition to the academic study of the tone of the text as one of the sections of computer linguistics, there are some applied studies aimed at improving the management decision-making process.

The use of recurrent and convolutional neural networks for sentiment analysis made it possible to achieve much higher accuracy compared to models based on logistic regression.

The author focused on the method of choosing the optimal architecture and hyperparameters of the neural network, which allows training the classification model on a public dataset containing estimates, and then to predict the text fragments from scientific and practical articles containing estimates with a given degree of accuracy.

The methodological approaches applied by the author can be presented in the following research framework of the study (Fig. 3.9.2). Let us consider in more detail each of the elements of the systematic framework.

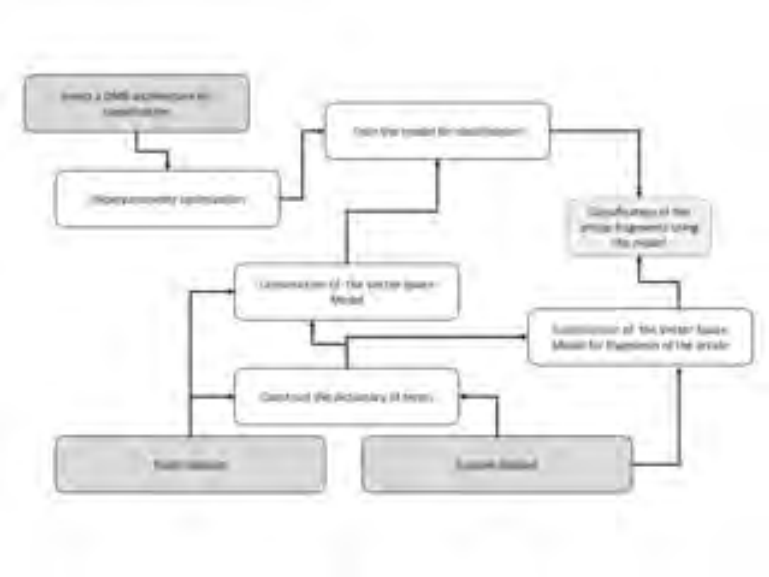


Figure 3.21 — Research framework for the study of emotional coloring of the texts.

Chapter 4. Approbation and Results

4.1 The set for the experiment for direct and inverse problems.

Two different domains are closely intertwined with each other and provide a comprehensive, in-depth look at the object of the study. Let us list these domains:

– The study of the STC via external presentation,

– The study of the STC from the inside.

These two tasks are oppositely directed. Let's take a closer look at them.

External manifestations include digital artifacts of the organization's activities - published scientific articles, conference materials, information sites on the Internet and news about the company. The study of digital artifacts is carried out using approaches based on the analysis of texts and co-authorship.

Then we are studying the STC from the inside; the research includes modeling of scientific activity, the efficiency of production processes, self-organization of small creative teams and models of scientific organization personnel.

Suppose we consider a particular organization with a certain number of employees, budget and work plan. Our interests are in the roots of the effectiveness of this organization. From this point of view, the following issues are essential for us:

- What is the intellectual potential of the organization? What research organization can perform independently, and which must be conducted in conjunction with other scientific organizations. It is evident that when carrying out joint research, communication costs arise and the study needs additional coordination. However, for efficiency is important not only the original boundary " can or can not," but also the distribution of time and effort.
- Performance is known to degrade under high load conditions. However, for questions of efficiency, this effect must be considered in the dynamics, as the return from degraded to regular also takes time. Also, it is important to segment the load by employee type. Beginners can be both overloaded and underloaded with work. Depends on the turnover of staff. However, a load of experts significantly more significantly affects the efficiency. The effects of mental fatigue of experts dramatically affect the efficiency.
- Scientific assets of the organization are exhausted? What is the dynamics of creating scientific assets? Are there any breakthrough trends in research conducted within the organization? Who is involved in the creation of a scientific justification?

The parameters of the organization are impossible to measure. We need essential measuring instruments. However, the effectiveness of STC depends on the parameters in principle.

The methods of estimation of these parameters developed by the author give methodological approaches to the clarification of the above questions.

The direct method of measurement in this study is to simulate the dynamics of the organizational environment to obtain digital artifacts. For this purpose, the author created personnel models, team building models, and STC productivity, models. The result of a multi-program experiment with these models is synthetic digital artifacts of the scientific organization: co-author, subjects, and directions of development.

The inverse method of the experiment analyzes the real digital artifacts of the STC. Namely, scientific articles, conference materials and from digital artifacts, the author builds a model of co-authorship, models of scientific topics, models of scientific directions and scientific schools in the organization.

4.2 The results of modeling of the process of publishing scientific articles

In this study, the publication activity of the scientific and technical center of Gazpromneft STC in the OnePetro electronic library of the Society of Petroleum Engineers (SPE)[1] was ana-

[1]SPE is a not-for-profit professional organization for oil and natural gas exploration and production (E&P) professionals. It was founded in

lyzed. The obtained dependence is shown in the figure (Fig. 4.1).

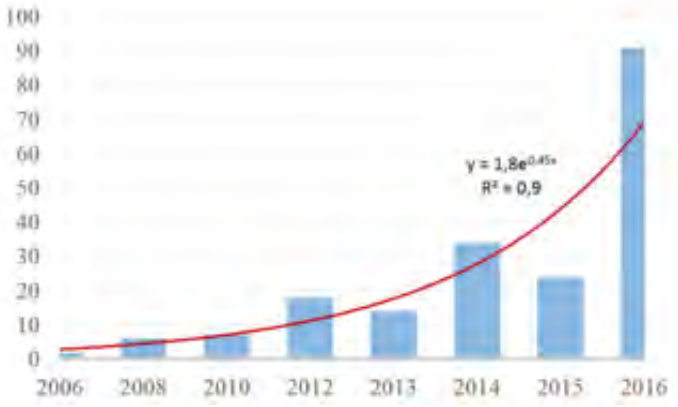


Figure 4.1 — The number of publications of employees of Gazpromneft STC.

The exponential growth of publications in one edition cannot continue indefinitely. Each edition has its limit of publications, manuscripts coming over the permissible volume of publications, increase competition for the right to be published. However, as a result of selection, some high-quality manuscripts are rejected by publishers. A simulation model was developed to study the publication process. The cognitive map of the publication process model is shown in the figure (Fig. 4.2).

The created model of the publishing process contains two stocks:

– Researchers,

1957, and today brings together more than 165,000 engineers, scientists, managers, and educators

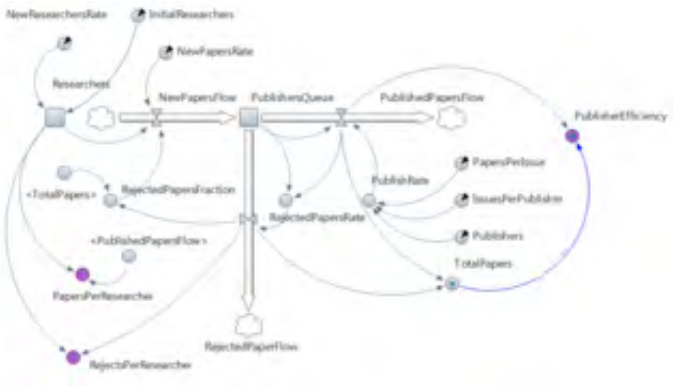


Figure 4.2 — Cognitive map of the publication process model.

– PublishersQueue – the stock of the manuscripts.

The model (4.2) is controlled by the following free parameters (Tab. 4):

Table 4 — Free parameters of the publishing process model.

Parameter Name	**Description**
Publishers	The number of publishers
PapersPerIssue	The number of articles in issue
IssuesPerPublisher	The number of issues per publisher per year
NewPapersRate	The speed of manuscripts creation
InitialResearchers	The initial number of researchers
NewResearchersRate	The rate of emergence of new researchers

A digital experiment was conducted by the cognitive map of the publication process model. Figure 4.3 presents the dependence of the efficiency of the publications from time to time with the different number of publishers.

The decline in the effectiveness of publications, as we can see, has a sharp, avalanche-like character. This type of publication efficiency behavior requires special attention in order not

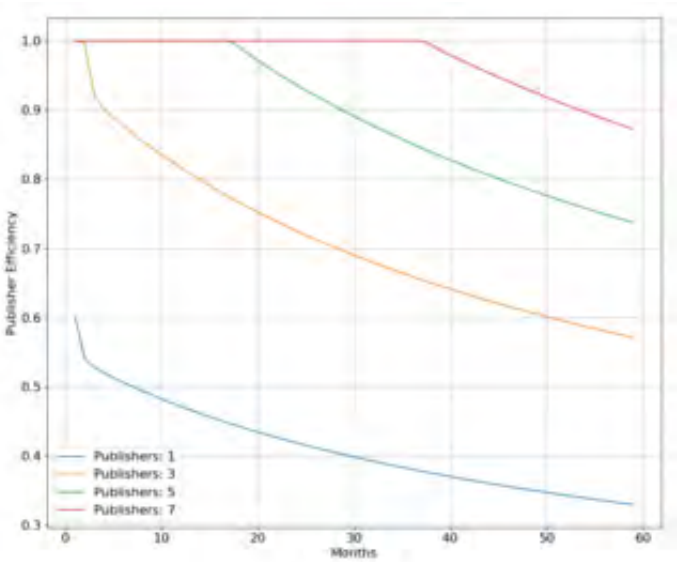


Figure 4.3 — The curve of the effectiveness of publications to time with a different number of publishers.

to miss the beginning of stagnation and to take organizational measures to increase the number of publishers involved in the publication process.

The principle of division of labor leads to an increase in the efficiency of processes. Hypothetically, extending the role model can improve the efficiency of the publishing process. Regardless of the number of co-authors, the publishing process defines the following roles:

- Producer – the carrier of the main idea of the research;
- Editor – changes the text of the manuscript;
- Reviewer – is responsible for the conclusions and results of the study;

- Translator – if the article is not in the authors' native language, technical translation and proofreading are required;
- Specialist in working with publishers – responsible for finding publishers and external communication;
- Designer – pictures, the presentation of the report;
- Speaker – presents the result orally at the conference (if needed and as many times as needed);
- Data Scientist – conducting computer calculations.

Given these roles, the research team does not change. The authors of the study are the scientists who conducted it. The quality of the manuscript getting better, and communication becomes more professional. Note that the functions of external and internal corporate communications are usually present in the organizational environment, but do not focus on the individual needs of researchers.

4.3 Measurement of the Intellectual Capital

Intellectual capital (IC) by its nature is a composite indicator of the productivity of a research organization whose main product is knowledge. The IC structure includes:

- Human capital
- Organizational capital

Human capital (HC) - includes knowledge and skills. Organizational capital (OC) includes technologies and processes. In other words, the HC characterizes the experience of employees, and the OC characterizes how employees apply their experience to the tasks in this organization.

In addition to creating intellectual capital, we can also consider its destruction – employees who conducted research leave and take with them knowledge.

Employees' contribution to intellectual capital is not equal. Next, the authors define the roles that belong to the "core team". The loss of employees at the core of the team dramatically impacts performance. The core includes employees with a high level of experience and the most popular skills in the organization.

There are quite a few approaches that describe the life cycle of an employee within an organization or position, but most studies agree on four main stages regarding the level of productivity:

1. the initial stage,
2. the accumulation of experience,
3. the productive stage,
4. the productivity decline.

At the same time, the stage of adaptation (initial stage and accumulation of experience) may differ depending on the type of activity and position level, but on average it takes up to six months for specialists and middle managers, about a year for senior managers.

The highest percentage of turnover among newcomers, so more attention should be paid to the social adaptation of new employees, the integration of newcomers in the processes and mentoring.

When a creative team breaks up, the brain drain can be different and doesn't always harm performance. In other words, sometimes the departure of an experienced, but having a different from the majority of the mental model of the employee, reduces the growth constraints of the IC.

There are two cases of staff turnover by employee's resignation and employer-initiated. From the IC point of view, both components have a negative impact. In Russian practice, there is a stable concept of staff turnover: an indicator that records the level of change in composition as a result of dismissal and transition to another job for personal reasons. The concept of turnover usually does not include the transition of an employee to another employer through transfer, which significantly distorts the Russian results compared to foreign ones. In different industries and industries, as well as at different levels of management, various values of staff turnover (from 2-5 to 80%) are considered the natural, which is due to the peculiarities of the business and categories of employees. For example, for retail and service sectors are characterized by very high rates of staff turnover, whereas for heavy industry, the standard relatively low value of staff turnover (5-10%). Typically, the level of staff turnover increases as the younger X-generations and Y-generations enter the labor market.

It is also important to note the relationship between burnout, fatigue and staff turnover, which has a negative impact on the productivity of the organization. The staff turnover has positive feedback reduced productivity. Organizations with high turnover typically have more problems with productivity and IC accumulation.

The most significant component of IC is the productivity of the organization, reflecting the ratio of effective staff to the total number of employees.

The author of this study has built a model of the IC based on the productivity of the organization. As an input for the IC model, a staff model was also built. The staff model usually solves the problem of predicting the number of personnel depending on the specific drivers of the population, usually an external (number of projects, tasks, customers, objects, services) from the current or specified productivity. The main problem of the staff models developed by the organizations, is a linear dependence of the personnel number of drivers and the lack of consideration of the factor of adaptation of the personnel (that is, the transition from beginners to experienced staff) as well as applicability only in a specific organization with its drivers and processes.

The objective of this experiment is to consider the behavior of the IC under load on the staff. A task execution model was created to assess changes in the IC under high load conditions. All three models are described next.

Figure 4.4 show the cognitive map of the staff model developed by the author of this study on the recommendations from [165].

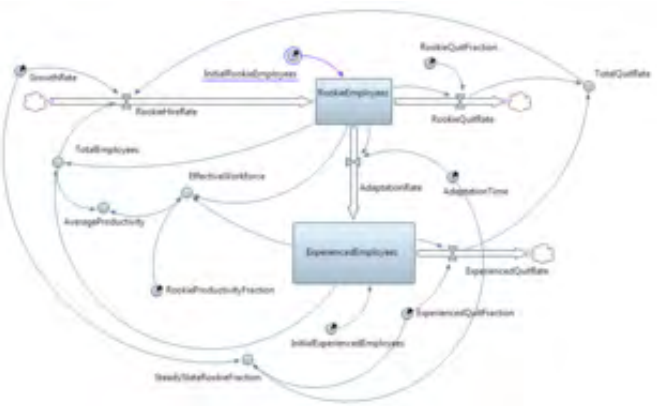


Figure 4.4 — The cognitive map of the staff model.

The staff model consists of two stocks - Rookie Employees (Beginners) and Experienced Employees and four flows:

– Rookie Hire Rate,
– Rookie Quit Rate,
– Adaptation Rate,
– Experienced Quit Rate.

Free parameters of the staff model are in te table 5:

Table 5 — The free parameters of the staff model.

Name	Description	Value
The speed of recruitment	Growth Rate	0.01 per week (from Total Employees)
The initial number of newcomers to the company	Initial Rookie Employees	40 people
The initial number of experienced employees in the company	Initial Experienced Employees	60 people
The time of adaptation of the newcomer to the experienced employee	AdaptationTime	50-100 weeks
The percentage contribution of the novice in staff productivity	Rookie Productivity Fraction	30-80%
The share of fired beginners	Rookie Quit Rate	0.01
Share of layoffs of experienced employees	experienced Quit Fraction	0.004

Dynamic variables of the staff model are in table 6:

Table 6 — The dynamic variables of the staff model.

Variable name	Formula
Total Employees	RookieEmployees + ExperiencedEmployees
Effective Workforce	ExperiencedEmployees + RookieProductivityFraction * RookieEmployees
Average Productivity	EffectiveWorkforce / TotalEmployees
Steady State Rookie Fraction	AdaptationTime * (ExperiencedQuitFraction + GrowthRate) / (1 + AdaptationTime * (ExperiencedQuitFraction + GrowthRate))
Total Quit Rate	RookieQuitRate + ExperiencedQuitRate

The flows listed in 4.3 calculated by the formulas from table 7:

Table 7 — The formulas for the staff model.

Name of the flow	Formula
Rookie Hire Rate	RookieHireRate = TotalQuitRate + TotalEmployees * GrowthRate
Rookie Quit Rate	RookieQuitRate = RookieEmployees * RookieQuitFraction
Adaptation Rate	AdaptationRate = RookieEmployees / AdaptationTime
Experienced Quit Rate	ExperiencedQuitRate = ExperiencedEmployees * ExperiencedQuitFraction

The staff model produce the input for the task model to the simulation of the workload. The figure 4.5 shows the cognitive map of the task model, developed by the authors of this study.

The task model consist of two stocks: ServiceBacklog (The task queue) and StandardTimePerTask (Standard time to complete the task). The table 8 shows the free parameters that manage the task model:

Table 8 — The free parameters of the task model.

Parameter name	Parameter value
Standard Workweek	40 hours
Target Delivery Delay	0.2 weeks
Initial Standard Time Per Task	1 person*hour/task
Minimum Delivery Delay	0.05 weeks

The table 9 shows the dynamic variables of the task model.

Figure 4.5 — The cognitive map of the task model.

Table 9 — The dynamic variables of the task model.

Variable name	Formula
Desired Workforce	DesiredCompletionRate * StandardTimePerTask / StandardWorkweek
Potential Completion Rate	EffectiveWorkforce * Workweek / TimePerTask
Time Per Task	StandardTimePerTask * EffectOfWorkPressureOnTimePerTask (WorkPressure)
Desired Completion Rate	ServiceBacklog / TargetDeliveryDelay
WorkPressure	DesiredWorkforce / EffectiveWorkforce
Workweek	StandardWorkweek * EffectOfWorkPressureOnWorkweek (WorkPressure)

Exogenous dynamic flows of new jobs (TaskArrivalRate) and the flow of completed jobs (TaskCompletionRate) control the task queue (ServiceBacklog). The equation 4.1 defines the equilibrium point for the task model.

$$EffectiveWorkforce = DesiredWorkforce \tag{4.1}$$

The staff model and the task model coupled via dynamic variables. Together, these models represent the dynamics of

skills and processes that characterize the intellectual capital of the organization, as we noted earlier. The table 10 shows the dynamic variables that couple two models.

Table 10 — The dynamic variables that couple the staff model and the task model.

Variable name	Action
Effective Workforce	Calculated in the staff model for the task model. Describes the number of employees for assignments.
Workweek	Calculated in the task model for the staff model. Describes the number of hours required to complete incoming tasks.

Lets examine the performance curves to identify the behavior of IC. The performance curve is a graphical representation of the change in learning rate of a particular activity. The figure 4.6 shows the performance curves for the staff model for various values of time of adaptation of newcomers. In the staff model, the dynamic variable that characterizes productivity is Average Productivity variable.

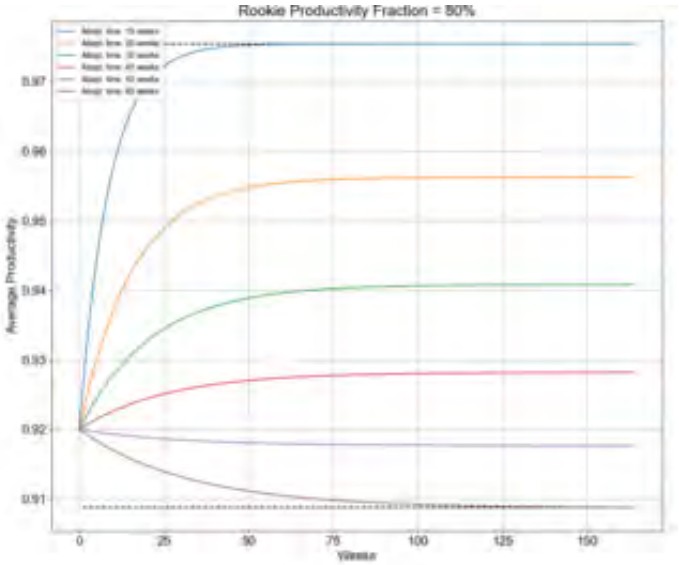


Figure 4.6 — The performance curves for different AdaptationTime.

As we can see from the figure 7 than Rookie Productivity Fraction is 80% and Adaptation Time is more than 50 weeks the learning curve goes to the lower asymptote, and at Adaptation Times over 60 weeks, the learning curve strives to the upper asymptote of the Average Productivity. Consequently, demonstrating a different nature of the performance behavior. In practice, this means that with a significant adaptation time of newcomers, organizational productivity falls, as the number of experienced employees in the team decreases regarding beginners, and the contribution to productivity from beginners is less than from experienced employees. On the other hand, the curves in the figure 4.7 shows that for the Adaptation

Time equal to 20 weeks, the productivity curves have a uniform character and differ in the rate of reaching the limit value – asymptote.

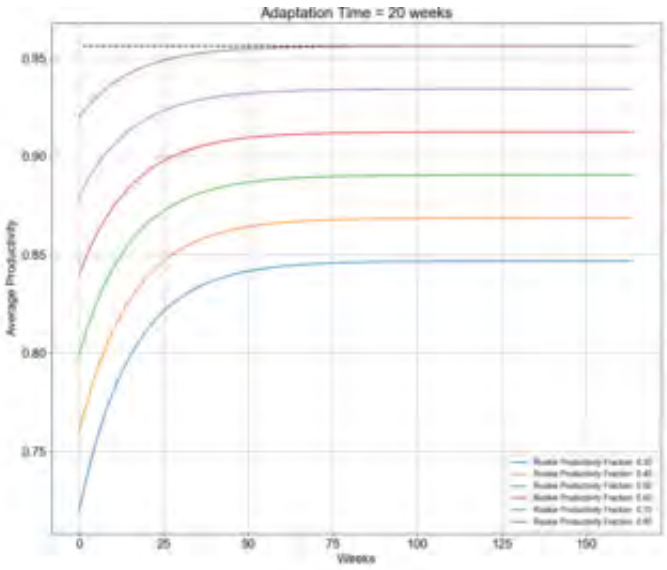


Figure 4.7 — The performance curves for different Rookie Productivity Fractions.

The small contribution of newcomers means that the complexity of the tasks does not imply the participation of untrained staff. On the other hand, large proportions of newcomer contributions mean that assignments allow even an inexperienced staff member to work with high returns approaching the returns of experienced staff.

The exogenous function was used to simulate different regimes of workload on the task model. The figure 4.8 shows the performance for different regimes.

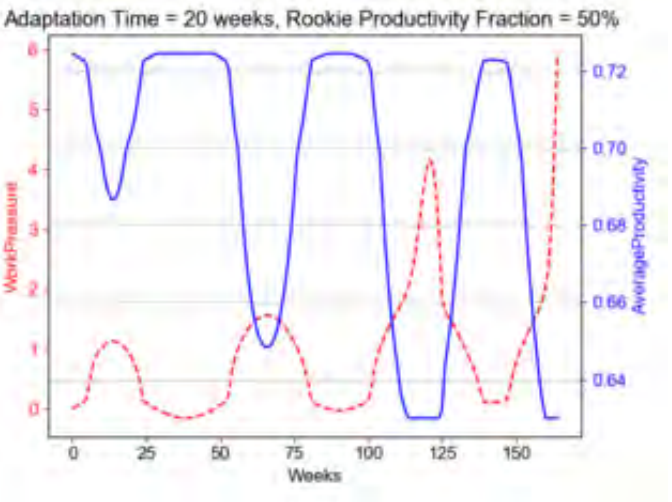


Figure 4.8 — The performance and Work Pressure curves for the IC model.

From the figure 4.8 we can see that in peak pressures the performance falls, but due to the adaptation of newcomers, the organization restores performance when the load falls. For different Adaptation Times in the IC model, the performance curves will have the form shown in the figure 4.3.

The figure 4.10 presents performance curves for an adaptation Time of 20 weeks, taking into account the load. We can observe different performance behavior before entering the asymptotes with different shares of newcomers, which reflects the fact that the possible inclusion of newcomers in the solution of tasks (before adaptation) characterizes these tasks as quite typical and straightforward.

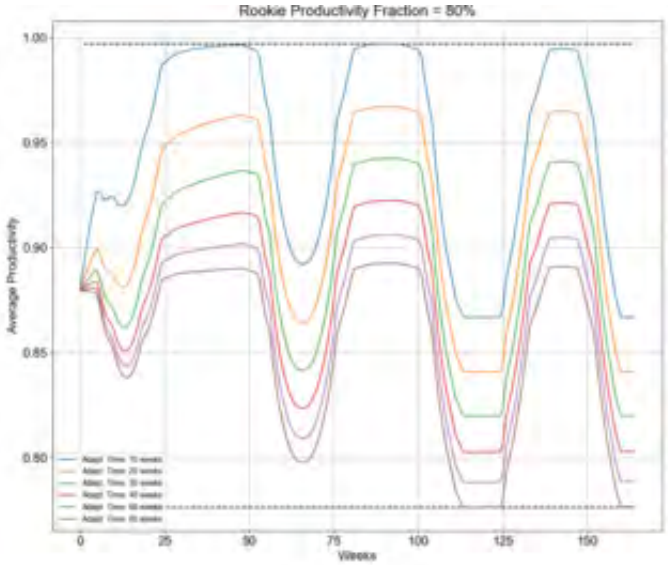


Figure 4.9 — The performance curves for the IC model with different Adaptation Times.

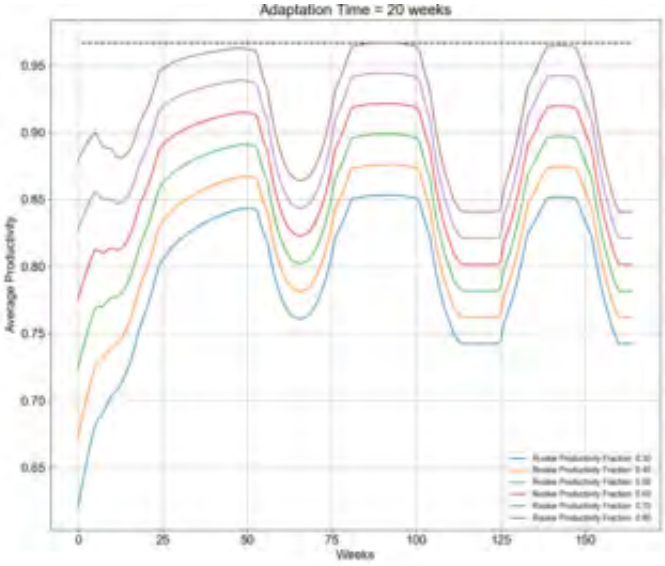


Figure 4.10 — The performance curves for different Rookie Productivity Fractions with pressure.

Note that with a short adaptation time of beginners (20 weeks) and a high share of newcomers in productivity (80%), the relative decline in productivity is lower than with a low high share of newcomers in productivity (30%). This observation confirms the fact that with the increasing pressure of short and simple tasks for beginners their productivity falls less than on complex tasks.

To simulate the effect of burnout and fatigue of employees in the conditions of long-term work in the mode of the extended week, the following dependencies are introduced in the IC model:

1. The effect of burnout is to increase the rate of dismissal of experienced employees, depending on the time of work in an extended week.
2. The effect of fatigue of employees is to reduce the productivity of employees depending on the time of work in an extended week.

The figure 4.11 shows the result of simulation of the IC model for 500 weeks. Such a long-term is chosen to show the effects of burnout and fatigue of the staff and as a consequence the decline in productivity caused by work in the extended working week.

The performance drop caused by prolonged high load has a dramatic effect on the IC. In conclusion, the figure 4.12 shows the curves of changes in human capital – experienced employees, beginners and the total number of employees. The curve of the required number of employees to perform incoming tasks

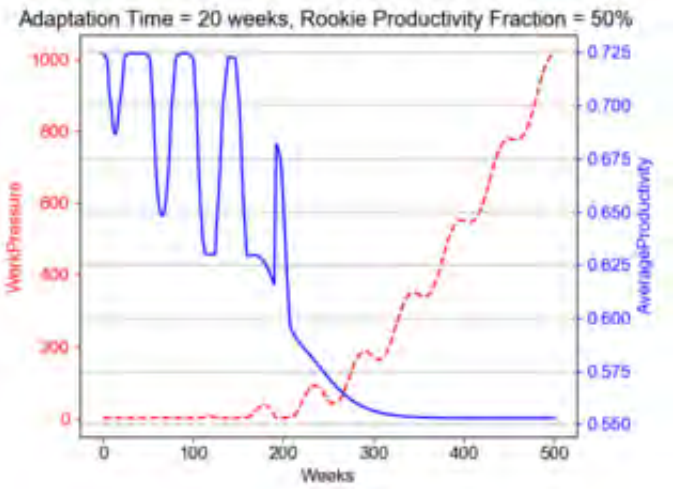


Figure 4.11 — The performance curves and pressure for the IC model with the extended working week.

is shown separately. We see that the number of newcomers is growing faster than the number of experienced employees.

The results of the experiment confirm the theoretical work on the study of the processes of intellectual capital management. The novelty of this study is to develop quantitative assessments that help to clarify the strategy of the intellectual capital management research organization.

The situation of workers in the conditions of high load considered by the author is typical for the Russian economy in modern conditions and is especially actual in oil and gas branch.

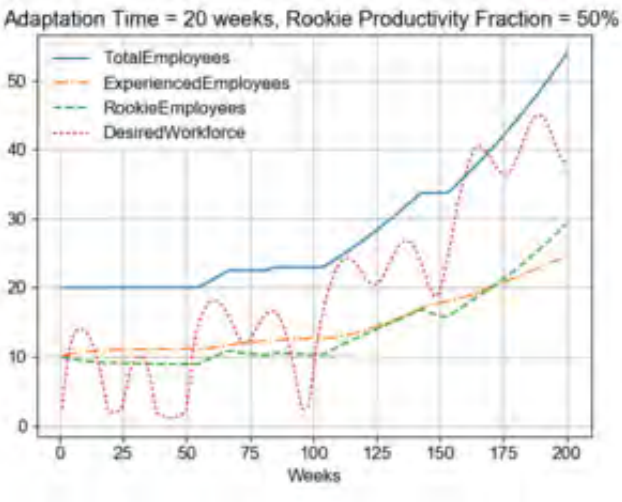


Figure 4.12 — The curves of changes in human capital.

4.4 The results of Team Building Modeling

For example, in industry research organization Ω laboratory work λ_i , where $i \in (1 \dots N_\lambda)$. We denote the set of laboratories $\Lambda = \{\lambda_i, \dots, \lambda_{N_\Lambda}\}$. The laboratories have researchers $a = \{a_i, \dots, a_{N_A}\}$.

We denote the set of topics t_i, where $I \in (1, \dots, N_T)$, by which the organization Ω leads the R&D as $T = \{t_1, \dots, t_{N_T}\}$. Then the activities of the organization Ω to perform R&D can be described by the following components (4.2):

$$\mathbb{M}_\Omega = \Big\{ S, \Xi, \Psi, E \Big\}, \text{ where } S = \{\Lambda, A, T, P, X\} \qquad (4.2)$$

In addition to the above defined components, there are in the 4.2 equation:

- $\Xi = \{\xi_1, \ldots, \xi_{N_\Xi}\}$ – connections between the subjects (team work, co-authorship, etc.),
- $\Psi = \{\psi_1, \ldots, \psi_{N_\Psi}\}$ – actions of the subjects (theme search, call for papers, etc.),
- $P = \{\rho_1, \ldots, \rho_{N_P}\}$ – researches,
- $X = \{\chi_1, \ldots, \chi_{N_X}\}$ – publishers.

Employees of the organization Ω perform research on topics T, create scientific articles and reports P for publication in journals and presentations at conferences X. When creating scientific articles P reviews of journals and conferences X are used. Conferences and editorial offices of journals X set priority topics T and accept manuscripts for publication on a specific schedule (abstracts, full text, reviewers' comments, presentation, publication) and from the most qualified and experienced researchers. The researcher has qualifications on the topics T, which can be represented as the n-dimensional vector $(c_1, \ldots, c_{N_T})$ and the experience of writing articles $(e_1, \ldots, e_{N_E})$, where c_i,$e_i \in \mathbb{R}$. Both qualifications and experience do not need to be normalized. Qualification grows with the successful implementation of research, and experience grows with the successful publication of articles on relevant topics.

Simulation modeling is a statistical experiment. Its results should be based on relevant statistical tests. The author chose a repetition method to compute confidence intervals and test hypotheses. Thus, each observation is presented as an inde-

pendent run of the model, in which the transition period is not taken into account. Then the average values of the sample are calculated. Since the runs are independent, a standard dispersion formula is applied. The advantage of this method is that each simulation run of the model is determined by its sequence of random numbers from the interval $(0, 1)$, which provides statistical independence of the obtained observations. The disadvantage is that initial transient conditions can strongly influence all observations.

GazpromNeft STC was taken as calibration organization for modeling. Six research topics $T = \{t_1, \ldots, t_{N_T}\}$, where $N_T = 6$, were chosen within the STC :

1. Development and exploitation of oil fields;
2. Geology and exploration;
3. Information technology in O&G;
4. Technologies of oil production;
5. Design of the field construction;
6. Drilling.

As the publisher of χ_1 selected edition of "Oil industry" produces the eponymous magazine since 1933. The authors chose the issue of the magazine for December 2016 (NX,12-2016), consisting entirely of articles by employees of Gazpromneft STC. The conference 16RPTC (SPE Russian Petroleum Technology Conference and Exhibition), held on October 24, 2016, in Moscow, was chosen as the conference χ_2. Thus, $x = \{\chi_1, \ldots, \chi_{N_X}\}$, where $N_x = 2$.

Currently, the analysis of social collaborations distinguishes two approaches:

- The structural approach focuses on the geometric shape of the network and the intensity of interactions (edge weight). In this case, Structural Theory and Network Exchange Theory are used to interpret the results;
- The dynamic approach focuses on changes in network structure over time.

The experiment at this stage aimed to confirm the adequacy of the structure of the components of the model $\mathbb{M}_{\Omega}$ on the example of the scientific and technical center of the oil and gas industry. When observing the visualization of agent behavior, the authors did not need to add new components to the model.

For the set conditions from $\mathbb{M}_{\Omega}$ a private model $\mathbb{M}_{STC}$ was created and a lot of running simulation of the model was carried out. One step of the simulation shown in the figure 4.13.

Moreover, by simulations, a database was created for the further study of the processes. The central database entities for the simulation of the $\mathbb{M}_{STC}$ model are shown in the figure 4.14.

Based on the simulation results, we obtained the following results (11) for the process of creating and publishing a scientific article.

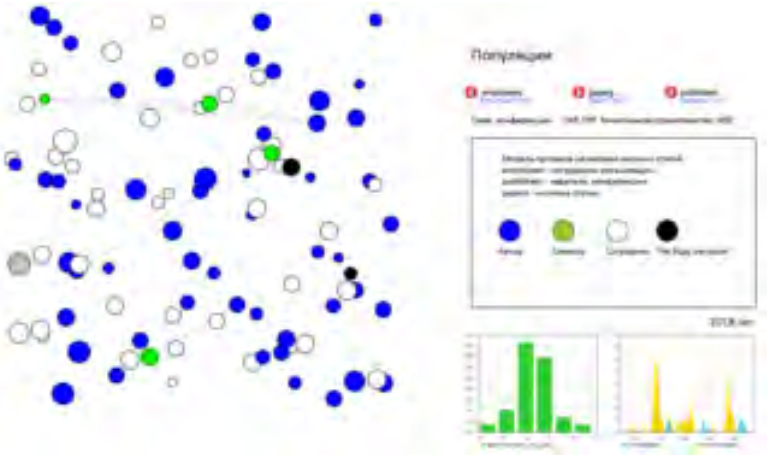


Figure 4.13 — One step of the simulation.

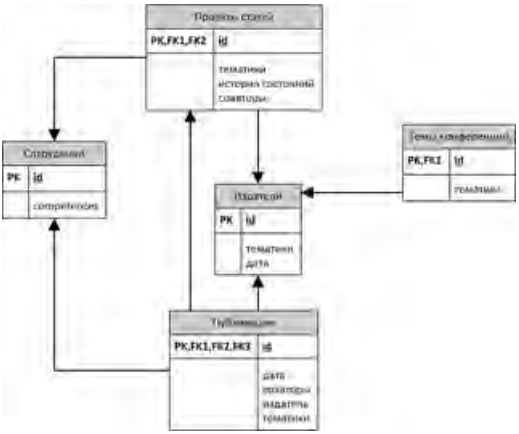


Figure 4.14 — The ERD for the simulation of the $\mathbb{M}_{STC}$ model.

Table 11 — The results of simulation of the $\mathbb{M}_{STC}$ model.

Variable name	Value
Average writing time	20 ± 2 weeks
Average number of co-authors	3.5 ± 1.0
Maximum number of co-authors	7 ± 3
Average number of articles per author per year	2 ± 0.5
The share of articles that did not meet the publication schedule	40 ± 10 %

The results are in agreement with the author's experience but need further verification. A joint bibliometric study of real empirical data was carried out using the author's method [166] to assess the correctness of the simulation results. Under the conditions of the experiment, a database of publications containing the following information fields was created:

1. Publication date;
2. Authors;
3. Title;
4. Key words according to the dictionary T;
5. Publisher according to the list X.

The database was consolidated from the articles of the publishing house "Oil industry" and publications from the electronic library of the community of oil and gas engineers SPE OnePetro. The analysis of publications is carried out, and the graph of co-authors is constructed 4.15.

Figure 4.15 — Co-authorship graph for Gazpromneft STC.

Based on the database of publications, the following parameters of the collaboration are calculated (12):

Table 12 — The results of the direct measurement of the STC activities.

Variable name	Value
Average number of published articles per author per year	2 ± 0.5
Average number of co-authors	2.8 ± 0.1
Maximum number of co-authors	10 ± 1

The obtained empirical results confirm the results of the simulation, which indicates the prospects for the use of simulation as an analog for modeling social processes in an organizational environment.

4.5 The result of optimization of the scientific activities

The task of finding the optimal parameters of the team of co-authors for the most productive contribution of scientific articles belongs to the class of optimization problems. The function that we want to minimize will depend on the following parameters:

- Number of employees in the organizational environment(N_o);
- The rate of appearance of new employees ($Vemp_{new}$);
- The rate of staff quiting ($Vemp_{fired}$);

- Maximum number of employee competencies ($Cmax_{emp}$);
- The maximum number of competencies necessary to achieve the goal of the study ($Cmax_{pub}$).

The vector of the optimum values consists of the following variables:

- The time to publish a scientific article (T_{pub});
- The fraction of employees who published articles ($Frac_{pub}$);
- The fraction of abandoned articles ($Frac_{notpub}$).

The following parameters represent the organization environment:

- Minimum and maximum number of employees in the organization (N_{omax},N_{omin});
- The rate of appearance of potential research targets (V_{pub});
- Time limits for writing an article (T_{eoc});
- Meeting speed ($V_{friending}$);
- The speed of a research target propagation (V_{go})

Based on the above parameters, the cost function $\mathcal{F}$ for optimization can be written in the following form (4.3).

$$\mathcal{F}\left\{\frac{1}{Frac_{pub}}, T_{pub}, Frac_{notpub}\right\} \rightarrow min \tag{4.3}$$

The equation 4.3 has the following constraints:

$$\begin{cases} N_o \in [N_{omin}, N_{omax}] \\ Cmax_{emp} \leqslant Cmax_{pub} \in [1, N_{comp}] \\ Vemp_{new} \geqslant Vemp_{fired} \geqslant 0 \end{cases} \tag{4.4}$$

The optimization experiment was conducted in AnyLogic environment for two models: with and without Scrum. Graphs of co-authorship stay the same. From the optimization experiment, the simulation of the co-authorship model developed by the author of this study was calibrated. The optimal parameters N_o, $Vemp_{new}$, $Vemp_{fired}$, $Cmax_{emp}$, $Cmax_{pub}$ were found for Gazpromneft STC. The optimal values of the parameters are given in Table 13.

Table 13 — Optimal values of the scientific activities.

Name	Value
Number of employees in the organization (N_o)	136
The rate of appearance of new employees ($Vemp_{new}$)	1 per week
Staff turnover ($Vemp_{fire}$)	1 employee per month
The maximum number of competencies of an employee ($Cmax_{emp}$)	4
The maximum number of competencies necessary to achieve the goal of the study ($Cmax_{pub}$)	5

The calibrated model became the basis for the study of the effect of the introduction of Scrum roles in the process of writing scientific articles. For the selected T_{pub} and $Frac_{notpub}$ was carried out many simulation runs of two types: using the methods of Scrum and without Scrum. The data analysis was performed in the statistical environment *Python*. The results of model runs for T_{pub} and $Frac_{notpub}$ shown in figures 4.16 and 4.17 respectively.

In both figures, the lines show the dependence of the moving average. To assess the impact of Scrum on the time of

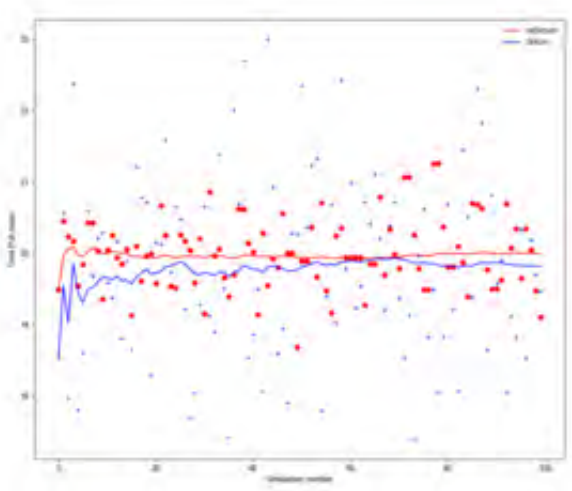


Figure 4.16 — The average time of publication of articles depending on the number of the run.

writing T_{pub}, the author compared the time of writing articles for two t-test samples to compare two independent samples. The results showed that at the level of 1% significance, the duration of writing articles using Scrum does not change.

- The average time to write a scientific article with Scrum was 19.90 weeks with a standard deviation of 3.33 weeks.
- The average time to write a scientific article without Scrum was 19.90 weeks with a standard deviation of 0.77 weeks.

The author also additionally used the nonparametric Mann – Whitney U-test in the case where the distribution of

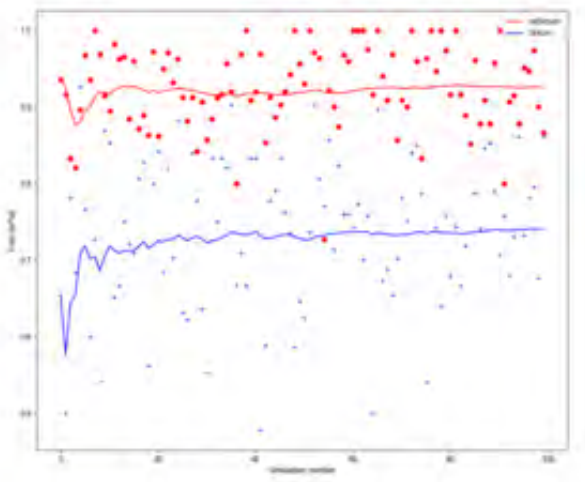


Figure 4.17 — The share of abandoned scientific articles depending on the run number.

features does not correspond to the normal distribution, the results of which were similar to the t-test. The results indicate that the use of Scrum does not accelerate the writing of articles, even if the function of writing articles is not subject to normal distribution.

Another indicator that can be used to measure Scrum productivity is the share of *abandoned* scientific papers. We estimated the proportion of *abandoned* scientific papers for teams that use Scrum and does not apply.

– The fraction of *abandoned* scientific papers in the Scrum teams is 0.74 with $\sigma = 0.02$;

– The fraction of *abandoned* scientific papers in the non-Scrum teams is 0.92 with $\sigma = 0.01$.

In other words, out of 100% of the articles started in teams using Scrum, 26% of the articles will be successful. In non-Scrum teams, only 8% of articles will be in the time frame of the publication process with the required quality.

4.6 Co-author Relationship Prediction

Collective co-authorship in writing scientific articles has deterministic and random structural components. In addition to the rational aspects in the formation of a team of co-authors of a particular scientific article, there are emotional components. In the temporary perspective, working groups of researchers are formed and break up the labor collective, and the composition of contractors who participate in joint industry collaborations for research are updated.

Despite the complexity of co-authorship, there are several classes of models for simulating co-authorship. Among them are models based on random graphs and models of co-authorship formation based on co-author competencies. Both mathematical tools have been developed and used for several decades separately. However, there are not so many practical applications of co-authorship models in corporate practice.

The author hypothesized that it is necessary to combine several different types of models in order to better understand the nature of scientific collaborations in a separate organization.

The author of this study set the task to develop a methodology for building a model of co-authorship for the scientific and technical center, taking into account the various structural components of co-authorship.

As a result, the author has developed a model using machine learning methods, random graphs, and competency models. From the developed model the forecast of development of co-authorship in writing scientific articles of Gazpromneft STC is made.

The practical value of the results of this study is as follows: Quantified the contribution of various structural components in the formation of co-authors in writing scientific articles.

Forecasting the development of co-authorship in writing scientific articles allows to carry out planning corporate resources to support the growth of scientific publications. Understanding the cluster structure of co-authorship allows aligning the areas of scientific activity under the strategic plan of the scientific and technical center.

Measurement of the activities of research organizations by the graph of co-authorship is a well-established practice. The researchers show opportunities to identify the highly productive authors and influential authors.

The publication activity of Gazpromneft STC was chosen as the object of research. The data were obtained from the onepetro open electronic library of the international community of petroleum engineers (SPE). After cleaning the data, 172 articles were left. The distribution of authors by year is shown in the figure 4.18.

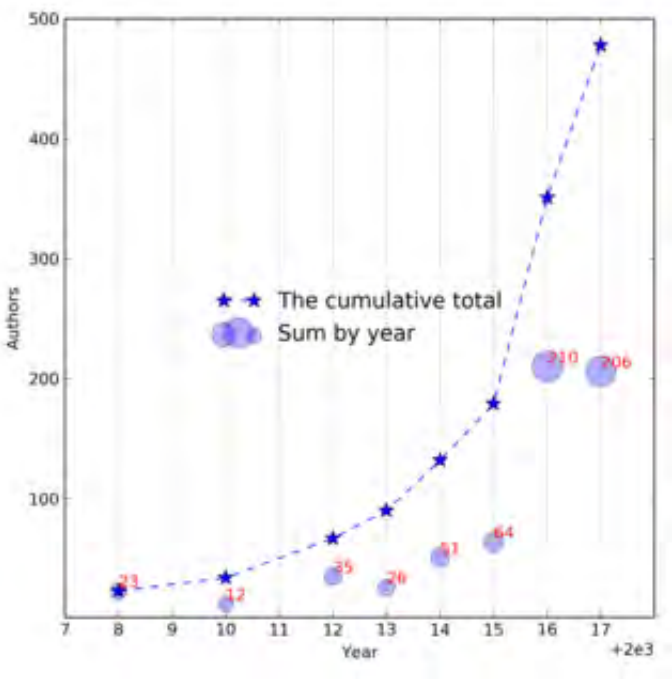


Figure 4.18 — Author allocation by year.

.

A direct answer to the placed research question can be an interpolation of the author quantity growth curve. We will receive the following dependency as a result of such assessment:

with authenticity giving a forecast of 585 authors in 2018. On the graph, we can also see that the number of authors in 2017 (207) is less than in 2016 (210), which could be a growth saturation and influence the forecast. Let us make a forecast based on the co-authorship graph. To do this, let us build a bipartite co-authorship graph [167] with node: author (479) and article (171). Authors possess technical competencies, while articles are characterized by their name, issue year and keywords.

The obtained co-authorship graph has 26 connected components, the biggest of which contains 556 nodes, while the rest have no more than eight. Small connected components relate to authors that wrote their first article. The existence of small connected components can be looked at as one of the components for co-authorship graph growth. The Table 14 shows the quantity and size of connected components for each year with an accumulating total.

Therefore, we can see that the co-authorship graph is progressing in quantity in the small connected components segment, while at the same time becoming more connected based on the number of nodes increasing in the main connected component. In order to increase the accuracy of forecasting it is expedient to take this structure into consideration.

Incremental co-authorship graph growth dynamic by year is depicted in Figure A.1. We specify that the co-authorship graph of 2017 is the sum of all depicted in Figure A.1.

From the figure Figure A.1 we can make a quality conclusion about the increase of quantity of annually added

Table 14 — The size of connected co-authorship graph components by year with an accumulating total.

Year	The size of connected components	Small components portion
2017	**556**, 8, 8, 8, 6, 5, 5, 5, 4, 4, 4, 4, 4, 3, 3, 3, 2, 2, 2, 2, 2, 2, 2, 2, 2	15%
2016	**367**, 8, 8, 8, 8, 8, 6, 5, 5, 5, 5, 4, 4, 4, 4, 3, 3, 3, 2, 2, 2, 2, 2, 2, 2, 2	23%
2015	**89**, 22, 21, 15, 12, 12, 8, 8, 8, 8, 6, 6, 5, 4, 3, 3, 2, 2, 2, 2, 2	63%
2014	**46**, 18, 15, 12, 12, 10, 8, 8, 8, 7, 6, 5, 4, 4, 3, 2, 2, 2, 2, 2	74%
2013	**23**, 15, 12, 11, 10, 8, 8, 7, 5, 4, 4, 4, 2, 2, 2	80%
2012	**15**, 14, 12, 11, 8, 8, 7, 4, 4, 4	83%
2010	**12**, 9, 8, 8, 4, 3	73%
2008	**12**, 8, 7, 3	60%

connections to the co-authorship graph. The change in growth leads to the co-authorship graph becoming more complex in 2016, which can be stated to be the "elbow effect" [168].

We will use the following graph node metrics in order to forecast authorship:

- Degree centrality
- Betweenness centrality
- Closeness centrality
- Harmonic centrality

– Clustering

The Figure 4.19 exposes the allocation of co-authorship graph node metrics.

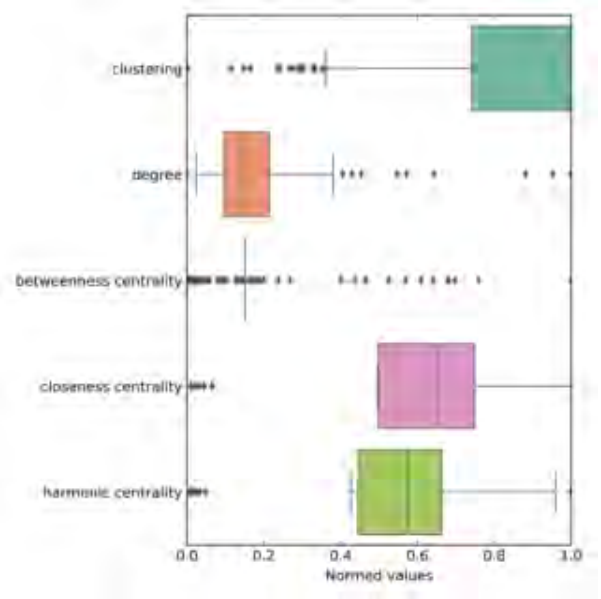


Figure 4.19 — Co-authorship graph node metrics.

We will use the binary classification model to forecast authorship. The model choice will be conducted based on the ROC-curve. Model learning will be conducted with the 2016 metrics. Model parameters are optimized with the help of cross-validation with a 5-fold folding. The following results were obtained as a result of comparing different classifiers (Table 15).

Multi-layer perceptron classifier based on a neural network with one layer consisting of 10 perceptions showed the best ROC AUC metric value.

Table 15 — Comparing classifier by the ROC AUC metric.

Model	ROC-AUC
KNeighborsClassifier	0.66
RidgeClassifier	0.73
RandomForestClassifier	0.72
SVM	0.70
Multi-layer perceptron	0.75

The execution report of the authorship forecast for 2018 based on the co-authorship graph of 2017 is presented in the Table 16.

Table 16 — Classification report of authorship forecast for 2018.

Labels	Precision	Recall	F1-score	Support
not author	0.80	0.98	0.88	66
author	0.80	0.20	0.32	20
Avg / Total	0.80	0.80	0.75	86

406 authors were predicted in 2018 as a result of forecasting. If we add employees who will write their first article in 2018 to the forecast, we will get an additional 15% based on

the assessment of the connected components growth dynamic. In total 467 employees will become authors in 2018.

4.7 Results of Clustering of R&D Trends

Planning of research and development trends in science and technology centers should be in line with the actual state of things. Such phenomena as organizational frigidity, research diversification and propensity for developing IT products are able to significantly impair any strategies and development trends.

However, feasibility of plans is an important attribute of development able to which significantly raise personnel's motivation for achieving best results. This is why setting achievable goals is of such importance. There are never enough quantitative tools for appraisal of research and development activities. Formal paperwork reporting on STC is not suitable for evaluation of researchers' involvement and dedication.

Instead, small formats of research works such as presentations at scientific and technical conferences or scientific articles in peer-reviewed scientific publications require much more informal approach from researchers. Analysis of a science and technology center's performance based on its publication activity is a common practice. Many studies analyze text corpus of scientific articles and make conclusions on development trends. Text data noise levels are quite high; even most advanced anal-

ysis methods based on word embedding are able to produce accurate predictions only if analyzed are huge text volumes which are seldom available in case of small organizations. Small research organizations suffer the most from inaccurate planning of research activities.

Author of this research propose to take advantages of articles (presentations) analysis based on co-authorship bipartite graph to extract research trends with the purpose of their further evaluation and planning.

From a formal point of view we have to solve the problem of unsupervised machine learning for co-authorship graph, attribute clusters to particular subjects and detect variations in clusters over the time.

Let us examine the conceptual meaning of the Betweenness centrality metrics applied to the problem of clustering of co-authorship graph in an STC. The Betweenness centrality metrics shows how important is a particular node for the graph's connectivity. Connections in a co-authorship graph reflect research collaboration. Co-authorship graphs are not always connected; usually they consist of several connected components of various sizes.

Connected components are natural clusters. Small connected components reflect primary initiatives – researchers' first articles. However the main connected component may contain up to 90% of a co authorship graph's nodes and call for a special approach to clustering.

To extract clusters from a main connected component of a co-authorship graph one may use the method of artificial removal of the nodes with the top-value Betweenness centrality metrics. As each of such nodes is removed a graph may break down into several disconnected components. The Figure 4.20 shows such separation model. On the left side of the figure depicted initially connected graph. On the right side of the picture the same graph with the node with the top-value Betweenness centrality metrics removed looks like two connected components.

Each of the components resulting from such separation can be analyzed for subjects homogeneity based on articles texts of which each component is formed. Several iterations would result in a set of clusters.

The method proposed by the authors is a heuristic one and requires examination by a particular formal criterion. Conventional criteria for the purposes of clustering are proximity metrics for a cluster components and distances between components in separate clusters.

Convergence of the authors' method is ensured through searching a minimum of functional errors in determining k clusters with equation 4.5.

$$\frac{WSS}{BSS} -> min \tag{4.5}$$

Where WSS_{c_i} – within-cluster variation for cluster C_i, m_i - centroid of C_i and $i \in [1..k]$ (4.6). The total WSS mea-

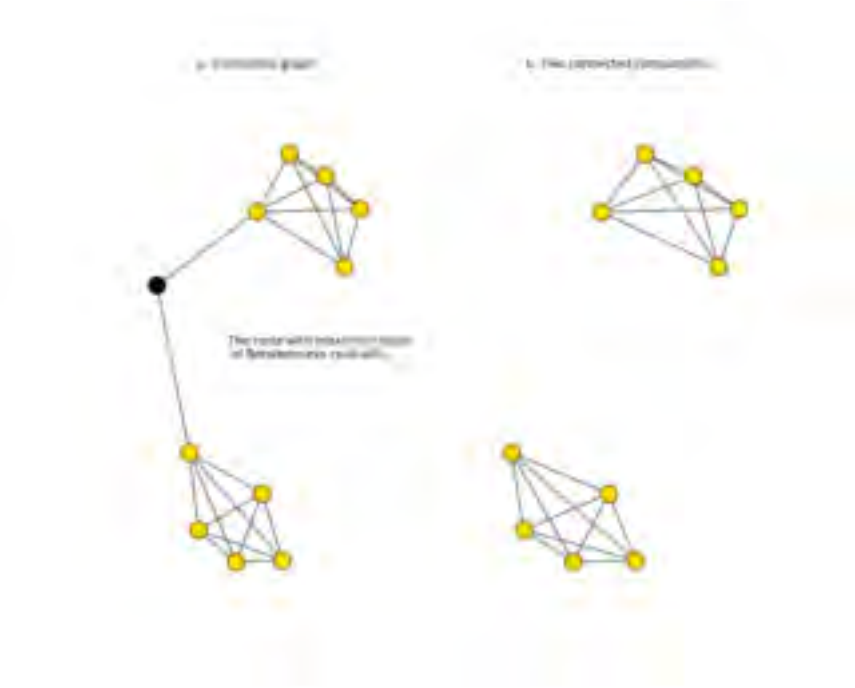


Figure 4.20 — Graph separation model

sures the compactness of the clustering and we want it to be as small as possible.

$$WSS = \sum_{i}^{k} \sum_{x \in C_i} \left(x - m_i\right)^2 \tag{4.6}$$

And *BSS* - weighted inter-cluster separation, measured by the between cluster sum of squares (4.7).

$$BSS = \sum_{j}^{k} \sum_{i}^{k} |C_i| \big(m_j - m_i\big)^2 \tag{4.7}$$

Where $|C_i|$ - is a cluster size.

Interdisciplinary researches lead to the situation where articles may fall into several subject categories, thus the resulting clusters would be intersecting and non-exclusive.

The Gazpromneft STC publication activity has been chosen as a research subject. The data has been obtained from the OnePetro open on-line library of the international Society of Petroleum Engineers (SPE). Upon cleansing 172 articles have been singled out.

Let us base our prediction on a co-authorship graph. For this purpose we build a co-authorship bipartite graph [167] with the nodes: author (479) and article (171). Authors have technical competences while articles have such attributes as title, year of publication and key words.

The resulting co-authorship graph has 26 connected components of which the strongest one has 556 nodes while the others have maximum eight nodes. Connected components with up to eight nodes represent the researchers' first articles.

Let us examine the strongest connected component (556 nodes). We extract a subgraph from the main co-authorship graph based on the nodes contained in the strongest connected component. The resulting subgraph is shown on the Figure 4.21 .

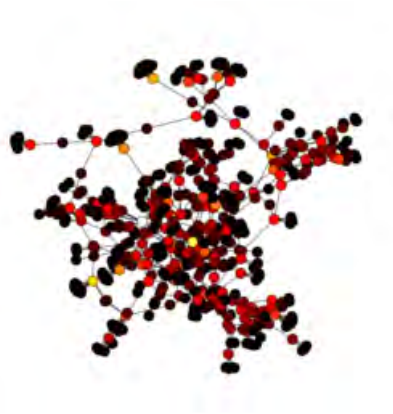


Figure 4.21 — Subgraph of the strongest connected component of the co-authorship graph of Gazpromneft STC.

Let us compute the Betweenness centrality metrics for the resulting subgraph. The obtained Betweenness centrality values are shown on the Figure A.2. Zero values for the Betweenness centrality are not shown.

As we can see on the Figure A.2 the values of the Betweenness centrality metrics in the third quartile belong to only 23 nodes which represent less than 5% of the total number of nodes.

Let us apply the algorithm of artificial removal of the nodes with the highest value of the Betweenness centrality metrics. The Figure A.3 shows correlation between the connected components number and the number of artificially removed nodes.

As the nodes get removed the graph can behave in two following modes:

1. Connectivity constraint (Mode I)
2. Exponential decay (Mode II)

Mode I is characterized by the graph's retaining its connectivity as the nodes with high values of the Betweenness centrality metrics get removed. It means that the removed nodes are not the only connections between clusters.

Mode II is characterized by following the exponential model of a graph's decay when each removed node causes exponential growth in emergence of new connected components.

Let us have a closer look at the second half of the Mode I of the algorithm when the graph has broken down into six connected components. These components' sizes are $511, 34, 1, 1, 1, 1$. Among them the component with 34 nodes shown on the Figure A.4 represents the most pronounced direction of research into *Subject 1*.

We have examined extraction of one cluster in detail. The complete algorithm of clusters extraction would consist of the following steps:

1. Building a co-authorship bipartite graph: G
2. Finding the Betweenness centrality metrics for the G graph
3. Finding a node with BC_{max} metrics (Betweenness centrality)
4. Removing the BC_{max} node (Betweenness centrality) from the G graph
5. Deriving a list of connected components of the G graph

6. Computing a quality metrics WSS and BSS of the retrieved clusters
7. Further the algorithm is iterated for each connected component
8. Algorithm is completed when all connected components represent clusters of acceptable quality.

For the selected co-authorship graph 16 clusters have been extracted. To compute values and W based on the articles texts we have applied the Vector Space Model (VSM). Each article is represented as a vector with the BM25 [169] metrics values for each word. Articles are considered as BOW ("bag of words"). For measuring distances between the articles' VSM we have applied cosine measure. The Figure 4.22 shows the clusters separability matrix. Clusters' numbers are on the axes. BSS function values are in the cells.

For the purposes of comparison of the resulting articles' clustering we have performed clustering with the KMeans algorithm which yielded similar results (Figure 4.23).

The articles corpus has been broken down into clusters using the KMeans algorithm. The resulting clusters allowed arranging authors into groups.

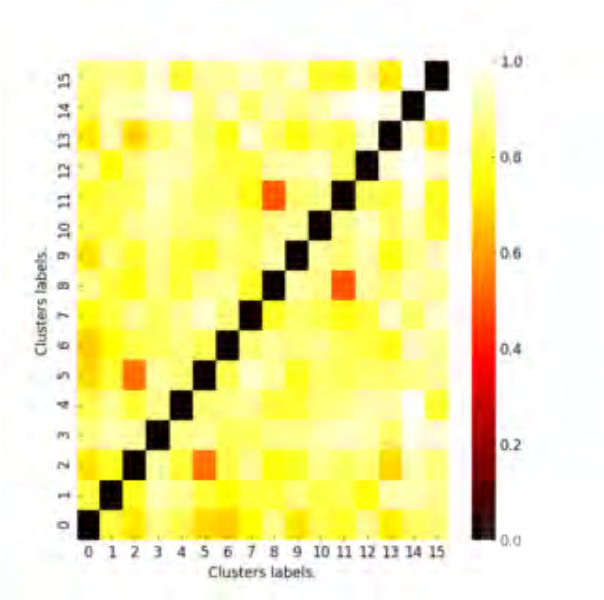


Figure 4.22 — Clusters separability matrix.

4.8 The probabilistic model of hidden topics based on the archive of the journal "Oil Industry".

The question of which way applied science and technology is moving is key to any scientific and technological field. Traditionally, the definition of development vectors was made and is made by experts in the subject area, but a significant increase in the volume of information and an increase in the number of development directions indicate the need to Refine and improve this tool and develop additional methods of research of

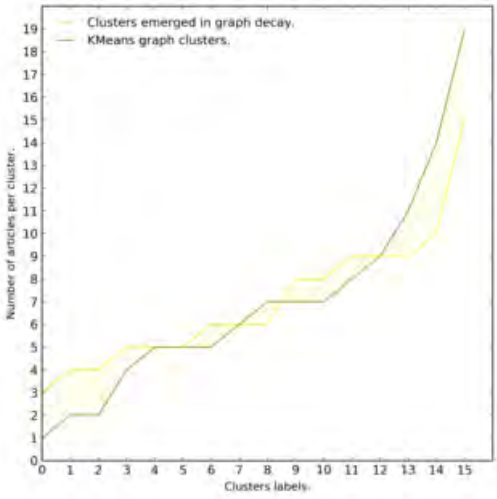


Figure 4.23 — Comparison of the clustering algorithm proposed in this article with the KMeans algorithm.

industry trends. In this study, the author analyzes the trends of the oil industry through automated text processing of scientific articles in the industry journal "Oil industry". Identifying the most common topics in the journal for the period from 2008 to 2016, we concluded that the importance of hard-to-recover reserves and the growing interest in the methods of development of such fields.

The journal "Oil industry" is devoted to oil and gas problems. It publishes articles on a wide range of issues in the oil and gas sector: economic, technical, technological, environmental and information. The publication dates back almost a century and has been published every month since 1920. All published articles are peer-reviewed. The journal is included

in The Russian science citation index (RSCI) and the international Scopus indexing system. Materials of the journal are in closed access and distributed by subscription. The main headings of the journal:

- oil and gas companies news;
- oil and gas industry;
- economics, management, law;
- geology and exploration;
- drilling;
- development and exploitation of oil fields;
- field design and construction;
- technology of oil production;
- oilfield equipment;
- transportation of oil;
- environmental and industrial safety;
- information technology.

For the study, the editors kindly provided the archives of the articles of the journal "Oil industry" for the period 2008-2016. The sample contains 108 issues of journals, with articles from 3517 authors. Thus, a complete sample was obtained, which contained materials in a variety of substantive areas. On average, in the issue of "Oil economy" about 20-25 articles. The issues of the journal were considered, as they were the unit of analysis. The authors of the journal's articles are researchers, engineers and industry experts, many of them are candidates and doctors of science.

The research process had the following steps:

1. Initially, the archives are presented as PDF files. Sometimes it was a single file (binder) with articles for the whole year, and sometimes different files with individual articles. In both cases, the files were intended for printing, that is, they contained a table of contents, page numbers, thematic inserts, and other editorial elements. For the analysis, only texts in the form of sentences were needed, so the author implemented a software module to bring all the data to such a format. It should be noted that, based on the chosen method, it was essential to keep the word order and division into sentences, while it was necessary to keep belonging to the issue, not to the article, as the minimum unit of interim analysis was chosen one issue.
2. At the second stage of the analysis, the words were reduced to the necessary forms. For the analysis and comparison of words by methods of frequency and probabilistic analysis, it is necessary to narrow possible variants of the use of word forms. There are several algorithms for solving text normalization task, in this case, stemming was used. Stemming is the procedure of finding the basis of the word, and the basis and the root of the word may differ from each other. One of the most common tools is Porter's stemmer, which, however, often trims the word more than necessary, making it difficult to get the correct word basis, such as "stolovaia"->"stol". Also, Porter's stemmer cannot

cope with all kinds of changes in the root of the word (for example, drop-down and fluent vowels), characteristic of the Russian language. Therefore, the author focused on the use of the technology of stemming company "Yandex" - *MyStem*. This program produces a morphological analysis of the text in Russian. It can build possible parses for words that are not included in the dictionary and offers several options for the basics of the word. However, the author found it necessary to maintain a reverse dictionary for the resulting word forms in order to maintain the relationship between the original word and the resulting word form. Abbreviations widely used in the oil and gas industry were treated as a separate branch. Definition of abbreviations was based on the dictionary of abbreviations established and supported in the company Gazprom-Neft in the framework of the project of Corporate Wikipedia [170].

3. At the third stage of the study, the formation of the dictionary was carried out. It is known that the most significant semantic weight are not single words, but combinations of words, in particular pairs of words – bigrams. The author used heuristic algorithms to distinguish bigrams. A matrix of bigrams in the neighborhood of five words for each of the sentences was compiled. Then we calculated the frequency of use of

each of the bigrams, after which were recorded 5% of the most common phrases.

4. In the fourth stage, the vocabularies of individual words and bigrams were combined for the overall treatment algorithms of allocation of subjects. The resulting dictionary was analyzed for the selection of high and low-frequency words for their filtering. Traditionally, the final generation of the dictionary is made with the help of stop words. The author did not use algorithms of stop words selection in this article. This decision was because the addition of a dictionary of stop words did not add accuracy to the study.
5. At the final fifth stage, a model of topics was built. The BigARTM [164] tool was used to build the topic model. At this stage, we obtained the matrix distribution of topics for the document (Θ) and the distribution of words for the topic (Φ). To improve the accuracy of the algorithm, the author applied an analytical approach that clarifies the regularization parameters.

The primary metric for detecting convergence of the topic model is the Perplexity metric. The curve of the dependence of Perplexity on the number of iterations on the body of texts is shown in figure 4.24.

From the figure 4.24, it can be seen that in three passes the topic model showed acceptable convergence and did not need further optimization.

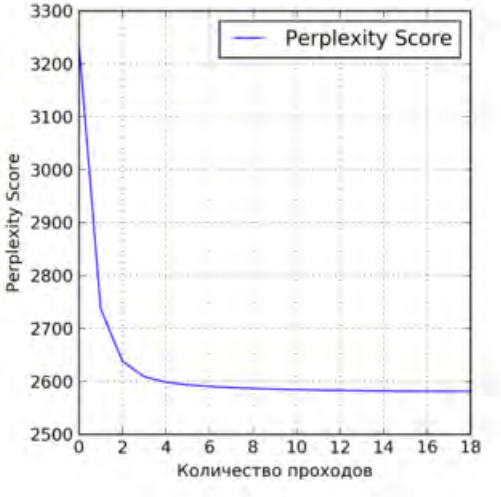


Figure 4.24 — The Perplexity score for the body of texts.

Essential quality metrics of the model are the degree of the sparseness of the matrices Φ and Θ. These metrics can be controlled by the parameters τ corresponding to the regularizers. In the figures 4.25 and 4.26 dependencies for the sparsity of Φ and Θ matrices are displayed.

Based on the dependencies shown in figures 4.25 and 4.26 the author chose the regularization parameters of the topic model, which allow achieving an optimal ratio between essential terms and noise.

The thematic model obtained as a result of this study can be presented in various forms. The level of noise terms makes it difficult to interpret the results, so there are six substantive topics from the planned twelve. The table 17 presents the topics calculated by the model.

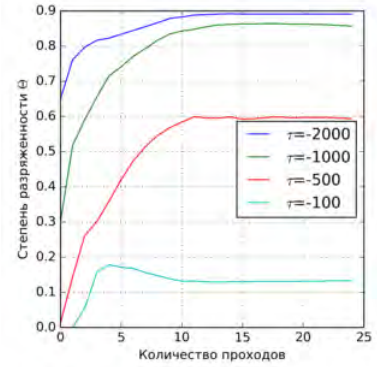


Figure 4.25 — The degree of the sparseness of Θ from τ dependence.

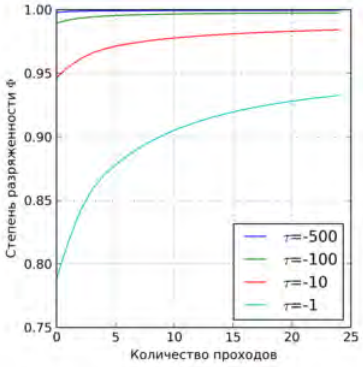


Figure 4.26 — The degree of the sparseness of Φ from τ dependence.

Table 17 — Fragment of the matrix Φ for terms with maximum probabilities.

Topic1	**Topic2**	**Topic3**	**Topic4**	**Topic5**	**Topic6**
электронный	ЭЦН	сдвиг	почва	нефтегазоносность	ингибитор
знание	УЭЦН	сигнал	добавка	свод	разлом
автоматизация	сероводород	окисление	композиция	компания	деформация
интегрировать	фациальный	разрушение	знание	впадина	трещиноватость
пользователь	гамма	деформация	агрегат	сепарация	исследовательский
архив	доломит	реологический	загрязнение	миграция	известняк
хранение	замер	песчаный	ПЗП	прогнозный	порода
доступ	депрессия	осадки	надежность	активность	политехнический
подразделение	агент	капиллярный	камень	филиал	штанга
платформа	каротаж	сечение	окисление	цемент	приемистость

It is safe to say that the terms collected in the column *Topic 1* characterize the subject of knowledge management. In column *Topic 2* are the subject of extraction. The other columns can also be quite unambiguously interpreted. Moreover, for machine processing, a set of terms is more important than its generalizing theme.

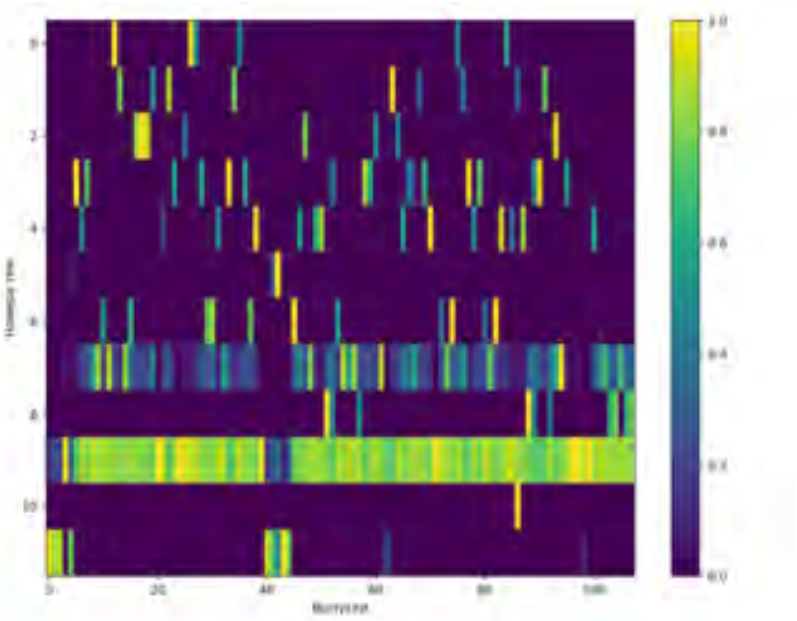


Figure 4.27 — Matrix Θ.

In the figure 4.27 the Θ matrix is presented, giving an idea of how the topics obtained are distributed in each of the analyzed releases. We can see that the theme with *Topic 9* is presented in all editions – it has shared information, greetings, and advertising. The resulting view also allows we to select the most relevant releases with a specific topic.

It is important to note that the method chosen by the author has shown a high speed of analysis, which makes it possible for use in online search processes. For example, on the website of the publishing house as a means of improving the search and giving recommendations to readers on articles with similar topics. It should also be noted that this technique can

be further improved and adapted for the analysis of significantly large amounts of dynamic data and highlight critical areas of technological development in both broader and narrower areas.

4.9 Topic Classification Through Topic Modeling with Additive Regularization for Collection of Scientific Papers

Topic Modeling is one of the current directions in statistical procession of natural languages. The Probabilistic Latent Semantic Analysis (PLSA) [171] of a collection of documents describes each topic through discrete probability distribution of words, while each document as a discrete distribution of topics.

Through this approach one can build an infinite number of various mathematical models for collections of documents. In practice for the purpose of Text Mining one needs "good" topic models: topics should be homogeneous and have unique content.

There are several dozens of various methods of building topic models. The most popular one is the method of building a topic model based on the Latent Dirichlet Allocation (LDA) suggested by D. Blei in 2003 according to the following formula:

$$p\left(d|w\right) = \sum_{t \in T} p\left(w|p\right)\ p\left(t|d\right) \tag{4.8}$$

, where T is a set of topics, $p(t)$ is an unknown distribution of topics over the entire collection, $p(d)$ is a prior documents distribution (estimation $\frac{n_d}{n}$), $p(w)$ is a prior distribution in a word set ($\frac{n_w}{n}$).

Based on the equation (4.8) we can introduce the following entities:

- $\theta_d = \left(p\left(t|d\right) : t \in T\right)$ – probability vectors of documents, $|T| \sim Dir\left(\theta, \alpha\right), \quad \alpha \in \mathbb{R}^{dim|T|}$,
- $\varphi_t = \left(p\left(w|t\right) : w \in W\right)$ – probability vectors of topics, $|W| \sim Dir\left(\varphi, \beta\right), \quad \beta \in \mathbb{R}^{dim|W|}$.

Practical experience shows that topical modeling based on the LDA model not always brings as well interpretable results as the ones mentioned in the original article [172]: topics contain "empty" words, it may be difficult to attribute the obtained topic to a particular class, the words contained in the topic most probably do not describe the topic.

Many practical researchers have noted that the LDA topic model does not overfitting itself as the PLSA does; however, from the Author's experience, this is not the case for the large collections of documents. It should be noted that the very concept of overfitting is not quite applicable to mathematical models, as the problem of generalization of a new text based on the topic model is not set in practice. As a rule practical researchers use ready sets of documents and solve a problem of getting information of this particular collection's hidden structure.

Topics extracted from collections of documents often contain noisy words and cannot be sufficiently interpreted, even by experts. The very need for experts for determine topics names makes automatic processing of documents collection very cumbersome.

Another issue in the topic modeling is that the obtained topics do not have any priorities or any other features which would make their interpretation easier. Such as a topic's degree of authenticity or its weight in a document. For this reason the obtained topics need additional heuristic modifications in order to make them "brushed up". Which suggests that such "brushing up" could be done even during the model training process. In other words, there are certain rules for obtaining θ_d and φ_t during the training process which may affect the quality of the obtained topics.

The essential rationale of the method described in this paper is setting aside the Dirichlet prior allocation and considering potential improving of the PLSA topic model in order to obtain more interpretable topics.

A series of researches like [173, 174, 175] support the presumption that the PLSA probabilistic topic model can be improved through additive regularization. Additive regularization of topic models (ARTM) is a multicriteria approach where a probabilistic topic model is optimized based on a weighted sum of criteria.

The paper [176] presents the method of developing an ARTM topic model which allows managing the training process

through adding to the authenticity function, at the training stage, such regulators such as:

$$R(\Phi, \Theta) = \sum_{t \in T} \sum_{w \in W} \beta_{wt} \ln(\varphi_{wt}) + \sum_{d \in D} \sum_{t \in T} \alpha_{td} \ln(\theta_{td}) \quad (4.9)$$

, where β_{wt} and α_{td} are regularization coefficients.

This approach to the management of the quality of probabilistic topic model has been proposed by Prof. K.V. Vorontsov.

The probabilistic topic model of text generation describes the process of documents collection formation through known distributions $p(w|t)$ and $p(t|d)$. The topic modeling problem is the reverse one: for a given collection D to find matrices of coefficients φ_{wt} and θtd with which the following condition is true:

$$p(w|d) \approx \sum_{t \in T} \varphi_{wt} \theta_{td} \quad (4.10)$$

The model training under the ARTM method is based on the reduced EM–algorithm which is used for iterated search of non-linear equations for φ_{wt} and θ_{td}.

Each topic's semantics is described as a φ_{wt} matrix. Therefore, to manage the quality of a topic model through regulators $\tau_i R_i(\Phi, \Theta)$ one can use various approaches in their sequenced combination so that each them would prepare the model for processing in the subsequent training iterations. Each approach can be described as a function setting values for τ_i coefficients depending on iteration step. However, how many iterations would be required for each approach, depends on

Figure 4.28 — The "Document-topic" space transformation.

such factors as the collections' vocabulary capacity and the sufficiency of documents.

Through the diversity of approaches $R_i(\Phi, \Theta)$ each topic's semantics φ_{wt} changes. On Figure 4.28 is shown how the "document-topic" space transforming in process of regularization.

The theory shows that it is possible to form the "black" and the "white" lists of topics through making matrices φ_{wt} and θ_{td} [177] smooth or sparse, determine "main" and "auxiliary" topics [178] as well as decorrelate topics by excluding repeated and linearly dependent topics [179]. However, experiments with particular collections of documents are still to support these theoretic assumptions.

It is important to note that in the process of training a convergence matrix will be the Perplexity calculated as follows:

$$P(D,\Phi,\Theta) = exp\left(\frac{-1}{n_d}\sum_{d\in D}\sum_{w\in d} n_{dw} \times \ln\left(\sum_{t\in T}\varphi_{wt}\theta_{td}\right)\right) \quad (4.11)$$

The *Perplexity* matrix is not normed and, therefore cannot be used in comparing various models convergence. The

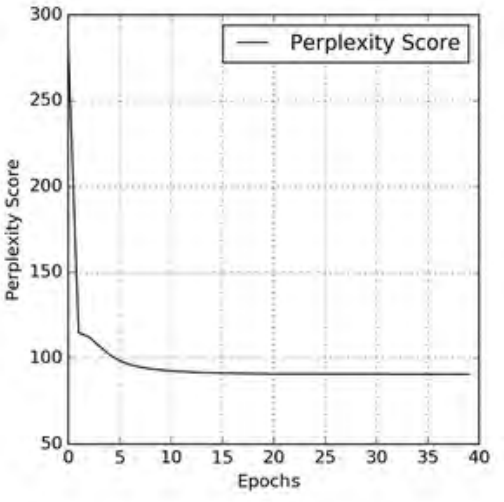


Figure 4.29 — Correlation of *Perplexity* to number of epochs.

general essence is, that the smaller *Perplexity*, the better a model. That is why in making decisions on sufficiency of convergence the criterion is whether *Perplexity* stops significantly decreasing as the number of training iterations grows.

The Author has used a collection of documents from the work [167], containing 1696 science and technology articles in English from the OnePetro e–library. All the articles in the collection are dedicated to oil and gas industry.

In the beginning, the Author has trained the PLSA topic model in order to establish the speed of convergence in respect to the *Perplexity* matrix. *Perplexity's* dependence on the number of the model training epochs is shown on Figure 4.29.

The dynamics of the *Perplexity* convergence shows that this model version gets converged pretty fast. From the 10th

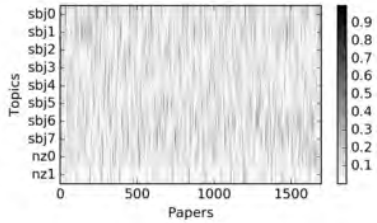


Figure 4.30 — The Θ matrix before regularization. Numbers of documents in the collection are marked on the x – axis.

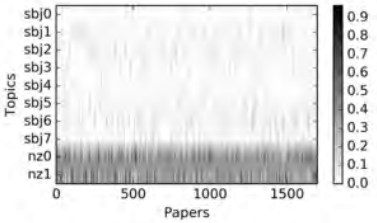


Figure 4.31 — The Θ matrix after regularization. Numbers of documents in the collection are marked on the x – axis.

iteration any further changes of the *Perplexity* metrics become insignificant.

In this experiment the Author's task was to divide topic into the basic and the auxiliary ones. From the perspective of regularization it means that the noisy topics would be smoothed while the basic ones get sparse in the process of the model's training. From the Equation (4.9) it follows that for smoothing coefficients β_{wt} and α_{td} have to be negative, and for sparsity – positive.

The obtained result can be presented in a graphical mode as a map of the matrix Θ coefficients. Let us consider the result of the model's distributing documents into particular topics. Distribution of topics for each document is shown in the Θ matrix. To obtain a more general idea of the θ matrix' transformation the author have represented it as charts before (Fig. 4.30) and after (Fig. 4.31) regularization.

Table 18 — Top10 terms forming auxiliary topics before and after regularization learning.

Before regularization		**After regularization**	
nz0	*nz1*	*nz0*	*nz1*
pump	wave	pump	stress
pipeline	seismic	sand	equation
esp	seg	completion	seismic
power	frequency	injection	wave
subsea	velocity	tubing	velocity
operating	waves	equipment	numerical
lift	amplitude	operating	elastic
equipment	x	downhole	array
installation	elastic	power	impedance
liquid	offshore	esp	direction

As we can see in the figure 4.30 and figure 4.31 the Θ matrix in the process of regularization becomes sparser in the basic topics ($sbj0 - sbj10$) and more compressed in the auxiliary topics ($nz0 - nz1$).

The vocabulary of the auxiliary topic is quite significant. The Table 18 below shows auxiliary topics before and after regularization.

Noteworthy is that the seismic ($nz1$) topic is among the auxiliary ones. According to an expert, such words as seismic, wave, velocity, elastic, seg, frequency and amplitude relate to

seismic. Articles on seismic are rare on the OnePetro platform and indeed can be considered secondary. After regularization learning several terms related to calculation appeared in the *nz1* but the seismic topic remained. In particular, the offshore term has gone to the basic topics. Situation with the *nz0* topic is quite clear. It includes the most frequently used words which are not actually noise.

Let us also pay attention to the quantitative analysis of the Θ matrix coefficients. For example the document No. 555 has the biggest topic weight 0.72 (*sbj6*). Probabilities of other primary topics for this document are close to zero. Thus, this document according to the model is fully focused on the sbj6 topic represented by words: recovery, injection, steam, core, viscosity, flooding, solvent, heavy, saturation, surfactant. With the help of an expert the *sbj6* was named "Chemical Enhanced Oil Recovery".

On the other hand, we can find in our sample of texts corpus that this document correlates to the article with the title "Low Tension Gas Process in High Salinity and Low Permeability Reservoirs". Here is an abstract from the public summary of this article on the OnePetro.org [2] :

[2]https://www.onepetro.org/conference-paper/SPE-179839-MS

Abstract
Chemical enhanced oil recovery (EOR) in carbonate reservoirs has always been technically and economically challenging. Conventional Alkaline-Surfactant-Polymer (ASP) flooding has limited application in low permeability (2-20 mD) and high salinity formations (200,000 ppm TDS) with a large concentration of divalent cations. Also injectivity into such low permeability reservoirs can be a significant problem with polymer solutions ....

As we can see from the public abstract from the article, its subject is highly specific. Moreover, based on the model we know that this article is indeed focused on this subject. When subscribing for this article one can be confident that it would not have other topics.

It is also worth noting that an expert's help was not essential for retrieving the *sbj6* topic's title; instead one could avail oneself of the fact that the document was represented by one single topic and find its title in the article's summary.

4.10 Analysis of Strong and Weak Ties in Oil&Gas Professional Community

The importance of weak social ties in professional communities is well studied and widely accepted. In our paper we analyze the structure of strong ties based on the co-authorship relation and use the formal concept analysis framework to figure out weak ties. The research is motivated by fast growing need in cross-disciplinary research, which requires experts from different areas to understand the bigger picture and identify potential fellows for collaborative research projects in nearest future.

The goal is to develop a methodology and tools for automated analysis of a collection of research papers available at the SPE digital library. On the basis of these analyzes one should be able to:

- figure out the most important and relevant research topics,
- assess the influence of different researchers and scientific schools,
- identify strong and weak ties in the professional community,

and use all of these in daily research management process. This paper is focused on the third item in the list. It continues our study of professional communities started in [180, 185, 184, 186, 183, 182, 181]

The analysis of social networks of co-authorship has a long history [187]. There are a plenty of studies examining the structure of co-authorship ties within diverse scientific fields and reveal specific collaboration patterns for the different disciplines [190, 189, 188, 192, 191]. Here we intend to uncover weak social ties in the Oil&Gas professional community. This is similar to the task of link prediction in social networks, see e.g. [193].

Weak ties within social networks is one of the key concepts. The idea of the differentiation of ties by their strength was firstly considered by sociologist Granovetter in [194], who empirically showed that weak ties (e.g. ties with not very close friends and relatives) are of a great importance in case of information propagation and knowledge diffusion. In case of Granovetter, weak ties were the source of the important information about working places and vacancies.

The identification of weak ties within a professional community has a great practical importance. Firstly, identification of people who are working on the same topic and substantial research idea is very important for information gathering and knowledge diffusion. Secondly, knowing the social environment, e.g. 'weak ties' within the community can be important in collaboration and cooperation establishment.

This study is based on materials of annual SPE Russian Oil and Gas Conference and Exhibition 2016. The main features of this event are as follows:

- Multi-disciplinary. The conference presentations, selected on the basic directions of development Oil & Gas industry. These areas are listed below.
- Periodic. This is an annual conference.
- Regional. The majority of the participants represented mainly the Russian companies.
- High selection criteria. The conference acceptance rate is approximately 15%. The selection process is conducted by Subject Matter Experts.
- The conference program consists of four parallel sections.
- At least one co-author must attend the event and present the work.

The data we work on is retrieved from open portal of Society of Petroleum Engineers (SPE) at `http://www.onepetro.org`.

Clean up and preparation of meta information was produced using Python on hybrid cluster at Gazpromneft STC. Text analysis was done using Python NLTK library. Statistical analysis was performed using SciPy library.

The collection comprises 404 articles written by 839 co-authors. It includes papers in the following areas:

1. Well construction – drilling and completion.
2. Static and dynamic modeling.
3. Hard-to-recover reserves.
4. Well and formation testing.
5. Field development monitoring and control.

6. Well intervention.
7. Shelf development experience and prospects.
8. Field geophysical survey/well logging.
9. Gas condensate and oil gas condensate field development.
10. Brownfields.
11. Geomechanics.
12. Oil and gas production - equipment and technologies.
13. Cores recovery, examination and analysis

In the retrieved data each publication record includes the following information:

– title and abstract of the article;
– the list of authors and their affiliations;
– year of publication.

The most time-consuming step was to prepare the data and make the data set clean and useful. Unfortunately, the portal does not have a directory for authors. As a result sometimes we had up to 6 different spellings of the same name in different articles.

Almost every paper in the collection is written jointly by a few authors. It usually takes at least several months to write a good paper, so in the context of professional community each publication could be considered as a proof of strong ties between the co-authors.

In figure A.5 nodes are authors, links correspond to the co-authorship relation. The graph has 839 nodes, 2315 ties, 127

components. The descriptive statistics for the co-authorship network is given below.

- Number of nodes: 839
- Number of strong ties: 2315
- Number of connected components: 127
- Size of the largest connected component: 198
- Size of the second largest connected component: 20

This and the other graphs in this paper are produced with yEd Graph Editor [195].

An inspection of the largest connected component shows that it mostly consists of participants of the well established collaborative program between Gazprom subsidiaries and Schlumberger. Otherwise the picture is very typical for a large industrial research conference, where the audience consists of big number of small cliques, which hardly communicate with each other.

As it was already mentioned above the goal of our work is to help the members of a professional community identify participants with similar interests and then convert weak ties into strong ones by establishing mutually beneficial collaborative research projects.

The importance of weak ties is well studied in the literature, see [194, 196]. In this paper we assume that two researches have weak ties if they both work with the same objects or concepts. We believe that if two persons work on the same substantial problem (e.g. they share same narrow research topic), they should at least know each others' works. We assume these

social ties are weak, because they are very much likely to know each other and even communicate, but the intensity of their interactions and communications is very much likely to be low, because they are not involved in joint projects.

The heuristics is implemented in the following way. First, we start from extracting keywords for each paper in order to create a formal context, i.a. object-attribute relation in which objects are words, attributes are authors, and the relation is «a keyword w is used by an author a». Second, the association rules with high characteristics of support and confidentiality are computed using Concept Explorer tool, see [198, 197].

Finally, for every association rule of the form:

$$a_1, \ldots, a_m \Rightarrow b_1, \ldots, b_k, \tag{4.12}$$

where $a_1, \ldots, a_m, b_1, \ldots, b_k$ are author IDs we assume that all members of the joint group $\{a_1, \ldots, a_m, b_1, \ldots, b_k\}$ are weakly connected.

As it was mentioned above our data set stores titles and abstracts of papers. As these texts are rather small we initially consider all words as equally important.

After the clean up the object-property table has 729 objects (keywords) and 839 attributes (authors).

On table 19 several examples of the association rules are presented. Each rule has two parts, antecedent and consequent, which are set of attributes. Support indicates the number of objects, which share these attributes. In our case, support is the number of keywords common for all authors in the set.

Table 19 — Examples of the computed association rules. Attributes are authors' IDs, support is the number of common keywords for these authors.

Support	Antecedent attributes	Confidence	Support	Consequent attributes
17	564;825	$= 94\% \Rightarrow$	16	133
16	335;636	$= 94\% \Rightarrow$	15	226;131;542;552
15	131;335	$= 100\% \Rightarrow$	15	226;542;552;636
16	101;436	$= 88\% \Rightarrow$	14	132
15	801;357;510	$= 93\% \Rightarrow$	14	8
15	333;754	$= 93\% \Rightarrow$	14	42;133
6	108;233	$= 83\% \Rightarrow$	5	754

In figure A.6 nodes are authors, co-authorship relation is represented by blue solid links, the dashed red edges correspond to the weak ties. Grey boxes set out previously disconnected fragments, which get bridged with the weak ties.

The main idea here is to interpret each association rule as an evidence of common interests for the involved authors. For example, from rule

$$15 \mid 333; 754 \ = 93\% \Rightarrow \ 14 \mid 42; 133$$

we conclude that members with IDs 333, 754, 42, and 133 work on the close subjects as they use 14 common keywords, so each two of them are considered weakly tied.

In general, each rule of a form

$$s \mid a_1; \ldots; a_n \ = c\% \Rightarrow \ s' \mid b_1; \ldots; b_m$$

produces C^2_{n+m} pairwise weak ties within the union set $\{a_1, \ldots, a_n, b_1, \ldots, b_m\}$.

For the data set of SPE papers the suggested procedure yielded the following. First, we have got 216 association rules with confidence greater than 80% and support at least 5 objects (keywords). Some of them are listed on 19. That resulted in 436 weak links out of which 149 were unique. Finally it turned out that the bigger part of them duplicates some of the existing strong ties and only 46 out of 149 suggest new connections. The network graph with the added weak ties is presented on figures A.6 and A.7.

Briefly, most of the isolated islands are not affected and remain isolated.Three cliques got connected to the largest component, see A.6. Another two joined the second largest component, see A.7.

The fact that out of 149 identified weak ties 103 are duplicates of the already established strong ties shows that the suggested heuristic is rather conservative, two thirds of the found connections are certainly relevant.

For the remaining new links we rely on expert opinion. For that visualization on figure A.8 was used together with the corresponding table of suggested candidate pairs for collaboration.

4.11 Deep analysis of publication texts

The use of artificial neural networks for text analysis has been developed in the mid-90s in the works of [199, 200, 201].

However, because of the high demands on computing resources for neural network training, it remained an academic discipline. Acceleration of research in this domain is connected with the growth of computing speed and the advent of new architectures of artificial neural networks like convolutional neural networks [202] and recurrent neural networks [203]. Learning neural networks has always required large amounts of labeled data. Moreover, with the advent of more layers with neurons, the need for labeled data has grown significantly. For example, to train an artificial neural network with one hundred thousand coefficients tens of thousands of marked texts are needed. So for the architecture of deep neural networks, the number of trained coefficients is millions [204, 205]. Therefore, training an artificial neural network on its data means allocating a certain amount of time and resources for labeling a dataset. In other words, each person classified object must be assigned to one of the classes "manually". Not so long ago, there were labeled text corpora with open access, for example, UMICH SI650 [3],

[3] https://www.kaggle.com/c/si650winter11

TreeBank [4], Twitter Sentiments [5], MPQA Opinion Corpus [6] and works on their analysis [206, 207, 208].

For this experiment, the author applied the method of *Transfer learning*. As a labeled dataset, the author selected reviews of movies [206]. This dataset contains 25 thousand positives and 25 thousand negative reviews. The dataset is thus balanced for training and validating the classification model. The length of reviews varies from 5 to 977 words and is shown in the figure 4.32.

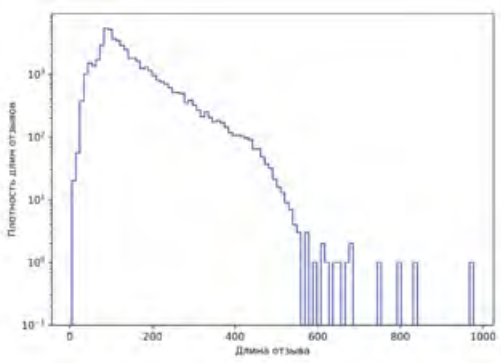


Figure 4.32 — The distribution of length of reviews.

In assembling the dictionary of reviews were discarded low-frequency words, that is, the words found in the documents are rare. The distribution of word frequencies by documents is shown in the figure 4.33.

[4] http://nlp.stanford.edu/sentiment/treebank.html
[5] http://www.sananalytics.com/lab/twitter-sentiment/
[6] http://mpqa.cs.pitt.edu

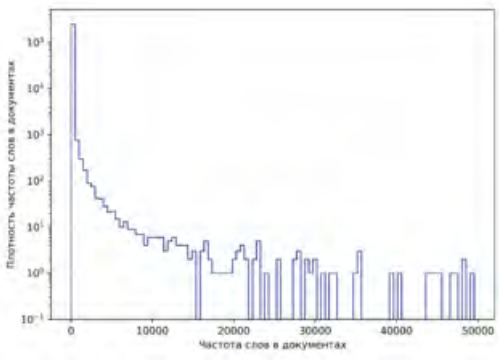


Figure 4.33 — Word frequency distribution by documents.

1696 scientific articles from the OnePetro were prepared as the dataset. To build a vector model of the text, we used the trained GloVe model. GloVe vectors with dimensions 100 and 300 were used. The advantage of using a trained vector text model is a significant reduction in the amount of computation. The number of trained parameters for creating a vector text model is several times greater than the number of parameters for the classification model architectures chosen by the author.

The author limited himself to a class of models based on artificial neural networks. Among the architectures of artificial neural networks used for text classification are CNN-LSTM and Stacked LSTM. The author chose the following three variants of model architectures for classification using artificial neural networks.

1. Recurrent neural network from a single LSTM layer. Further we will call this architecture RNN and sepa-

rately indicate the number of elements in the LSTM layer.

2. A convolutional neural network from a single layer Dropout-Conv1D-Conv1D-MaxPooling and a recurrent neural network from a single layer of LSTM elements. Further we will call this architecture CNN-LSTM and separately indicate the number of elements and the parameters of the convolutional layers.
3. Recurrent neural network of two LSTM layers. Further we will call this architecture RNN-2 and separately indicate the number of elements in the LSTM layer.

For the considered classification model architectures, the author chose the following essential hyperparameters:

– Model type: RNN, CNN-LSTM, RNN-2.
– Dictionary dimensionality. Depending on the filters of low-frequency words, the dimension of the dictionary changed from 2,000 to 200,000 words.
– VSM dimensionality: 100 and 300
– Fragment length: 80, 128 и 196.

The training of classification models was carried out in parallel on several servers with GPU. The dataset contained an equal amount of positive and negative reviews, so the Accuracy metric was chosen to assess the quality of learning. The cross-entropy function made the optimization of model parameters for classification. To accelerate learning, the author applied the method of early stopping based on the Accuracy metric on

the validation data set. Learning curves for the RNN type classification model are shown in the figures 4.34 and 4.35.

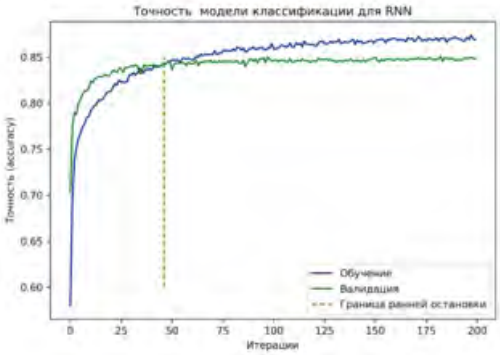


Figure 4.34 — The learning curves for the RNN model.

From the dependence 4.34 of the Accuracy metric for the training and validation dataset, it can be seen that in the area of the 42nd training iteration, the Accuracy metric stops increasing for the validation dataset. This phenomenon means that the model is starting to overfit and learning should be stopped. This classification model architecture does not allow for improved accuracy on this dataset.

It is clear that the value of cross-entropy in the validation data set does not begin to decrease in the area of the 37th iteration. That is, a little earlier than the Accuracy metric begins to degrade.

Based on the above techniques, classification models with different architectures and hyperparameters were trained. The learning outcomes are shown in the table 20.

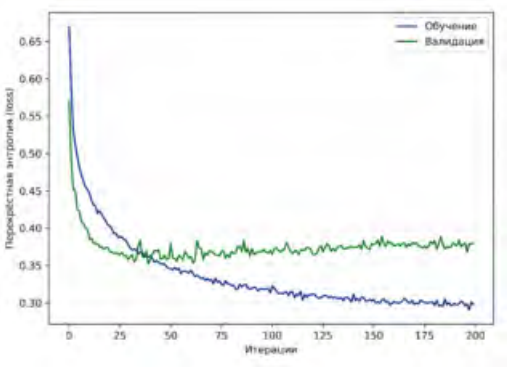


Figure 4.35 — Loss function for the RNN model.

Table 20 — The learning outcomes.

Model type	Number of parameters, thousand.	Fragment length	VSM dimensionality	Dictionary dimensionality	Validation accuracy
CNN+RNN	63	128	100	2 300	0,85
CNN+RNN	63	196	100	2 300	0,87
CNN+RNN	63	196	100	23 400	0,86
RNN	69	128	100	2 300	0,87
RNN	722	196	300	23 400	0,88
RNN	81	128	100	47 969	0,85
RNN-2	161	128	100	2 300	0,86
RNN-2	161	196	100	2 300	0,87
RNN-2	1 443	196	300	23 000	0,87
RNN-2	1 443	196	300	248 739	0,85
RNN-2	1 443	80	300	248 739	0,85

The best value of the Accuracy metric on the validation dataset showed the RNN model with a dictionary of 23 thousand words and a vector textual dimension of 300. Note that in the test dataset the Accuracy metric value for this model was 88%.

The resulting model was used to predict the tonality of scientific articles from OnePetro. Each scientific article was divided into fragments of a length of 196 words to assess emotional coloring. Then fragments of articles were collected back to get an emotional map of the entire article. Thus, it was possible to

identify fragments of an article with abnormal emotional colors, such as disappointment and satisfaction.

This study does not take into account the semantics of the text, so the subject of emotional coloring was not automatically determined. Selected fragments of the article must be analyzed with the help of an expert. However, such an approach to annotating an article allowed finding difficult-to-find fragments. The table 21 shows examples of emotional fragments of articles.

Table 21 — Identified emotional fragments of articles.

the results from pilot tests which were using as injectant are disappointing and the results from pilot tests which were using natural gases are encouraging
to sum up diffusion mechanism for in pilot tests had not been well recognized which in turn did not enhance oil production rate in those wells
the outstanding result from this study
using the other forward model result dramatically bad

The author also developed a color representation of the emotional coloring of articles depending on the probability of assigning a piece of text to a positive or negative emotional coloring.

On the x-axis, the ordinal number of the article is plotted; on the y-axis, the emotional coloring of the article fragments. The color scale displays a digital characteristic of emotionality: negative (-1), positive (+1).

In the picture, the emotionality of fragments of articles is displayed as a map. For each article on the X-axis, the emotionality of each fragment is shown sequentially along the Y-axis.

Scientific articles use academic vocabulary, and it would be naive to expect a degree of emotion in them comparable to reviews of movies. However, modern text-processing concepts based on the analysis of context allow us to isolate and classify

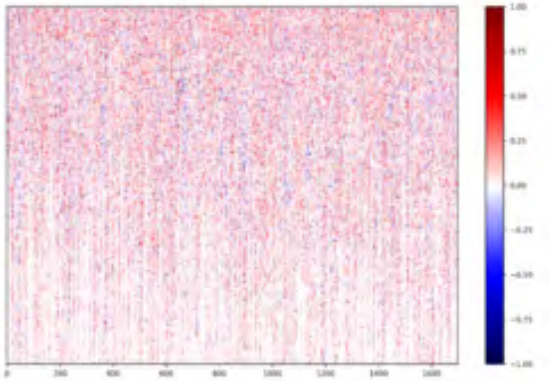


Figure 4.36 — The polarity map of the articles emotionality.

changes in emotionality precisely enough to process even scientific articles. The author believes that the conducted study opens up the possibility of creating additional tools for annotation and classification of scientific texts.

Chapter 5. Conclusions

In recent years, the question of the trajectory of the development of the oil and gas complex, as well as the entire energy system, is becoming increasingly appealing, both from experts and from the broad public [209, 210]. Several factors facilitate this situation.

- First, the pace of economic development leads to a significant increase in global energy consumption. As noted in the report of the Analytical Center under the Government of the Russian Federation, a significant increase in energy consumption occurs at the expense of developing countries, mainly in the Asia-Pacific region, while in developed countries the volume of electricity generation is stable, and consumption trends are similar to trends in general economic growth and recessions.
- Secondly, there is a change in the structure of hydrocarbon reserves. As noted in the "Energy Strategy of Russia for the Period until 2035" (formulated in 2015), the domestic oil industry is faced with such a problem as "an increase in production costs due to the prevalence of *hard to recover reserves* (HTRR) which complicates the retention of the achieved levels of oil production". At the same time, one of the tasks posed to the oil sector is the development of HTRR in vol-

umes up to 17% of the total production, which can be solved by developing mining technologies.

– Finally, thirdly, renewable energy sources play an increasing role in the energy sector, which affects the structure of energy markets. Experts, politicians, and citizens are increasingly concerned about environmental and climate challenges, which indicates the need for diversification of energy resources. Additionally, it is worth noting the negative impact of external economic and political restrictions on the raw materials sector of the Russian economy.

Thus, the energy sector is in the process of continuous transformation, and one of the critical issues on the agenda of the oil community is the optimization of geological exploration, production, and use of energy carriers.

There are several ways to analyze the trajectory of a change in scientific, technical and technological processes of oil production. The most obvious is a survey of experts specializing in mining.

Methods of expert surveys are widely used in various studies in which other forms of research are impossible or difficult to access due to the lack of objective data. Thus, the vast majority of foresight research is being implemented. The advantages of the expert survey include their relative simplicity and accessibility, as well as the possibility of using it in the absence of information about the phenomenon being studied.

At the same time, the apparent disadvantage of the expert survey is the possible subjectivity and limitations of the experts, their commitment to a certain point of view. As noted in the work of Bakhtin and co-authors [209], over the past years the volumes of expert-analytical and scientific literature, as well as information in general, have been overgrowing, so the task of receiving, filtering, processing and reflexive perception of all information becomes virtually impossible. At the same time, the expert needs to develop and improve in various substantive areas, which requires even greater labor and temporary investments. This situation demonstrates the need to develop and form additional feedback, which is designed to help the expert and professional community to analyze vast amounts of information and to extract the most relevant aspects from it, in particular, to identify technological trends.

With the development of automated methods for processing unstructured data, in particular, text data, thematic modeling of scientific texts is gaining popularity [69]. As was demonstrated in the work of Blea and Lafferty, thematic modeling turns out to be promising tools for tracking trends in such scientific fields as nuclear physics and neuroscience [211], technologies of the agro-industrial complex [209] , and so on. The study of automatically selected topics in the time perspective illustrates the changing interest of the scientific community to various objects and subjects of study. The advantage of this method is the possibility of automated processing of vast amounts of information and the identification of latent topics

of texts. At the same time, thematic modeling cannot be called an exclusively automated method, since the topics obtained as a result of machine classification should later be manually reviewed and worked out by subject-matter experts. Thus, thematic modeling can be considered as a method containing the advantages of both automated text processing and expert evaluation. The implementation of this method in the application to various substantive tasks will allow forming a dialogue between science and strategy at a fundamentally new level.

Topic modeling allows us to quickly process a significant amount of text to narrow the found concepts to small significant pieces of text - topics. Every topic seems to be a set of words, and a possible interpretation depends on the quality of this representation. The author showed the effectiveness of the approach to improving the interpretability of topics based on sequential regularization.

The applied methods of managing the relationship "density-sparseness" open up the possibility of setting the model on the subject area of the texts. The author showed the principles of creating and customizing a model of topics that allow intellectual search (exploration) of highly focused sources of knowledge. Clustering of topics was tested using two methods for vectorization of words (FastText, GloVe) and two methods for reducing the dimension of a vector space (TSNE, MDS). The results are presented in the form of diagrams and confidently show the presence of clusters. The approach to the analysis of textual information based on the modeling of topics is widely

used in the internal processes of GazpromNeft STC for optimizing knowledge management processes, identifying the most promising areas of research and finding opinion leaders in specific scientific areas. It is important to note that the method chosen by the author showed a high analysis rate, which makes it possible for use in online search processes. For example, on the publisher's site as a means of improving search and giving recommendations to readers on articles with similar topics. It should also be noted that the developed methodology can be further improved and adapted for the analysis of substantially large arrays of dynamic data and the identification of critical areas of technological development in both broader and narrower areas.

Existing forecasts of scientific and technological development (including foresight forecasts) for the most part extrapolate existing trends for the long term. Thus, considerable interest is acquired by works in which it becomes possible to identify new technological trends that can significantly alter the structure of markets. By themselves, individual technologies should not be considered as separate and isolated initiatives. Many technological trends are developing in parallel, which is the result of venture capital policy, technological development, and other related factors.

Because of this, an important area of analysis of technological trends is the study of the coevolution of the development of several technologies at once. It is the study of the totality of

scientific and technical initiatives that will allow a meaningful analysis of the direction of technological development.

In the section 4.3, the author presents a new perspective on the process of publishing scientific articles. Productivity indicators and strategies for managing the productivity of the publication process have been determined. The organizational environment should serve as a tool for increasing the efficiency of the primary production processes. The recognition by a research organization of the fact that the publication of scientific articles is one of the primary production processes means that it is necessary to create individual units aimed at supporting the effectiveness of this process. A measure of the maturity of the process is the degree of division of labor of its participants. A scientist should be engaged in his direct duties - research and is not obliged to delve into the details of the processes of arranging business trips, the ergonomics of presentations and the subtleties of communication with publishers.

The author has developed a role model, which will allow scientists to unload the scientists from the formal labor costs of publishing research results and in some cases avoid the appearance of "guest" co-authors.

Due to the limitation on the volume of publications in the publisher's edition, organizations need to expand the list of publishers publishing their researchers in order to maintain the growth rate of the number of published articles.

The productivity indicator expressing the share of articles rejected by the publisher is an essential characteristic of the pub-

lication of research results not only at the organizational but also at the industry level. The possibility of analyzing this indicator makes it possible to assess the sufficiency of the capacity of the market of scientific publishing houses and the degree of competition for publication in publications with a high impact factor.

In the section 4.5, the author summarizes and elaborates the formalization of the process of self-organizing teams to achieve a specific goal – writing scientific articles. The study developed a detailed algorithm for the formation of a graph of co-authors widely used in various studies. The basic theoretical statements are formulated and proved, definitions of the staffing of the team and the failed scientific article are given. The hypothesis about the invariance of the graph of the co-authors concerning the introduction of Scrum roles in the process of writing articles was formulated and confirmed. As a result of the optimization experiment conducted by the author, the optimal values of the parameters for the constructed article writing model were found. According to the results made on the optimized co-authorship model developed by the author, the effect of introducing flexible methods (Scrum) into the process of writing scientific articles by small teams of co-authors is as follows:

- The average time of writing a scientific article (T_{pub}) does not change;
- The average share of abandoned scientific articles ($Frac_{notpub}$) decreases.

The overall influence of Scrum on the process of writing scientific articles by a team of co-authors is positive. The productivity of teams formed on the complementary principle becomes higher from the use of flexible methods and Scrum, in particular.

In the experiment described in the section 4.11, a hypothesis was confirmed about the possibility of isolating emotionally-colored text fragments from scientific articles. Scientific articles use academic vocabulary, and it would be naive to expect a degree of emotion in them comparable to reviews of movies. However, modern text-processing concepts based on the analysis of context make it possible to isolate and classify changes in emotionality precisely enough to process even scientific articles. The author believes that the conducted study opens up the possibility of creating additional tools for annotation and classification of scientific texts. Recurrent neural networks showed the best quality assessment of the allocation of emotionally colored text fragments. Accuracy metric accuracy for them was 88%. It is important to note that concerning learning speed, recurrent networks permanently lose to convolutional networks. The author sees an explanation of the difference in performance in that a higher degree of parallel computing is possible for learning convolutional neural networks. Whereas for recurrent neural networks it is necessary to maintain a sequence of previous neuron states. In further studies, the author plans to explore the applicability of emotionally colored text fragments for the task of classifying texts as attributes. Also in the

opinion of the authors, the analysis of the syntax of emotionally colored text fragments is of scientific interest.

In the experiment described in the section 4.6, an analysis of the dynamics of the co-authorship graph for one organization was carried out from public data on publications. The bipartite graph of co-authorship, methodically substantiated by the author in the section 3.7.1, was chosen as the primary analytical tool. A multi-component approach to predicting changes in the properties of a graph of co-authorship has been applied. The analysis of small related components revealed their share in the annual increase in the number of authors. Note that the share of small components in the considered column of co-authors decreases with time, which is a structural limitation of the growth of the organization in question.

In 2016, the "Effect of the elbow" was discovered – a sharp complication of the nature of the growth of the co-authorship graph over the years. The author made a direct growth forecast based on the growing trend of the authors over the years and a revised forecast for the growth of the graph of co-authorship based on modeling using machine learning methods. A comparison of the accuracy of classifiers determined the classifier based on the neural network as the most accurate for this task. The forecast made from the model showed a result (467) significantly smaller than the result based on the trend (585). As a result of the study, the author concluded that there is essential information in the structure of the graph of co-authors on the development of the graph of co-authors, which determines

the growth forecast. That allows determining the significant signs of the formation of new collaborations, as well as the regression prediction of new links between the already formed research teams.

The use of vectorization methods for graph models in combination with feature extraction will improve the accuracy of predicting the emergence of new connections, as well as qualitatively measure publication activity based on publicly available metrics of journals and conferences. The author also proposed a method for identifying areas of scientific research based on a graph of co-authorship. The proposed method is related to top-down clustering algorithms. The *Betweenness centrality* metric was chosen as the criterion for the selection of clusters. As a criterion for checking the quality of clusters, the metric of the closeness of cluster members and the remoteness metric of various clusters based on the topics of scientific articles included in the graph of authorship were selected.

The result of the application of the proposed method is an enlarged vision of the scientific direction of the organization's development, made from public data on the publication activity of employees. The method developed by the author for identifying research directions by a graph of co-authorship was tested at the GazpromNeft STC. As a result, 16 clusters characterizing the activities of the organization were identified. Essential features of the developed method for identifying research directions based on the graph of co-authorship are as follows:

- The recursiveness of the algorithm allows working with graphs of various orders;
- The "greedy" algorithm for determining the quality of clusters allows you to adjust the optimization at each step;
- The use of bipartite graph construction co-authors allows us to analyze different projections;
- Research based on public data provides ample opportunities for use in business intelligence.

The novelty of the method of identifying research directions from a graph of co-authorship proposed by the author consists in the use of a duplex construction of a graph of co-authorship and a dynamic clustering model using structural metrics of the graph of co-authorship and proximity metrics of texts of scientific articles.

The results of the simulation experiment are consistent with empirical observations. Based on this, the author concluded that the work of researchers could be modeled using an agent-based approach.

The author has created the staff model in the organization and the model of tasks. Based on the interaction of these models, the author has built a model of productivity, which reflects the changes of IC for R&D organization.

The author has constructed dependences of IC on various terms of adaptation of beginners and various levels of workload, have shown the asymptotic behavior of IC that allows to model situations of different types of tasks, characteristics of

the organization: staff turnover, the speed of adaptation, the complexity of tasks and so forth.

The study analyzed how the IC is affected by the workload on the staff. It is shown how productivity decreases with time at high loads and the need to work longer than 40 hours a week. The author modeled the effects of "burnout" and "fatigue" of staff from the long-term high workload.

An experiment on multi-agent simulation was conducted, in which the agents were laboratory researchers interacting with each other and producing scientific articles as a result of their work.

In addition to the broad mathematical model of NTC $\mathbb{M}_{\Omega}$, the author created a private $\mathbb{M}_{GPN}$ model by calibrating the broad model on the data of NTC Gazpromneft.

To create a compact description of the computer experiment to predict the effectiveness of STC from a private model $\mathbb{M}_{GPN}$, the author used the *Anylogic* software. Based on the created private model, the author identified the following free parameters:

- Maximum number of co-authors
- Average number of co-authors in articles
- Distribution of the number of co-authors
- Number of articles per year per employee

Solving such a problem is an essential step towards identifying the mechanisms of collective work and the formation of a community high-tech product.

List of abbreviations and acronyms

W	A dictionary, a list of words.
D	A collection of documents.
n_d	Number of documents in collection.
n_w	Number of words in dictionary.
n_dw	Number of words in document.
w_i	Word with order number i.
d_i	Document with order number i.
T	A list of topics.
t_i	Topic from the list with order number i
Θ	A matrix "topics–documents" with dimension TxD.
θ_{td}	An element of matrix Θ for topic t and document d. A number in range from 0 to 1.
Φ	A matrix "words–topics" with dimension WxT.
φ_{wt}	An element of matrix Φ for word w and topic t. A number in range from 0 to 1.
$\mathbb{M}_\Omega$	A model of a social system.
MAP	A maximum a posteriori estimation.
MLE	A maximum likelihood estimation.
DC	Degree centrality
BC	Betweenness centrality
CC	Closeness centrality
HC	Harmonic centrality
CN	Common Neighbours

SI	Salton Index
JI	Jaccard Index
HPI	Hub Promoted Index
HDI	Hub Depressed Index
LHN1	Leicht-Holme-Newman Index
PA	Preferential Attachment Index
AAI	Adamic-Adar Index
RAI	Resource Allocation Index
ПКК	Full teams code
ОКК	Rest teams code
LDA	Latent Dirichlet Allocation
$\mathcal{KL}$	Kullback–Leibler divergence
TSNE	t-distributed Stochastic Neighbor Embedding
DBI	Davies Bouldin Index
cDBI	Cosine Davies Bouldin Index
GloVe	Global Vectors representation
FastText	A software library for text representations and text classifiers.

Glossary of Terms

Maximum probability of improvement (MPI) : $\mu(x) = P(\hat{f}(x) \geqslant f^* + \varepsilon = \Phi\left(\frac{\mathbb{E}\hat{f}(x) - f^* - \varepsilon}{Var[\hat{f}(x)]}\right)$

Upper confidence bound (UCB) : $\mu(x) = \mathbb{E}\hat{f}(x) + \eta Var[\hat{f}(x)]$

Expected improvement (EI) : $\mu(x) = \mathbb{E}\max(f(x) - f^*,0) = Var[\hat{(}f(x)] \cdot [z\Phi(z) + \varphi(z)]$, где $z = \frac{\mathbb{E}\hat{f}(x) - m(x)}{Var[\hat{f}(x)]}$

Scrum : Scrum is an agile process framework for managing complex knowledge work, with an initial emphasis on software development.

Failed scientific article (FSA) : An article which did not fit into the time frame of the publication process with the required quality.

Cosine measure of distance : $\frac{v_1 \cdot v_2}{\|v_1\|_2 * \|v_2\|_2}$

Euclidean measure of distance : $\|v_1 - v_2\|_2$

ROC AUC : A receiver operating characteristic curve, or ROC curve, is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied.

T4C : An author's technique for revealing the deep properties of text collections.

References

[1] Michael E Porter. — *Competitive advantage of nations: creating and sustaining superior performance.* — Vol. 2. — Simon and Schuster, 2011.

[2] Robert E Quinn and John Rohrbaugh. — "A spatial model of effectiveness criteria: Towards a competing values approach to organizational analysis". — In: *Management science* 29.3 (1983), p. 363—377.

[3] Richard A Wolfe. — "Organizational innovation: Review, critique and suggested research directions". — In: *Journal of management studies* 31.3 (1994), p. 405—431.

[4] Richard L Daft, Jonathan Murphy, and Hugh Willmott. — *Organization theory and design.* — Cengage learning EMEA, 2010.

[5] NV Lipchiu and KI Lipchiu. — "Metodologija nauchnogo issledovanija: uchebnoe posobie". — In: *Krasnodar: KubGAU* (2013).

[6] MA Mkrtchjan. — "Fazy perehodnogo perioda ot gruppovogo sposoba obuchenija k kollektivnomu". — In: *Kollektivnyj sposob obuchenija* 2 (1995), p. 8—11.

[7] NV Danilevskaja. — "Ocenka kak istochnik dinamiki tekstoobrazovanija v nauchnoj kommunikacii". — In: *Mezhdunarodnyj nauchno – issledovatel'skij zhurnal* 12 (2016), p. 27—30.

[8] Michael Shayne Gary and Robert E Wood. — "Unpacking mental models through laboratory experiments". — In: *System Dynamics Review* 32.2 (2016), p. 101—129.

[9] AV Sidorenkov. — "Gruppovaja splochennost' i neformal'nye podgruppy". — In: *Psihologicheskij zhurnal* 27.1 (2006), p. 44—53.

[10] Dedre Gentner and Albert L Stevens. — *Mental models*. — Psychology Press, 2014.

[11] Frederick W Taylor. — "Scientific management". — In: *The Sociological Review* 7.3 (1914), p. 266—269.

[12] Lev Davidovich Landau and Evgenii Mikhailovich Lifshitz. — *Course of theoretical physics*. — Elsevier, 2013.

[13] TS Koroleva, IA Vasil'ev, and IO Torzhkov. — "Kriterii ocenki jeffektivnosti dejatel'nosti nauchnyh uchrezhdenij". — In: *Trudy Sankt-Peterburgskogo nauchno-issledovatel'skogo instituta lesnogo hozjajstva* 2 (2014), p. 94.

[14] Nicholas S Vonortas. — "New directions for US science and technology policy: the view from the R&D assessment front". — In: *Science and Public Policy* 22.1 (1995), p. 19—28.

[15] Reinhilde Veugelers. — "Collaboration in R&D: an assessment of theoretical and empirical findings". — In: *De economist* 146.3 (1998), p. 419—443.

[16] Dries Faems, Bart Van Looy, and Koenraad Debackere. — "Interorganizational collaboration and innovation: Toward a portfolio approach". — In: *Journal of product innovation management* 22.3 (2005), p. 238—250.

[17] Thomas J Allen et al. — "Managing the flow of technology: Technology transfer and the dissemination of technological information within the R&D organization". — In: *MIT Press Books* 1 (1984).

[18] Raymond A Noe, John R Hollenbeck, Barry Gerhart, and Patrick M Wright. — *Human resource management.* — China People's University Press, 2006.

[19] Charles Horton Cooley. — *Social organization.* — Transaction Publishers, 1956.

[20] Stanley Wasserman and Katherine Faust. — *Social network analysis: Methods and applications.* — Vol. 8. — Cambridge university press, 1994.

[21] Bernhard Ganter, Gerd Stumme, and Rudolf Wille. — *Formal Concept Analysis: foundations and applications.* — Vol. 3626. — springer, 2005.

[22] Cynthia F Kurtz. — *Collective Network Analysis.* — 2009.

[23] Vaclav Snasel, Zdenek Horak, Jana Kocibova, and Ajith Abraham. — "Analyzing social networks using FCA: complexity aspects". — In: *Proceedings of the 2009*

IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology-Volume 03. — IEEE Computer Society. 2009, — P. 38—41.

[24] Dmitry Gnatyshak, Dmitry I Ignatov, Alexander Semenov, and Jonas Poelmans. — "Gaining insight in social networks with biclustering and triclustering". — In: *international conference on business informatics research.* — Springer. 2012, — P. 162—171.

[25] Sergei Kuznetsov, Sergei Obiedkov, and Camille Roth. — "Reducing the representation complexity of lattice-based taxonomies". — In: *International Conference on Conceptual Structures.* — Springer. 2007, — P. 241—254.

[26] Jonas Poelmans, Paul Elzinga, Dmitry I Ignatov, and Sergei O Kuznetsov. — "Semi-automated knowledge discovery: identifying and profiling human trafficking". — In: *International Journal of General Systems* 41.8 (2012), p. 774—804.

[27] Jonas Poelmans, Dmitry I Ignatov, Sergei O Kuznetsov, and Guido Dedene. — "Formal concept analysis in knowledge processing: A survey on applications". — In: *Expert systems with applications* 40.16 (2013), p. 6538—6560.

[28] Sergei Obiedkov and Camille Roth. — *Social Network Analysis and Conceptual Structures: Exploring Opportunities: Proceedings, Clermont-Ferrand, France, February 2007.* — Universite Blaise Pascal, Laboratoire Limos, 2007.

[29] Marie-Aude Aufaure and Benedicte Le Grand. — "Advances in FCA-based applications for social networks analysis". — In: *International Journal of Conceptual Structures and Smart Applications (IJCSSA)* 1.1 (2013), p. 73—89.

[30] Ruggero G Pensa and Jean-Francois Boulicaut. — "Towards fault-tolerant Formal Concept Analysis". — In: *Congress of the Italian Association for Artificial Intelligence*. — Springer. 2005, — P. 212—223.

[31] Frederick P Morgeson and David A Hofmann. — "The structure and function of collective constructs: Implications for multilevel research and theory development". — In: *Academy of management review* 24.2 (1999), p. 249—265.

[32] Richard Klimoski and Susan Mohammed. — "Team mental model: Construct or metaphor?" — In: *Journal of management* 20.2 (1994), p. 403—437.

[33] James P Walsh. — "Managerial and organizational cognition: Notes from a trip down memory lane". — In: *Organization science* 6.3 (1995), p. 280—321.

[34] Susan T Fiske and Shelley E Taylor. — *Social cognition: From brains to culture*. — Sage, 2013.

[35] Henry P Sims and Dennis A Gioia. — *The thinking organization*. — Jossey-Bass Inc Pub, 1986.

[36] Arthur P Brief and H Kirk Downey. — "Cognitive and organizational structures: A conceptual analysis of implicit organizing theories". — In: *Human relations* 36.12 (1983), p. 1065—1089.

[37] M Hall. — "Shale vs tight". — In: *Agile Geoscience* (2011).

[38] John E Mathieu, Tonia S Heffner, Gerald F Goodwin, Eduardo Salas, and Janis A Cannon-Bowers. — "The influence of shared mental models on team process and performance." — In: *Journal of applied psychology* 85.2 (2000), p. 273.

[39] Beng-Chong Lim and Katherine J Klein. — "Team mental models and team performance: A field study of the effects of team mental model similarity and accuracy". — In: *Journal of Organizational Behavior: The International Journal of Industrial, Occupational and Organizational Psychology and Behavior* 27.4 (2006), p. 403—418.

[40] Robert G Harper. — "Power, dominance, and nonverbal behavior: An overview". — In: *Power, dominance, and nonverbal behavior*. — Springer, 1985, — P. 29—48.

[41] Christian Steglich, Tom AB Snijders, and Michael Pearson. — "8. Dynamic Networks and Behavior: Separating Selection from Influence". — In: *Sociological methodology* 40.1 (2010), p. 329—393.

[42] Tom AB Snijders, Gerhard G Van de Bunt, and Christian EG Steglich. — "Introduction to stochastic actor-based models for network dynamics". — In: *Social networks* 32.1 (2010), p. 44—60.

[43] Miller McPherson, Lynn Smith-Lovin, and James M Cook. — "Birds of a feather: Homophily in social networks". — In: *Annual review of sociology* 27.1 (2001), p. 415—444.

[44] Per Block and Thomas Grund. — "Multidimensional homophily in friendship networks". — In: *Network Science* 2.2 (2014), p. 189—212.

[45] Joseph Weizenbaum. — "ELIZA—a computer program for the study of natural language communication between man and machine". — In: *Communications of the ACM* 9.1 (1966), p. 36—45.

[46] Henry Kucera and Winthrop Nelson Francis. — *Computational analysis of present – day American English.* — Dartmouth Publishing Group, 1967.

[47] Stephen Cole Kleene. — *Representation of events in nerve nets and finite automata.* — Tech. rep. — RAND PROJECT AIR FORCE SANTA MONICA CA, 1951.

[48] Ken Thompson. — "Programming techniques: Regular expression search algorithm". — In: *Communications of the ACM* 11.6 (1968), p. 419—422.

[49] Ben Eisner, Tim Rocktaschel, Isabelle Augenstein, Matko Bosnjak, and Sebastian Riedel. — "emoji2vec: Learning emoji representations from their description". — In: *arXiv preprint arXiv:1609.08359* (2016).

[50] Brendan O'Connor, Ramnath Balasubramanyan, Bryan R Routledge, Noah A Smith, et al. — "From tweets to polls: Linking text sentiment to public opinion time series." — In: *Icwsm* 11.122-129 (2010), p. 1—2.

[51] Joachim Bingel and Anders Sgaard. — "Identifying beneficial task relations for multi-task learning in deep neural networks". — In: *arXiv preprint arXiv:1702.08303* (2017).

[52] Rahul Jha, Amjad-Abu Jbara, Vahed Qazvinian, and Dragomir R Radev. — "NLP-driven citation analysis for scientometrics". — In: *Natural Language Engineering* 23.1 (2017), p. 93—130.

[53] Julie Beth Lovins. — "Development of a stemming algorithm". — In: *Mech. Translat. & Comp. Linguistics* 11.2 (1968), p. 22—31.

[54] Ilya Segalovich. — "A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine." — In: *MLMTA*. — Citeseer. 2003, — P. 273—280.

[55] Serge Sharoff and Joakim Nivre. — "The proper place of men and machines in language technology: Processing Russian without any linguistic knowledge". — In: *Proc.*

Dialogue 2011, Russian Conference on Computational Linguistics. — 2011.

[56] Mikhail Korobov. — "Morphological analyzer and generator for Russian and Ukrainian languages". — In: *International Conference on Analysis of Images, Social Networks and Texts.* — Springer. 2015, — P. 320—332.

[57] Peter Willett. — "The Porter stemming algorithm: then and now". — In: *Program* 40.3 (2006), p. 219—223.

[58] Martin F Porter. — *Snowball: A language for stemming algorithms.* — 2001.

[59] David Packard. — "Computer–assisted morphological analysis of ancient Greek". — In: *COLING 1973 Volume 2: Computational And Mathematical Linguistics: Proceedings of the International Conference on Computational Linguistics.* — Vol. 2. — 1973.

[60] Steven Bird, Ewan Klein, and Edward Loper. — *Natural language processing with Python: analyzing text with the natural language toolkit.* — " O'Reilly Media, Inc.", 2009.

[61] Holger Schwenk and Jean-Luc Gauvain. — "Connectionist language modeling for large vocabulary continuous speech recognition". — In: *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on.* — Vol. 1. — IEEE. 2002, — P. I—765.

[62] William J Teahan and John G Cleary. — "The entropy of English using PPM-based models". — In: *dcc*. — IEEE. 1996, — P. 53.

[63] WJ Teahan and John G Cleary. — "Models of English text". — In: *Data Compression Conference, 1997. DCC'97. Proceedings*. — IEEE. 1997, — P. 12—21.

[64] Lalit R Bahl, Frederick Jelinek, and Robert L Mercer. — "A maximum likelihood approach to continuous speech recognition". — In: *IEEE transactions on pattern analysis and machine intelligence* 2 (1983), p. 179—190.

[65] Lalit Bahl, Peter Brown, Peter De Souza, and Robert Mercer. — "Maximum mutual information estimation of hidden Markov model parameters for speech recognition". — In: *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'86.* — Vol. 11. — IEEE. 1986, — P. 49—52.

[66] Ar Averbuch, L Bahl, R Bakis, P Brown, G Daggett, S Das, K Davies, S De Gennaro, P De Souza, E Epstein, et al. — "Experiments with the TANGORA 20,000 word speech recognizer". — In: *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'87.* — Vol. 12. — IEEE. 1987, — P. 701—704.

[67] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. — "Distributed representations of words and phrases and their compositionality". — In:

Advances in neural information processing systems. — 2013, — P. 3111—3119.

[68] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. — "Efficient estimation of word representations in vector space". — In: *arXiv preprint arXiv:1301.3781* (2013).

[69] David M Blei and John D Lafferty. — "Dynamic topic models". — In: *Proceedings of the 23rd international conference on Machine learning.* — ACM. 2006, — P. 113—120.

[70] Paul H Algoet, Thomas M Cover, et al. — "Asymptotic optimality and asymptotic equipartition properties of log-optimum investment". — In: *The Annals of Probability* 16.2 (1988), p. 876—898.

[71] Thomas M Cover and Joy A Thomas. — "Entropy, relative entropy and mutual information". — In: *Elements of information theory* 2 (1991), p. 1—55.

[72] E ShannClaudeon. — "Prediction and entropy of printed English". — In: *Bell system technical journal* 30.1 (1951), p. 50—64.

[73] Thomas Cover and Roger King. — "A convergent gambling estimate of the entropy of English". — In: *IEEE Transactions on Information Theory* 24.4 (1978), p. 413—421.

[74] Peter F Brown, Vincent J Della Pietra, Robert L Mercer, Stephen A Della Pietra, and Jennifer C Lai. — "An estimate of an upper bound for the entropy of English". — In: *Computational Linguistics* 18.1 (1992), p. 31—40.

[75] William W Cohen, Robert E Schapire, and Yoram Singer. — "Learning to order things". — In: *Advances in Neural Information Processing Systems*. — 1998, — P. 451—457.

[76] Ronan Collobert, Jason Weston, Leon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. — "Natural language processing (almost) from scratch". — In: *Journal of Machine Learning Research* 12.8 (2011), p. 2493—2537.

[77] Jeffrey Pennington, Richard Socher, and Christopher Manning. — "Glove: Global vectors for word representation". — In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. — 2014, — P. 1532—1543.

[78] Melvin Earl Maron. — "Automatic indexing: an experimental inquiry". — In: *Journal of the ACM (JACM)* 8.3 (1961), p. 404—417.

[79] Thomas Bayes, Richard Price, and John Canton. — "An essay towards solving a problem in the doctrine of chances". — In: (1763).

[80] Frederick Mosteller and David L Wallace. — "Inference in an authorship problem: A comparative study of discrimination methods applied to the authorship of the disputed federalist papers". — In: *Journal of the American Statistical Association* 58.302 (1963), p. 275—309.

[81] Mehran Sahami, Susan Dumais, David Heckerman, and Eric Horvitz. — "A Bayesian approach to filtering junk e-mail". — In: *Learning for Text Categorization: Papers from the 1998 workshop.* — Vol. 62. — Madison, Wisconsin. 1998, — P. 98—105.

[82] Vangelis Metsis, Ion Androutsopoulos, and Georgios Paliouras. — "Spam filtering with naive bayes-which naive bayes?" — In: *CEAS.* — Vol. 17. — Mountain View, CA. 2006, — P. 28—69.

[83] Sida Wang and Christopher D Manning. — "Baselines and bigrams: Simple, good sentiment and topic classification". — In: *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2.* — Association for Computational Linguistics. 2012, — P. 90—94.

[84] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. — "Thumbs up?: sentiment classification using machine learning techniques". — In: *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10.* — Association for Computational Linguistics. 2002, — P. 79—86.

[85] Andrew McCallum, Kamal Nigam, et al. — "A comparison of event models for naive bayes text classification". — In: *AAAI-98 workshop on learning for text categorization.* — Vol. 752. — 1. — Citeseer. 1998, — P. 41—48.

[86] Bo Pang, Lillian Lee, et al. — "Opinion mining and sentiment analysis". — In: *Foundations and Trends® in Information Retrieval* 2.2 (2008), p. 1—135.

[87] Bing Liu and Lei Zhang. — "A survey of opinion mining and sentiment analysis". — In: *Mining text data.* — Springer, 2012, — P. 415—463.

[88] Efstathios Stamatatos. — "A survey of modern authorship attribution methods". — In: *Journal of the American Society for information Science and Technology* 60.3 (2009), p. 538—556.

[89] Hinrich Schutze, Christopher D Manning, and Prabhakar Raghavan. — *Introduction to information retrieval.* — Vol. 39. — Cambridge University Press, 2008.

[90] Ledell Wu, Adam Fisch, Sumit Chopra, Keith Adams, Antoine Bordes, and Jason Weston. — "Starspace: Embed all the things!" — In: *arXiv preprint arXiv:1709.03856* (2017).

[91] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. — "Enriching word vectors with subword information". — In: *arXiv preprint arXiv:1607.04606* (2016).

[92] Matteo Pagliardini, Prakhar Gupta, and Martin Jaggi. — "Unsupervised learning of sentence embeddings using compositional n-gram features". — In: *arXiv preprint arXiv:1703.02507* (2017).

[93] Gregory Finley, Stephanie Farmer, and Serguei Pakhomov. — "What analogies reveal about word vectors and their compositionality". — In: *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (* SEM 2017)*. — 2017, — P. 1—11.

[94] Alexander Panchenko, Fide Marten, Eugen Ruppert, Stefano Faralli, Dmitry Ustalov, Simone Paolo Ponzetto, and Chris Biemann. — "Unsupervised, knowledge-free, and interpretable word sense disambiguation". — In: *arXiv preprint arXiv:1707.06878* (2017).

[95] Maria Pelevina, Nikolay Arefyev, Chris Biemann, and Alexander Panchenko. — "Making sense of word embeddings". — In: *arXiv preprint arXiv:1708.03390* (2017).

[96] Sergey Bartunov, Dmitry Kondrashkin, Anton Osokin, and Dmitry Vetrov. — "Breaking sticks and ambiguities with adaptive skip-gram". — In: *Artificial Intelligence and Statistics*. — 2016, — P. 130—138.

[97] Eric H Huang, Richard Socher, Christopher D Manning, and Andrew Y Ng. — "Improving word representations via global context and multiple word prototypes". — In: *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume*

1. — Association for Computational Linguistics. 2012, — P. 873—882.

[98] Anna Gladkova and Aleksandr Drozd. — "Intrinsic evaluations of word embeddings: What can we do better?" — In: *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP.* — 2016, — P. 36—42.

[99] Sebastian Schuster, Joakim Nivre, and Christopher D. Manning. — "Sentences with Gapping: Parsing and Reconstructing Elided Predicates". — In: *North American Chapter of the Association of Computational Linguistics (NAACL).* — 2018. — URL: https://nlp.stanford.edu/pubs/schuster2018gapping.pdf.

[100] Mihail Eric and Christopher D Manning. — "A copy-augmented sequence-to-sequence architecture gives good performance on task-oriented dialogue". — In: *arXiv preprint arXiv:1701.04024* (2017).

[101] Sida I Wang, Samuel Ginn, Percy Liang, and Christoper D Manning. — "Naturalizing a programming language via interactive learning". — In: *arXiv preprint arXiv:1704.06956* (2017).

[102] Jiwei Li, Will Monroe, Tianlin Shi, Sebastien Jean, Alan Ritter, and Dan Jurafsky. — "Adversarial learning for neural dialogue generation". — In: *arXiv preprint arXiv:1701.06547* (2017).

[103] Ziang Xie, Sida I Wang, Jiwei Li, Daniel Levy, Aiming Nie, Dan Jurafsky, and Andrew Y Ng. — "Data noising as smoothing in neural network language models". — In: *arXiv preprint arXiv:1703.02573* (2017).

[104] Wouter De Nooy, Andrej Mrvar, and Vladimir Batagelj. — *Exploratory social network analysis with Pajek.* — Cambridge University Press, 2018.

[105] Jacob Levy Moreno. — "Who shall survive? Foundations of sociometry, group psychotherapy and socio-drama". — In: (1953).

[106] Maksim Tsvetovat and Alexander Kouznetsov. — *Social Network Analysis for Startups: Finding connections on the social web.* — " O'Reilly Media, Inc.", 2011.

[107] Scott Atherley, Clarence Dillon, and Vince Kane. — "A Model of Policy Formation Through Simulated Annealing: The Impact of Preference Alignment on Productivity and Satisfaction". — In: *International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction.* — Springer. 2015, — P. 93—100.

[108] Wayne Porter, Camber Warren, and Rob Schroeder. — "Mapping the Navy Innovation Network Using Social Network Analysis". — In: (2018).

[109] Reza Zafarani, Mohammad Ali Abbasi, and Huan Liu. — *Social media mining: an introduction.* — Cambridge University Press, 2014.

[110] Anvar Kurmukov, Yulia Dodonova, and Leonid Zhukov. — "Classification of normal and pathological brain networks based on similarity in graph partitions". — In: *Data Mining Workshops (ICDMW), 2016 IEEE 16th International Conference on.* — IEEE. 2016, — P. 107—112.

[111] Ilya Makarov, Oleg Bulanov, and Leonid E Zhukov. — "Co-author recommender system". — In: *International Conference on Network Analysis.* — Springer. 2016, — P. 251—257.

[112] Ilya Makarov, Olga Gerasimova, Pavel Sulimov, and Leonid E Zhukov. — "Recommending Co-authorship via Network Embeddings and Feature Engineering: The case of National Research University Higher School of Economics". — In: *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries.* — ACM. 2018, — P. 365—366.

[113] Duncan J Watts and Steven H Strogatz. — "Collective dynamics of 'small-world'networks". — In: *nature* 393.6684 (1998), p. 440.

[114] Albert-Laszlo Barabasi and Reka Albert. — "Emergence of scaling in random networks". — In: *science* 286.5439 (1999), p. 509—512.

[115] Santo Fortunato. — "Community detection in graphs". — In: *Physics reports* 486.3 (2010), p. 75—174.

[116] Andrea Lancichinetti and Santo Fortunato. — "Community detection algorithms: a comparative analysis". — In: *Physical review E* 80.5 (2009), p. 056117.

[117] Vincent W Zheng, Sandro Cavallari, Hongyun Cai, Kevin Chen-Chuan Chang, and Erik Cambria. — "From node embedding to community embedding". — In: *arXiv preprint arXiv:1610.09950* (2016).

[118] Zemin Liu, Vincent W Zheng, Zhou Zhao, Fanwei Zhu, Kevin Chen-Chuan Chang, Minghui Wu, and Jing Ying. — "Semantic Proximity Search on Heterogeneous Graph by Proximity Embedding." — In: *AAAI*. — 2017, — P. 154—160.

[119] Nicholas C Mullins. — "The development of specialties in social science: The case of ethnomethodology". — In: *Science Studies* 3.3 (1973), p. 245—273.

[120] Pham Minh Chuan, Mumtaz Ali, Tran Dinh Khang, Nilanjan Dey, et al. — "Link prediction in co-authorship networks based on hybrid content similarity metric". — In: *Applied Intelligence* 48.8 (2018), p. 2470—2486.

[121] Yang Chen, Cong Ding, Jiyao Hu, Ruichuan Chen, Pan Hui, and Xiaoming Fu. — "Building and analyzing a global co-authorship network using Google Scholar data". — In: *Proceedings of the 26th International Conference on World Wide Web Companion*. — International World Wide Web Conferences Steering Committee. 2017, — P. 1219—1224.

[123] Nikolai Alekseevich Mazov, Vadim Nikolaevich Gureev, and Mikhail Ivanovich Epov. — "Russian publications and journals on Earth sciences in international databases". — In: *Herald of the Russian Academy of Sciences* 85.1 (2015), p. 20—25.

[124] Aleksandr Ivanovich Shherbakov. — "Jeffektivnost' nauchnoj dejatel'nosti v SSSR". — In: (1982).

[125] OA Ovchinnikov. — "K metodologii ocenki nauchnoj dejatel'nosti v nauchnyh i obrazovatel'nyh uchrezhdenijah Rossijskoj Federacii". — In: *Vestnik Moskovskogo universiteta MVD Rossii* 3 (2009), p. 48—51.

[126] Konstantin Fursov, Roshhina Jana, and Oksana Balmush. — "Faktory rezul'tativnosti nauchnoj dejatel'nosti: mikrourovnevyj analiz". — In: *Forsajt* 10.2 (2016).

[127] Natal'ja Shmatko and Galina Volkova. — "Sluzhba ili sluzhenie? Motivacionnye patterny rossijskih uchenyh". — In: *Forsajt* 11.2 (2017).

[128] Nils Brede Moe, Torgeir Dingsoyr, and Tore Dybaa. — "Understanding self-organizing teams in agile software development". — In: *Software Engineering, 2008. ASWEC 2008. 19th Australian Conference on.* — IEEE. 2008, — P. 76—85.

[129] Dmitrij Aleksandrovich Novikov. — "Matematicheskie modeli formirovanija i funkcionirovanija komand". —

In: *M.: Izdatel'stvo fiziko-matematicheskoj literatury* (2008).

[130] Damir Kajrzhanovich Bejl'hanov and Irina Jur'evna Kvjatkovskaja. — "Ispol'zovanie modeli kompetencij v processe komandoobrazovanija". — In: *Tehnicheskie nauki-ot teorii k praktike* 30 (2014), p. 7—12.

[131] Alex Bavelas. — "A mathematical model for group structures". — In: *Human organization* 7.3 (1948), p. 16—30.

[132] Przemyslaw Rozewski and Bartlomiej Malachowski. — "Competence management in knowledge-based organisation: case study based on higher education organisation". — In: *Knowledge Science, Engineering and Management* (2009), p. 358—369.

[133] Philip Leifeld, Sandra Wankmuller, Valentin TZ Berger, Karin Ingold, and Christiane Steiner. — "Collaboration patterns in the German political science co-authorship network". — In: *PloS one* 12.4 (2017), e0174671.

[134] Mehmet Ali Koseoglu, Fevzi Okumus, Eka Diraksa Putra, Mehmet Yildiz, and Ismail Cagri Dogan. — "Authorship trends, collaboration patterns, and co-authorship networks in lodging studies (1990–2016)". — In: *Journal of Hospitality Marketing & Management* 27.5 (2018), p. 561—582.

[135] Tung Manh Ho, Ha Viet Nguyen, Thu-Trang Vuong, Quang-Minh Dam, Hiep-Hung Pham, and Quan-Hoang Vuong. —

"with basic network measures of 2008-2017 Scopus data". — In: (2017).

[136] Horng-Jinh Chang and Whe-Min Wang. — "The Hidden Power of Social-Linkage in the Office: A Co-authorship Network Analysis". — In: *Proceedings of the 4th Multidisciplinary International Social Networks Conference on ZZZ*. — ACM. 2017, — P. 4.

[137] Tanveer Ahmed, Adeel Ahmed, Mubashir Ali, and Muhammad Kamran. — "Analysis of co-authorship in computer networks using centrality measures". — In: *Communication, Computing and Digital Systems (C-CODE), International Conference on*. — IEEE. 2017, — P. 54—57.

[138] Ionut Cristian Paraschiv, Mihai Dascalu, Stefan Trausan-Matu, Nicolae Nistor, Ambar Murillo Montes De Oca, and Danielle S McNamara. — "Semantic Similarity versus Co-authorship Networks: A Detailed Comparison". — In: *Control Systems and Computer Science (CSCS), 2017 21st International Conference on*. — IEEE. 2017, — P. 566—570.

[140] Giovanna Guimaraes Gielfi, Andre Tosi Furtado, Andre Sica de Campos, and Robert JW Tijssen. — "University-industry research collaboration in the Brazilian oil industry: the case of Petrobras". — In: *Rev. Bras. Inov* 16.2 (2017), p. 325—350.

[142] Martin Fowler and Jim Highsmith. — "The agile manifesto". — In: *Software Development* 9.8 (2001), p. 28—35.

[143] Nancy A Bonner, Nisha Kulangara, Sridhar Nerur, and James TC Teng. — "An Empirical Investigation of the Perceived Benefits of Agile Methodologies Using an Innovation-Theoretical model". — In: *Journal of Database Management (JDM)* 27.3 (2016), p. 38—63.

[144] Rashina Hoda and Latha K Murugesan. — "Multi-level agile project management challenges: A self-organizing team perspective". — In: *Journal of Systems and Software* 117 (2016), p. 245—257.

[145] Nils Brede Moe, Torgeir Dingsoyr, and Tore Dybaa. — "Overcoming barriers to self-management in software teams". — In: *IEEE software* 26.6 (2009).

[146] Omar A Alnuaimi, Lionel P Robert, and Likoebe M Maruping. — "Team size, dispersion, and social loafing in technology-supported teams: A perspective on the theory of moral disengagement". — In: *Journal of Management Information Systems* 27.1 (2010), p. 203—230.

[147] Roger Guimera, Brian Uzzi, Jarrett Spiro, and Luis A Nunes Amaral. — "Team assembly mechanisms determine collaboration network structure and team performance". — In: *Science* 308.5722 (2005), p. 697—702.

[148] Joseph Berger, Susan J Rosenholtz, and Morris Zelditch Jr. — "Status organizing processes". — In: *Annual review of sociology* 6.1 (1980), p. 479—508.

[149] Joseph Berger, Bernard P Cohen, and Morris Zelditch Jr. — "Status characteristics and social interaction". — In: *American Sociological Review* (1972), p. 241—255.

[150] Barton H Hamilton, Jack A Nickerson, and Hideo Owan. — "Team incentives and worker heterogeneity: An empirical analysis of the impact of teams on productivity and participation". — In: *Journal of political Economy* 111.3 (2003), p. 465—497.

[151] Andrea Prat. — "Should a team be homogeneous?" — In: *European Economic Review* 46.7 (2002), p. 1187—1207.

[152] Barry Boehm and Richard Turner. — "People factors in software management: lessons from comparing agile and plan-driven methods". — In: *Crosstalk-The Journal of Defense Software Engineering,(Dec 2003)* (2003).

[153] Maria Paasivaara, Ville T Heikkila, and Casper Lassenius. — "Experiences in scaling the product owner role in large-scale globally distributed scrum". — In: *Global Software Engineering (ICGSE), 2012 IEEE Seventh International Conference on.* — IEEE. 2012, — P. 174—178.

[154] Scott I Tannenbaum and Christopher P Cerasoli. — "Do team and individual debriefs enhance performance? A meta-analysis". — In: *Human factors* 55.1 (2013), p. 231—245.

[155] Gayle W Hill. — "Group versus individual performance: Are N+ 1 heads better than one?" — In: *Psychological bulletin* 91.3 (1982), p. 517.

[156] Maria Tims, Arnold B Bakker, Daantje Derks, and Willem Van Rhenen. — "Job crafting at the team and individual level: Implications for work engagement and performance". — In: *Group & Organization Management* 38.4 (2013), p. 427—454.

[157] David H Cropley and James C Kaufman. — "Measuring functional creativity: Non-expert raters and the Creative Solution Diagnosis Scale". — In: *The Journal of Creative Behavior* 46.2 (2012), p. 119—137.

[158] Philip W Jackson and Samuel Messick. — "The person, the product, and the response: conceptual problems in the assessment of creativity 1". — In: *Journal of personality* 33.3 (1965), p. 309—329.

[159] Daniel Olguin Olguin, Benjamin N Waber, BN Taemie Kim, Akshay Mohan, Koji Ara, and Alex Pentland. — "Sensible organizations: Technology and methodology for automatically measuring organizational behavior". — In: — Institute of Electrical and Electronics Engineers. 2008.

[160] Jeff Sutherland and Ken Schwaber. — "The scrum guide". — In: *The definitive guide to scrum: The rules of the game. Scrum. org* 268 (2013).

[161] Florian Pereme, Bertrand Rose, Virginie Goepp, Jean Pierre Radoux, and Abdelkrim Belhaoua. — "Toward An integrative CSDS based model of industrial R&D division efficiency". — In: *IFAC-PapersOnLine* 49.12 (2016), p. 1785—1790.

[162] Bryan D Edwards, Eric Anthony Day, Winfred Arthur Jr, and Suzanne T Bell. — "Relationships among team ability composition, team mental models, and team performance." — In: *Journal of Applied Psychology* 91.3 (2006), p. 727.

[163] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. — "Bag of tricks for efficient text classification". — In: *arXiv preprint arXiv:1607.01759* (2016).

[164] Anastasia Ianina, Lev Golitsyn, and Konstantin Vorontsov. — "Multi-objective topic modeling for exploratory search in tech news". — In: *Conference on Artificial Intelligence and Natural Language*. — Springer. 2017, — P. 181—193.

[165] Rogelio Oliva and John D Sterman. — "Death spirals and virtuous cycles". — In: *Handbook of Service Science*. — Springer, 2010, — P. 321—358.

[168] David J Ketchen Jr and Christopher L Shook. — "The application of cluster analysis in strategic management research: an analysis and critique". — In: *Strategic management journal* (1996), p. 441—458.

[169] Yuanhua Lv and ChengXiang Zhai. — "Adaptive term frequency normalization for BM25". — In: *Proceedings of the 20th ACM international conference on Information and knowledge management.* — ACM. 2011, — P. 1985—1988.

[171] Thomas Hofmann. — "Probabilistic latent semantic indexing". — In: *ACM SIGIR Forum.* — Vol. 51. — 2. — ACM. 2017, — P. 211—218.

[172] David M Blei. — "Probabilistic topic models". — In: *Communications of the ACM* 55.4 (2012), p. 77—84.

[173] Konstantin Vorontsov and Anna Potapenko. — "Additive regularization of topic models". — In: *Machine Learning* 101.1-3 (2015), p. 303—323.

[174] KV Vorontsov. — "Additive regularization for topic models of text collections". — In: *Doklady Mathematics.* — Vol. 89. — 3. — Springer. 2014, — P. 301—304.

[175] Konstantin Vorontsov, Anna Potapenko, and Alexander Plavin. — "Additive regularization of topic models for topic selection and sparse factorization". — In: *International Symposium on Statistical Learning and Data Sciences.* — Springer. 2015, — P. 193—202.

[176] KV Vorontsov and AA Potapenko. — "Modifikacii EM-algoritma dlya veroyatnostnogo tematicheskogo modelirovaniya". — In: *Mashinnoe obuchenie i analiz dannyh* 1.6 (2013), p. 657—686.

[177] Konstantin Vorontsov, Oleksandr Frei, Murat Apishev, Peter Romov, and Marina Dudarenko. — "BigARTM: open source library for regularized multimodal topic modeling of large collections". — In: *International Conference on Analysis of Images, Social Networks and Texts*. — Springer. 2015, — P. 370—381.

[178] Konstantin Vorontsov, Oleksandr Frei, Murat Apishev, Peter Romov, Marina Suvorova, and Anastasia Yanina. — "Non-Bayesian additive regularization for multimodal topic modeling of large collections". — In: *Proceedings of the 2015 Workshop on Topic Models: Post-Processing and Applications*. — ACM. 2015, — P. 29—37.

[179] Konstantin Vorontsov and Anna Potapenko. — "Tutorial on probabilistic topic modeling: Additive regularization for stochastic matrix factorization". — In: *International Conference on Analysis of Images, Social Networks and Texts*. — Springer. 2014, — P. 29—46.

[180] Krasnov Fedor and Yavorskiy Rostislav. — "Measurement of maturity level of a professional community". — In: *Business Informatics* 23.1 (2013).

[181] Dokuka Sofia, Yavorskiy Rostislav, and Krasnov Fedor. — "The Structure of Organization: the Coauthorship Network Case". — In: *Analysis of Images, Social Networks and Texts. 5th International Conference, AIST 2016, Yekaterinburg, Russia, April 7-9, 2016, Revised*

Selected Papers. Communications in Computer and Information Science. — Springer International Publishing. 2017, — P. 93—101.

[182] Alexandra Barysheva, Mikhail Petrov, and Rostislav Yavorskiy. — "Building Profiles of Blog Users Based on Comment Graph Analysis: The Habrahabr. ru Case". — In: *International Conference on Analysis of Images, Social Networks and Texts.* — Springer. 2015, — P. 257—262.

[183] Alexandra Barysheva, Anna Golubtsova, and Rostislav Yavorskiy. — "Profiling Less Active Users in Online Communities." — In: *SNAFCA@ ICFCA.* — 2015.

[187] Haiyan Hou, Hildrun Kretschmer, and Zeyuan Liu. — "The structure of scientific collaboration networks in Scientometrics". — In: *Scientometrics* 75.2 (2008), p. 189—202.

[188] Ying Ding. — "Scientific collaboration and endorsement: Network analysis of coauthorship and citation networks". — In: *Journal of informetrics* 5.1 (2011), p. 187—203.

[189] Albert-Laszlo Barabasi, Hawoong Jeong, Zoltan Neda, Erzsebet Ravasz, Andras Schubert, and Tamas Vicsek. — "Evolution of the social network of scientific collaborations". — In: *Physica A: Statistical mechanics and its applications* 311.3-4 (2002), p. 590—614.

[190] Francisco Jose Acedo, Carmen Barroso, Cristobal Casanueva, and Jose Luis Galan. — "Co-authorship in management and organizational studies: An empirical and network analysis". — In: *Journal of Management Studies* 43.5 (2006), p. 957—983.

[191] Marko A Rodriguez and Alberto Pepe. — "On the relationship between the structural and socioacademic communities of a coauthorship network". — In: *Journal of Informetrics* 2.3 (2008), p. 195—201.

[192] Mark EJ Newman. — "The structure of scientific collaboration networks". — In: *Proceedings of the National Academy of Sciences* 98.2 (2001), p. 404—409.

[193] Peng Wang, BaoWen Xu, YuRong Wu, and XiaoYu Zhou. — "Link prediction in social networks: the state-of-the-art". — In: *Science China Information Sciences* 58.1 (2015), p. 1—38.

[194] Mark S Granovetter. — "The strength of weak ties". — In: *American journal of sociology* (1973), p. 1360—1380.

[195] Roland Wiese, Markus Eiglsperger, and Michael Kaufmann. — "Yfiles —– visualization and automatic layout of graphs". — In: *Graph Drawing Software*. — Springer, 2004, — P. 173—191.

[196] Mark Granovetter. — "The strength of weak ties: A network theory revisited". — In: *Sociological theory* 1.1 (1983), p. 201—233.

[197] Serhiy A Yevtushenko. — "System of data analysis "Concept Explorer"". — In: *Proceedings of the 7th national conference on Artificial Intelligence KII.* — Vol. 2000. — 2000.

[198] Serhiy Yevtushenko, Julian Tane, Tim B Kaiser, Sergei Obiedkov, Joachim Hereth, and Heiko ReppeEPPE. — *ConExp-The Concept Explorer.* — 2006.

[199] Hwee Tou Ng, Wei Boon Goh, and Kok Leong Low. — "Feature selection, perceptron learning, and a usability case study for text categorization". — In: *ACM SIGIR Forum.* — Vol. 31. — SI. — ACM. 1997, — P. 67—73.

[200] Savio LY Lam and Dik Lun Lee. — "Feature reduction for neural network based text categorization". — In: *Database Systems for Advanced Applications, 1999. Proceedings., 6th International Conference on.* — IEEE. 1999, — P. 195—202.

[201] Wai Lam, Miguel Ruiz, and Padmini Srinivasan. — "Automatic text categorization and its application to text retrieval". — In: *IEEE Transactions on Knowledge & Data Engineering* 6 (1999), p. 865—879.

[202] Yoon Kim. — "Convolutional neural networks for sentence classification". — In: *arXiv preprint arXiv:1408.5882* (2014).

[203] Tomas Mikolov, Martin Karafiat, Lukas Burget, Jan Cernocky, and Sanjeev Khudanpur. — "Recurrent neural network based language model". — In: *Eleventh Annual*

Conference of the International Speech Communication Association. — 2010.

[204] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. — "Imagenet classification with deep convolutional neural networks". — In: *Advances in neural information processing systems.* — 2012, — P. 1097—1105.

[205] George A Miller. — "WordNet: a lexical database for English". — In: *Communications of the ACM* 38.11 (1995), p. 39—41.

[206] Andrew L Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. — "Learning word vectors for sentiment analysis". — In: *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies-volume 1.* — Association for Computational Linguistics. 2011, — P. 142—150.

[207] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Ng, and Christopher Potts. — "Recursive deep models for semantic compositionality over a sentiment treebank". — In: *Proceedings of the 2013 conference on empirical methods in natural language processing.* — 2013, — P. 1631—1642.

[208] Cem Akkaya, Janyce Wiebe, and Rada Mihalcea. — "Subjectivity word sense disambiguation". — In: *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1.* —

Association for Computational Linguistics. 2009, — P. 190—199.

[209] Pavel Bakhtin, Ozcan Saritas, Alexander Chulok, Ilya Kuzminov, and Anton Timofeev. — "Trend monitoring for linking science and strategy". — In: *Scientometrics* 111.3 (2017), p. 2059—2075.

[210] Ilya Kuzminov, Alexey Bereznoy, and Pavel Bakhtin. — "Global energy challenges and the national economy: stress scenarios for Russia". — In: *foresight* 19.2 (2017), p. 174—197.

[211] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. — "Latent Dirichlet Allocation". — In: *J. Mach. Learn. Res.* 3 (Mar. 2003), p. 993—1022. — URL: http://dl.acm.org/citation.cfm?id=944919.944937.

[212] Stuart J Russell and Peter Norvig. — "Artificial intelligence: a modern approach (International Edition)". — In: (2002).

[213] Willem JM Levelt. — *Speaking: From intention to articulation*. — Vol. 1. — MIT press, 1993.

[214] NV Lipchiu and KI Lipchiu. — "Metodologiya nauchnogo issledovaniya: uchebnoe posobie". — In: *Krasnodar: KubGAU* (2013).

[215] MA Mkrtchyan. — "Fazy perekhodnogo perioda ot gruppovogo sposoba obucheniya k kollektivnomu". — In: *Kollektivnyj sposob obucheniya* 2 (1995), p. 8—11.

[216] NV Danilevskaya. — “Ocenka kak istochnik dinamiki tekstoobrazovaniya v nauchnoj kommunikacii”. — In: *Mezhdunarodnyj nauchno-issledovatel'skij zhurnal* 12 (2016), p. 27—30.

[217] IV Kleshheva. — “Ocenka jeffektivnosti nauchno-issledovatel'skoj dejatel'nosti studentov”. — In: *SPb: NIU ITMO* (2014).

[218] VI Levin. — “Vozmozhna li pravil'naja ocenka vklada uchenogo v nauku s pomoshh'ju indeksa hirsha? primery”. — In: *Matematicheskie metody v tehnike i tehnologijah-MMTT* 6 (2016), p. 100—102.

[219] Frederick Winslow Taylor. — *Scientific management.* — Routledge, 2004.

[220] Fangfang Wei, Guijie Zhang, Yuqiang Feng, Luning Liu, and Zhen Shao. — “A co-authorship network-based method for understanding the evolution of a research area: A case of information systems research”. — In: *Malaysian Journal of Library & Information Science* 22.2 (2017), p. 1—14.

[221] David Andres Munoz, Juan Pablo Queupil, and Pablo Fraser. — “Assessing collaboration networks in educational research: A co-authorship-based social network analysis approach”. — In: *International Journal of Educational Management* 30.3 (2016), p. 416—436.

[222] Antonio Perianes-Rodriguez, Carlos Olmeda-Gomez, and Felix Moya-Anegon. — "Detecting, identifying and visualizing research groups in co-authorship networks". — In: *Scientometrics* 82.2 (2010), p. 307—319.

[223] Eldon Y Li, Chien Hsiang Liao, and Hsiuju Rebecca Yen. — "Co-authorship networks and research impact: A social capital perspective". — In: *Research Policy* 42.9 (2013), p. 1515—1530.

[224] Katy Borner, Luca Dall'Asta, Weimao Ke, and Alessandro Vespignani. — "Studying the emerging global brain: Analyzing and visualizing the impact of co-authorship teams". — In: *Complexity* 10.4 (2005), p. 57—67.

[225] Alireza Abbasi, Kon Shing Kenneth Chung, and Liaquat Hossain. — "Egocentric analysis of co-authorship network structure, position and performance". — In: *Information Processing & Management* 48.4 (2012), p. 671—679.

[226] Cynthia F Kurtz. — *Collective Network Analysis.* — 2009.

[227] Haiyan Hou, Hildrun Kretschmer, and Zeyuan Liu. — "The structure of scientific collaboration networks in Scientometrics". — In: *Scientometrics* 75.2 (2007), p. 189—202.

[228] Alistair Cockburn and Jim Highsmith. — "Agile software development, the people factor". — In: *Computer* 34.11 (2001), p. 131—133.

[229] John P Hausknecht and Jacob A Holwerda. — "When does employee turnover matter? Dynamic member configurations, productive capacity, and collective performance". — In: *Organization Science* 24.1 (2013), p. 210—225.

[230] Ulrik Brandes. — "A faster algorithm for betweenness centrality". — In: *Journal of mathematical sociology* 25.2 (2001), p. 163—177.

[231] Ulrik Brandes. — "On variants of shortest-path betweenness centrality and their generic computation". — In: *Social Networks* 30.2 (2008), p. 136—145.

[232] Ulrik Brandes and Christian Pich. — "Centrality estimation in large networks". — In: *International Journal of Bifurcation and Chaos* 17.07 (2007), p. 2303—2318.

[233] Ulrik Brandes and Daniel Fleischer. — "Centrality measures based on current flow". — In: *Annual Symposium on Theoretical Aspects of Computer Science*. — Springer. 2005, — P. 533—544.

[234] Mark EJ Newman. — "A measure of betweenness centrality based on random walks". — In: *Social networks* 27.1 (2005), p. 39—54.

[235] Ilya Segalovich. — "A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine". — In: — MLMTA03, 2003. — URL: http://cache-mskstoredata01.cdn.yandex.net/download.yandex.ru/company/iseg-las-vegas.pdf.

[236] Thomas Hofmann. — "Probabilistic Latent Semantic Indexing". — In: *Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. — SIGIR '99. — Berkeley, California, USA: ACM, 1999, — P. 50—57. — ISBN: 1-58113-096-1. — DOI: 10.1145/312624.312649. — URL: http://doi.acm.org/10.1145/312624.312649.

[237] Danish Contractor, Tanveer A. Faruquie, and L. Venkata Subramaniam. — "Unsupervised Cleansing of Noisy Text". — In: *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*. — COLING '10. — Beijing, China: Association for Computational Linguistics, 2010, — P. 189—196. — URL: http://dl.acm.org/citation.cfm?id=1944566.1944588.

[238] DeLesley Hutchins & Peter Norvig Moshe Looks Marcello Herreshoff. — "DEEP LEARNING WITH DYNAMIC COMPUTATION GRAPHS". — In: ICLR 2017 (2017).

[239] Laurens van der Maaten. — "Visualizing Data using t-SNE". — In: Journal of Machine Learning Research 9(Nov):2579-2605 (2008).

[240] B. J. Frey and D. Dueck. — "Clustering by Passing Messages Between Data Points". — In: *Science* 315.5814 (Feb. 2007), p. 972—976. — DOI: 10.1126/science.1136800. — URL: https://doi.org/10.1126%2Fscience.1136800.

[241] Andrej Borshhev. — “Primenenie imitacionnogo modelirovanija v Rossii-sostojanie na 2007g.” — In: *Biznes-informatika* 4 (2008).

[242] Andrei Borshchev. — *The big book of simulation modeling: multimethod modeling with AnyLogic 6.* — AnyLogic North America, 2013.

[243] Filip Agneessens and Rafael Wittek. — “Where do intra-organizational advice relations come from? The role of informal status and social capital in social exchange”. — In: *Social Networks* 34.3 (2012), p. 333—345.

[244] Filip Agneessens and Rafael Wittek. — “Social capital and employee well-being: disentangling intrapersonal and interpersonal selection and influence mechanisms”. — In: *Revue francaise de sociologie* 49.3 (2008), p. 613—637.

[245] Brigham S Anderson, Carter Butts, and Kathleen Carley. — “The interaction of size and density with graph-level indices”. — In: *Social Networks* 21.3 (1999), p. 239—267.

[246] Lea Ellwardt, Christian Steglich, and Rafael Wittek. — “The co-evolution of gossip and friendship in workplace social networks”. — In: *Social Networks* 34.4 (2012), p. 623—633.

[247] David Krackhardt. — “The ties that torture: Simmelian tie analysis in organizations”. — In: *Research in the Sociology of Organizations* 16.1 (1999), p. 183—210.

[248] David Krackhardt. — "Assessing the political landscape: Structure, cognition, and power in organizations". — In: *Administrative science quarterly* (1990), p. 342—369.

[249] David Krackhardt and Jeffrey R Hanson. — "Informal networks". — In: *Harvard business review* 71.4 (1993), p. 104—111.

[250] Christina Prell. — *Social network analysis: History, theory and methodology.* — Sage, 2012.

[251] R Core Team. — *R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2013.* — 2014.

[252] Konstantin Ushakov. — "Diagnostics of an Educational Institution Real Structure". — In: *Educational Studies* 4 (2013), p. 247—260.

[253] "Agile Scientific". — In: (2007). — URL: ---%20Rezhim%20dostupa:%20%5Curl%7Bhttps://agilescientific.com/articles/%7D,%20svobodnyj..

[254] Laura A King, Lori McKee Walker, and Sheri J Broyles. — "Creativity and the five-factor model". — In: *Journal of research in personality* 30.2 (1996), p. 189—203.

[255] VS Tikin. — "Jeffektivnost'-ne kojefficient". — In: *Jekonomicheskie nauki* 7 (2009), p. 94—97.

[256] Majkl Porter. — *Mezhdunarodnaja konkurencija: konkurentnye preimushhestva stran.* — Al'pina Publisher, 1993.

[257] Jason D Rennie, Lawrence Shih, Jaime Teevan, and David R Karger. — "Tackling the poor assumptions of naive bayes text classifiers". — In: *Proceedings of the 20th international conference on machine learning (ICML-03)*. — 2003, — P. 616—623.

[258] Akshar Bharati, K Prakash Rao, Rajeev Sangal, and SM Bendre. — "Basic statistical analysis of corpus and cross comparison among corpora". — In: *Technical Report of Indian Institute of Information Technology* (2000).

[259] Roland Hodler, Sorawoot Srisuma, Alberto Vesperoni, and Noemie Zurlinden. — "Measuring Ethnic Stratification and its Effect on Trust in Africa". — In: (2018).

[260] Paul Rayson. — "Matrix: A statistical method and software tool for linguistic analysis through corpus comparison". — PhD thesis. Lancaster University, 2003.

[261] Paul Rayson and Roger Garside. — "Comparing corpora using frequency profiling". — In: *Proceedings of the workshop on Comparing corpora-Volume 9*. — Association for Computational Linguistics. 2000, — P. 1—6.

[262] Teresa Mihwa Chung. — "A corpus comparison approach for terminology extraction". — In: *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication* 9.2 (2003), p. 221—246.

[263] Maria Jose Marin Perez. — "Measuring the degree of specialisation of sub-technical legal terms through corpus comparison". — In: *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication* 22.1 (2016), p. 80—102.

[264] Fedor Krasnov. — "Seismic Patents (BoW)". — In: (2019). — DOI: http://dx.doi.org/10.17632/x5z5r6br2z.1.

[265] Dieter Pawelczak. — "Benefits and drawbacks of source code plagiarism detection in engineering education". — In: *2018 IEEE Global Engineering Education Conference (EDUCON)*. — IEEE. 2018, — P. 1048—1056.

[266] Frederic Jack. — "Study on the Different Forms of Plagiarism in Textual Data and Image: Internal and External Detection". — In: *Advanced Metaheuristic Methods in Big Data Retrieval and Analytics*. — IGI Global, 2019, — P. 75—90.

[267] Derek Nadhan, Maryam Gholami Mayani, and Rolv Rommetveit. — "Drilling with Digital Twins". — In: *IADC/SPE Asia Pacific Drilling Technology Conference and Exhibition*. — Bangkok, Thailand: Society of Petroleum Engineers, 2018, — P. 18. — ISBN: 978-1-61399-574-7. — DOI: 10.2118/191388-MS. — URL: https://doi.org/10.2118/191388-MS.

[268] Jan Van Os. — *The Digital Twin throughout the Lifecycle.* — Providence, Rhode Island, USA, 2018. — URL: https://doi.org/.

[269] Maryam Gholami Mayani, Rolv Rommetveit, Sven Inge Oedegaard, and Morten Svendsen. — *Drilling Automated Realtime Monitoring Using Digital Twin.* — Abu Dhabi, UAE, 2018. — DOI: 10.2118/192807-MS. — URL: https://doi.org/10.2118/192807-MS.

[270] Are Follesdal Tjonn. — *Digital Twin Through the Life of a Field.* — Abu Dhabi, UAE, 2018. — DOI: 10.2118/193203-MS. — URL: https://doi.org/10.2118/193203-MS.

[271] Tushar Poddar. — *Digital Twin Bridging Intelligence Among Man, Machine and Environment.* — Kuala Lumpur, Malaysia, 2018. — DOI: 10.4043/28480-MS. — URL: https://doi.org/10.4043/28480-MS.

[272] Gurtej Saini, Pradeepkumar Ashok, Eric van Oort, and Matthew R Isbell. — *Accelerating Well Construction Using a Digital Twin Demonstrated on Unconventional Well Data in North America.* — Houston, Texas, USA, 2018. — DOI: 10.15530/URTEC-2018-2902186. — URL: https://doi.org/10.15530/URTEC-2018-2902186.

[273] Partha Sharma, David Knezevic, Phuong Huynh, and Grzegorz Malinowski. — *RB-FEA Based Digital Twin*

for Structural Integrity Assessment of Offshore Structures. — Houston, Texas, USA, 2018. — DOI: 10.4043/29005-MS. — URL: https://doi.org/10.4043/29005-MS.

[274] Elgonda La Grange. — *A Roadmap for Adopting a Digital Lifecycle Approach to Offshore Oil and Gas Production.* — Houston, Texas, USA, 2018. — DOI: 10.4043/28669-MS. — URL: https://doi.org/10.4043/28669-MS.

[275] Peter F Brown, Vincent J Della Pietra, Stephen A Della Pietra, and Robert L Mercer. — "The mathematics of statistical machine translation: Parameter estimation". — In: *Computational linguistics* 19.2 (1993), p. 263—311.

[276] Jorg Tiedemann. — "Improved sentence alignment for movie subtitles". — In: *Proceedings of RANLP.* — Vol. 7. — 2007.

[277] Einav Itamar and Alon Itai. — "Using Movie Subtitles for Creating a Large-Scale Bilingual Corpora." — In: *LREC.* — 2008.

[278] Jorg Tiedemann. — "Parallel Data, Tools and Interfaces in OPUS." — In: *Lrec.* — Vol. 2012. — 2012, — P. 2214—2218.

[279] Mehdi Mohammadi and Nasser GhasemAghaee. — "Building bilingual parallel corpora based on wikipedia". — In: *2010 Second International Conference on Computer Engineering and Applications.* — Vol. 2. — IEEE. 2010, — P. 264—268.

[280] Warren Weaver. — “Translation”. — In: *Machine translation of languages* 14 (1955), p. 15—23.

[281] Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M Rush. — “Opennmt: Open-source toolkit for neural machine translation”. — In: *arXiv preprint arXiv:1701.02810* (2017).

[282] Denny Britz, Anna Goldie, Minh-Thang Luong, and Quoc Le. — “Massive exploration of neural machine translation architectures”. — In: *arXiv preprint arXiv:1703.03906* (2017).

[283] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. — “Sequence to sequence learning with neural networks”. — In: *Advances in neural information processing systems*. — 2014, — P. 3104—3112.

[284] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. — “BLEU: a method for automatic evaluation of machine translation”. — In: *Proceedings of the 40th annual meeting on association for computational linguistics*. — Association for Computational Linguistics. 2002, — P. 311—318.

[285] Philipp Koehn, Marcello Federico, Wade Shen, Nicola Bertoldi, Ondrej Bojar, Chris Callison-Burch, Brooke Cowan, Chris Dyer, Hieu Hoang, Richard Zens, et al. — “Open source toolkit for statistical machine translation: Factored translation models and confusion

network decoding". — In: *Final Report of the 2006 JHU Summer Workshop*. — 2006.

[286] Spence Green, Daniel Cer, and Christopher Manning. — "Phrasal: A toolkit for new directions in statistical machine translation". — In: *Proceedings of the Ninth Workshop on Statistical Machine Translation*. — 2014, — P. 114—121.

[287] Will Monroe, Jennifer Hu, Andrew Jong, and Christopher Potts. — "Generating bilingual pragmatic color references". — In: *arXiv preprint arXiv:1803.03917* (2018).

[288] David M Blei, Andrew Y Ng, and Michael I Jordan. — "Latent dirichlet allocation". — In: *Journal of machine Learning research* 3.1 (2003), p. 993—1022.

[289] Sergei Koltsov, Sergei Pashakhin, and Sofia Dokuka. — "A Full-Cycle Methodology for News Topic Modeling and User Feedback Research". — In: *International Conference on Social Informatics*. — Springer. 2018, — P. 308—321.

[290] Yanir Seroussi, Fabian Bohnert, and Ingrid Zukerman. — "Authorship attribution with author-aware topic models". — In: *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*. — Association for Computational Linguistics. 2012, — P. 264—269.

[291] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. — "Latent Dirichlet Allocation". — In: *J. Mach. Learn. Res.* 3 (Mar. 2003), p. 993—1022. — URL: http://dl.acm.org/citation.cfm?id=944919.944937.

[292] Konstantin Vorontsov, Anna Potapenko, and Alexander Plavin. — "Additive Regularization of Topic Models for Topic Selection and Sparse Factorization". — In: *Statistical Learning and Data Sciences*. — Springer International Publishing, 2015, — P. 193—202. — DOI: 10.1007/978-3-319-17091-6_14. — URL: https://doi.org/10.1007/978-3-319-17091-6_14.

[293] Sergei Koltsov, Sergei Pashakhin, and Sofia Dokuka. — "A Full-Cycle Methodology for News Topic Modeling and User Feedback Research". — In: *Social Informatics*. — Ed. by Steffen Staab, Olessia Koltsova, and Dmitry I. Ignatov. — Cham: Springer International Publishing, 2018, — P. 308—321. — DOI: 10.1007/978-3-030-01129-1_19. — URL: https://doi.org/10.1007/978-3-030-01129-1_19.

[294] Yanir Seroussi, Fabian Bohnert, and Ingrid Zukerman. — "Authorship Attribution with Author-aware Topic Models". — In: *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers - Volume 2*. — ACL '12. — Jeju Island, Korea: Association for Computational Linguistics, 2012, —

P. 264—269. — URL: http://dl.acm.org/citation.cfm?id=2390665.2390728.

[295] Debin Fang, Haixia Yang, Baojun Gao, and Xiaojun Li. — "Discovering research topics from library electronic references using latent Dirichlet allocation". — In: *Library Hi Tech* 36.3 (Feb. 2018), p. 400—410. — DOI: 10.1108/LHT-06-2017-0132. — URL: https://app.dimensions.ai/details/publication/pub.1101114990.

[296] David Binkley, Daniel Heinz, Dawn Lawrie, and Justin Overfelt. — "Understanding LDA in Source Code Analysis". — In: *Proceedings of the 22Nd International Conference on Program Comprehension*. — ICPC 2014. — Hyderabad, India: ACM, 2014, — P. 26—36. — ISBN: 978-1-4503-2879-1. — DOI: 10.1145/2597008.2597150. — URL: http://doi.acm.org/10.1145/2597008.2597150.

[297] Amritanshu Agrawal, Wei Fu, and Tim Menzies. — "What is wrong with topic modeling? And how to fix it using search-based software engineering". — In: *Information and Software Technology* 98 (Jan. 2018), p. 74—88. — DOI: 10.1016/j.infsof.2018.02.005. — URL: https://doi.org/10.1016/j.infsof.2018.02.005.

[298] Rainer Storn and Kenneth Price. — "Differential Evolution – A Simple and Efficient Heuristic for Global Optimization over Continuous Spaces". — In: *J. of Global Optimization* 11.4 (Dec. 1997), p. 341—359. — DOI: 10.

1023/A:1008202821328. — URL: https://doi.org/10.1023/A:1008202821328.

[299] Arthur Asuncion, Max Welling, Padhraic Smyth, and Yee Whye Teh. — "On Smoothing and Inference for Topic Models". — In: *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*. — UAI '09. — Montreal, Quebec, Canada: AUAI Press, 2009, — P. 27—34. — ISBN: 978-0-9749039-5-8. — URL: http://dl.acm.org/citation.cfm?id=1795114.1795118.

[300] Hanna M. Wallach, Iain Murray, Ruslan Salakhutdinov, and David Mimno. — "Evaluation Methods for Topic Models". — In: *Proceedings of the 26th Annual International Conference on Machine Learning*. — ICML '09. — Montreal, Quebec, Canada: ACM, 2009, — P. 1105—1112. — ISBN: 978-1-60558-516-1. — DOI: 10.1145/1553374.1553515. — URL: http://doi.acm.org/10.1145/1553374.1553515.

[301] Jonathan Chang, Jordan Boyd-Graber, Sean Gerrish, Chong Wang, and David M. Blei. — "Reading Tea Leaves: How Humans Interpret Topic Models". — In: *Proceedings of the 22Nd International Conference on Neural Information Processing Systems*. — NIPS'09. — Vancouver, British Columbia, Canada: Curran Associates Inc., 2009, — P. 288—296. — ISBN: 978-1-61567-911-9. — URL: http://dl.acm.org/citation.cfm?id=2984093.2984126.

[302] Sergei Koltcov, Olessia Koltsova, and Sergey Nikolenko. — "Latent Dirichlet Allocation: Stability and Applications to Studies of User-generated Content". — In: *Proceedings of the 2014 ACM Conference on Web Science*. — WebSci '14. — Bloomington, Indiana, USA: ACM, 2014, — P. 161—165. — ISBN: 978-1-4503-2622-3. — DOI: 10.1145/2615569.2615680. — URL: http://doi.acm.org/10.1145/2615569.2615680.

[303] David Mimno and David Blei. — "Bayesian Checking for Topic Models". — In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. — EMNLP '11. — Edinburgh, United Kingdom: Association for Computational Linguistics, 2011, — P. 227—237. — ISBN: 978-1-937284-11-4. — URL: http://dl.acm.org/citation.cfm?id=2145432.2145459.

[304] Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. — "Sharing Clusters Among Related Groups: Hierarchical Dirichlet Processes". — In: *Proceedings of the 17th International Conference on Neural Information Processing Systems*. — NIPS'04. — Vancouver, British Columbia, Canada: MIT Press, 2004, — P. 1385—1392. — URL: http://dl.acm.org/citation.cfm?id=2976040.2976214.

[305] David M. Blei, Thomas L. Griffiths, and Michael I. Jordan. — "The Nested Chinese Restaurant Process and Bayesian Nonparametric Inference of Topic Hierar-

chies". — In: *J. ACM* 57.2 (Feb. 2010), 7:1—7:30. — DOI: 10.1145/1667053.1667056. — URL: http://doi.acm.org/10.1145/1667053.1667056.

[306] David M. Blei, Michael I. Jordan, Thomas L. Griffiths, and Joshua B. Tenenbaum. — "Hierarchical Topic Models and the Nested Chinese Restaurant Process". — In: *Proceedings of the 16th International Conference on Neural Information Processing Systems*. — NIPS'03. — Whistler, British Columbia, Canada: MIT Press, 2003, — P. 17—24. — URL: http://dl.acm.org/citation.cfm?id=2981345.2981348.

[307] Michael Bryant and Erik B. Sudderth. — "Truly Nonparametric Online Variational Inference for Hierarchical Dirichlet Processes". — In: *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 2*. — NIPS'12. — Lake Tahoe, Nevada: Curran Associates Inc., 2012, — P. 2699—2707. — URL: http://dl.acm.org/citation.cfm?id=2999325.2999436.

[308] Marco Rossetti, Fabio Stella, and Markus Zanker. — "Towards Explaining Latent Factors with Topic Models in Collaborative Recommender Systems". — In: *2013 24th International Workshop on Database and Expert Systems Applications*. — IEEE, Sept. 2013. — DOI: 10.1109/DEXA.2013.26. — URL: https://doi.org/10.1109/dexa.2013.26.

[309] David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. — "Automatic Evaluation of Topic Coherence". — In: *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. — HLT '10. — Los Angeles, California: Association for Computational Linguistics, 2010, — P. 100—108. — ISBN: 1-932432-65-5. — URL: http://dl.acm.org/citation.cfm?id=1857999.1858011.

[310] Sergei Koltcov. — "Application of Renyi and Tsallis entropies to topic modeling optimization". — In: *Physica A: Statistical Mechanics and its Applications* 512 (Dec. 2018), p. 1192—1204. — DOI: 10.1016/j.physa.2018.08.050. — URL: https://doi.org/10.1016/j.physa.2018.08.050.

[311] Xin Bing, Florentina Bunea, and Marten H. Wegkamp. — "A fast algorithm with minimax optimal guarantees for topic models with an unknown number of topics". — In: *CoRR* abs/1805.06837 (May 2018).

[312] Zachary C. Lipton. — "The Mythos of Model Interpretability". — In: *Queue* 16.3 (June 2018), 30:31—30:57. — DOI: 10.1145/3236386.3241340. — URL: http://doi.acm.org/10.1145/3236386.3241340.

[313] M. El-Assady, R. Sevastjanova, F. Sperrle, D. Keim, and C. Collins. — "Progressive Learning of Topic Modeling Parameters: A Visual Analytics Framework". — In: *IEEE*

Transactions on Visualization and Computer Graphics 24.1 (Jan. 2018), p. 382—391. — DOI: 10.1109/TVCG.2017.2745080.

[314] Sergey I. Nikolenko, Sergei Koltcov, and Olessia Koltsova. — "Topic modelling for qualitative studies". — In: *Journal of Information Science* 43.1 (July 2016), p. 88—102. — DOI: 10.1177/0165551515617393. — URL: https://doi.org/10.1177/0165551515617393.

[315] Kayhan Batmanghelich, Ardavan Saeedi, Karthik Narasimhan and Sam Gershman. — "Nonparametric Spherical Topic Modeling with Word Embeddings". — In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. — Association for Computational Linguistics, 2016. — DOI: 10.18653/v1/p16-2087. — URL: https://doi.org/10.18653/v1/p16-2087.

[316] Jarvan Law, Hankz Hankui Zhuo, JunHua He, and Erhu Rong. — "LTSG: Latent Topical Skip-Gram for Mutually Improving Topic Model and Vector Representations". — In: *Pattern Recognition and Computer Vision*. — Springer International Publishing, 2018, — P. 375—387. — DOI: 10.1007/978-3-030-03338-5_32. — URL: https://doi.org/10.1007/978-3-030-03338-5_32.

[317] Rajarshi Das, Manzil Zaheer, and Chris Dyer. — "Gaussian LDA for Topic Models with Word Embeddings". —

In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. — Beijing, China: Association for Computational Linguistics, 2015, — P. 795—804. — DOI: 10.3115/v1/P15-1077. — URL: http://aclweb.org/anthology/P15-1077.

[318] Dat Nguyen, Richard Billingsley, Lan Du, and Mark Johnson. — "Improving Topic Models with Latent Feature Word Representations". — In: *Transactions of the Association for Computational Linguistics* 3 (2015), p. 299—313. — URL: http://aclweb.org/anthology/Q15-1022.

[319] Mika V. Mantyla, Maelick Claes, and Umar Farooq. — "Measuring LDA Topic Stability from Clusters of Replicated Runs". — In: *Proceedings of the 12th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement*. — ESEM '18. — Oulu, Finland: ACM, 2018, — 49:1—49:4. — ISBN: 978-1-4503-5823-1. — DOI: 10.1145/3239235.3267435. — URL: http://doi.acm.org/10.1145/3239235.3267435.

[320] V. Mehta, R. S. Caceres, and K. M. Carter. — "Evaluating topic quality using model clustering". — In: *2014 IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*. — Dec. 2014, — P. 178—185. —

DOI: 10.1109/cidm.2014.7008665. — URL: https://doi.org/10.1109/cidm.2014.7008665.

[321] James C. Bezdek. — "Cluster Validity with Fuzzy Sets". — In: *Journal of Cybernetics* 3.3 (1973), p. 58—73. — DOI: 10.1080/01969727308546047. — URL: https://doi.org/10.1080/01969727308546047.

[322] J. C. Dunn. — "Well-Separated Clusters and Optimal Fuzzy Partitions". — In: *Journal of Cybernetics* 4.1 (Jan. 1974), p. 95—104. — DOI: 10.1080/01969727408546059. — URL: https://doi.org/10.1080/01969727408546059.

[323] David L. Davies and Donald W. Bouldin. — "A Cluster Separation Measure". — In: *IEEE Trans. Pattern Anal. Mach. Intell.* 1.2 (Feb. 1979), p. 224—227. — DOI: 10.1109/TPAMI.1979.4766909. — URL: http://dx.doi.org/10.1109/TPAMI.1979.4766909.

[324] Maria Halkidi, Yannis Batistakis, and Michalis Vazirgiannis. — "Clustering Validity Checking Methods: Part II". — In: *SIGMOD Rec.* 31.3 (Sept. 2002), p. 19—27. — DOI: 10.1145/601858.601862. — URL: http://doi.acm.org/10.1145/601858.601862.

[325] Xuanli Lisa Xie and Gerardo Beni. — "A Validity Measure for Fuzzy Clustering". — In: *IEEE Trans. Pattern Anal. Mach. Intell.* 13.8 (Aug. 1991), p. 841—847. — DOI: 10.1109/34.85677. — URL: http://dx.doi.org/10.1109/34.85677.

[326] Peter Rousseeuw. — "Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis". — In: *J. Comput. Appl. Math.* 20.1 (Nov. 1987), p. 53—65. — DOI: 10.1016/0377-0427(87)90125-7. — URL: http://dx.doi.org/10.1016/0377-0427(87)90125-7.

[327] Jeffrey Pennington, Richard Socher, and Christopher Manning. — "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. — URL: http://aclweb.org/anthology/D14-1162.

[328] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. — "Enriching Word Vectors with Subword Information". — In: *Transactions of the Association for Computational Linguistics* 5 (2017), p. 135—146. — URL: http://aclweb.org/anthology/Q17-1010.

[329] Ledell Yu Wu, Adam Fisch, Sumit Chopra, Keith Adams, Antoine Bordes, and Jason Weston. — "StarSpace: Embed All The Things!" — In: *AAAI*. — 2018.

[330] P. V. Bicalho, T. d. O. Cunha, F. H. J. Mourao, G. L. Pappa, and W. Meira. — "Generating Cohesive Semantic Topics from Latent Factors". — In: *2014 Brazilian Conference on Intelligent Systems*. — IEEE, Oct.

2014, — P. 271—276. — DOI: 10.1109/bracis.2014.56. — URL: https://doi.org/10.1109/bracis.2014.56.

[331] Adrian Kuhn, Stephane Ducasse, and Tudor Girba. — "Semantic clustering: Identifying topics in source code". — In: *Information and Software Technology* 49.3 (Mar. 2007), p. 230—243. — DOI: 10.1016/j.infsof.2006.10.017. — URL: https://doi.org/10.1016/j.infsof.2006.10.017.

[332] Jason Chuang, Margaret E. Roberts, Brandon M. Stewart, Rebecca Weiss, Dustin Tingley, Justin Grimmer, and Jeffrey Heer. — "TopicCheck: Interactive Alignment for Assessing Topic Model Stability". — In: *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. — Denver, Colorado: Association for Computational Linguistics, 2015, — P. 175—184. — DOI: 10.3115/v1/N15-1018. — URL: http://aclweb.org/anthology/N15-1018.

[333] Derek Greene, Derek O'Callaghan, and Padraig Cunningham. — "How Many Topics? Stability Analysis for Topic Models". — In: *Machine Learning and Knowledge Discovery in Databases*. — Springer: Berlin, Heidelberg, 2014, — P. 498—513. — DOI: 10.1007/978-3-662-44848-9_32. — URL: https://doi.org/10.1007/978-3-662-44848-9_32.

[334] Sergei Koltcov, Sergey I. Nikolenko, Olessia Koltsova, Vladimir Filippov, and Svetlana Bodrunova. — "Stable Topic Modeling with Local Density Regularization". — In: *Internet Science*. — Springer International Publishing, 2016, — P. 176—188. — DOI: 10.1007/978-3-319-45982-0_16. — URL: https://doi.org/10.1007/978-3-319-45982-0_16.

[335] I. Borg and P. Groenen. — "Modern Multidimensional Scaling: Theory and Applications". — In: *Journal of Educational Measurement* 40.3 (Sept. 2003), p. 277—280. — DOI: 10.1111/j.1745-3984.2003.tb01108.x. — URL: https://doi.org/10.1111/j.1745-3984.2003.tb01108.x.

[336] T. Calinski and J Harabasz. — "A dendrite method for cluster analysis". — In: *Communications in Statistics* 3.1 (1974), p. 1—27. — DOI: 10.1080/03610927408827101. — URL: https://www.tandfonline.com/doi/abs/10.1080/03610927408827101.

[337] Xiaoqiang Lu, Xiangtao Zheng, and Xuelong Li. — "Latent semantic minimal hashing for image retrieval". — In: *IEEE Transactions on Image Processing* 26.1 (2016), p. 355—368.

[338] Jarvan Law, Hankz Hankui Zhuo, Junhua He, and Erhu Rong. — "LTSG: Latent Topical Skip-Gram for mutually learning topic model and vector representations". — In: *arXiv preprint arXiv:1702.07117* (2017).

Author's publications

Author's publications in HAC editions

[141] Fedor Krasnov. — "Model' processa publikacij nauchno-prakticheskih statej po special'nosti 25.00 «Nauki o Zemle»". — In: *Internet-zhurnal «NAUKOVEDENIE»* 9.5 (2017).

[167] F.V. Krasnov. — "Analiz metodov postroenija grafa soavtorstva: podhod na osnove dvudol'nogo grafa". — In: *International Journal of Open Information Technologies* 6.2 (2018), p. 31—37.

[339] Fedor Krasnov and Oleg Ushmaev. — "Razvedka skrytyh napravlenij issledovanij v neftegazovoj otrasli s pomoshh'ju analiza biblioteki OnePetro". — In: *International Journal of Open Information Technologies* 6.5 (2018).

[340] Fedor Krasnov and Mars Khasanov. — "Allocation of the scientific directions of development of science and technologies center in oil and gas industry based on the co-authorship network". — In: *International Journal of Open Information Technologies* 6.4 (2018), p. 1—6.

[341] Fedor Krasnov, Sofia Dokuka, and Rostislav Yavorskiy. — "Team assembly in R&D: A review of imitating modeling approach for science and technology center in Oil&Gaz industry". — In: *International*

Journal of Open Information Technologies 6.1 (2018), p. 17—24.

[342] Irina Kuchma and FV Krasnov. — "Upravlenie znanijami: otkrytye cifrovye obrazovatel'nye resursy i arhivy: materialy mezhdunarodnogo seminara. Jaroslavl', 14 oktjabrja 2011 g." — In: (2011).

[343] FV Krasnov and N.V. Kurchakova. — "Modelirovanie izmenenij intellektual'nogo kapitala v uslovijah povyshennoj nagruzki na personal nauchno-issledovatel'skoj organizacii". — In: *Internet-zhurnal «NAUKOVEDENIE»* 9.6 (2017).

[344] F.V. Krasnov. — "Neftegazorazvedka bez bol'shih dannyh". — In: *Zhurnal "Otkrytye sistemy. SUBD"* 4 (2015).

[345] F.V. Krasnov and S.V. Dokuka. — "Modelirovanie i ocenka vlijanija ot primenenija karkasa SCRUM v processe napisanija nauchnyh statej". — In: *Internet-zhurnal «NAUKOVEDENIE»* 9.6 (2017).

[346] F.V. Krasnov and S.V. Dokuka. — "Verojatnostnaja model' skrytyh tem na osnove arhiva zhurnala «Neftjanoe Hozjajstvo»". — In: *Vestnik Evrazijskoj nauki* 2 (2018).

[347] F.V. Krasnov and I.A. Makarov. — "Prognozirovanie razvitija soavtorstva v napisanii nauchnyh statej nauchno-tehnicheskogo centra Gazpromneft' na osnove modeli". — In: *Vestnik Evrazijskoj nauki* 1 (2018).

[348] F.V. Krasnov, S.V. Doduka, and R.Je. Javorskij. — "Komandoobrazovanija v nauchnoj dejatel'nosti: analiz podhodov na osnovanii imitacionnoj modeli dlja nauchno-tehnicheskogo centra v neftegazovoj otrasli". — In: *International Journal of Open Information Technologies* 6.1 (2018), p. 17—24.

[349] F.V. Krasnov. — "Analiz tonal'nosti teksta nauchno-prakticheskih statej po neftegazovoj tematike s pomoshh'ju iskusstvennyh nejronnyh setej". — In: *Vestnik Evrazijskoj nauki* 3 (2018).

[350] Fedor Krasnov and Alexander Butorin. — "Spectral Inversion in Estimation of Change in the Dominant Frequency of the Wave Field". — In: *International Journal of Open Information Technologies* 7.3 (2019), p. 42—49.

[351] Alexander Butorin and Fedor Krasnov. — "Modern Approaches to Numerical Modeling of Microseismic Events". — In: *International Journal of Open Information Technologies* 7.3 (2019), p. 7—16.

[352] Fedor Krasnov. — "Evaluation of Optimal Number of Topics of Topic Model: An Approach Based on the Quality of Clusters". — In: *International Journal of Open Information Technologies* 7.2 (2019), p. 8—15.

[353] Fedor Krasnov and Alexander Butorin. — "Optimization Methodology for the Selection of Frequencies to Produce an RGB Representation of the Results of Spectral

Decomposition". — In: *International Journal of Open Information Technologies* 6.11 (2018), p. 21—27.

[354] Fedor Krasnov and Oleg Ushmaev. — "Exploration of Hidden Research Directions in Oil and Gas Industry via Full Text Analysis of OnePetro Digital Library". — In: *International Journal of Open Information Technologies* 6.5 (2018), p. 7—14.

[355] Fedor Krasnov, Alexander Butorin, and Alexander Sitnikov. — "Automatic Detection of Channels in Seismic Images via Deep Convolutional Neural Networks Learning". — In: *International Journal of Open Information Technologies* 6.3 (2018), p. 20—26.

[356] Fedor Krasnov, Nikolay Glavnov, and Alexander Sitnikov. — "A review of two algorithms for proxy model of enhanced oil recovery". — In: *International Journal of Open Information Technologies* 5.10 (2017), p. 18—23.

[357] AV Butorin and FV Krasnov. — "Vozmozhnosti ispol'zovanija rezul'tatov spektral'noj inversii pri interpretacii sejsmicheskih dannyh". — In: *Geofizika* 4 (2017), p. 2—7.

[358] FV Krasnov, NG Glavnov, and AN Sitnikov. — "Primenenie mashinnogo obuchenija po ansamblju reshajushhih pravil dlja vychislenija prognoza dopolnitel'nogo kojefficienta izvlechenija nefti". — In: *International Journal of Open Information Technologies* 5.10 (2017).

[359] F. Krasnov, A. Ershov, and A. Margarit. — "CIFROVAJa PLATFORMA RAZVEDKI I DOBYChI UGLEVODORODOV". — In: *Otkrytye sistemy. SUBD.* 2 (2019), p. 36—39.

[360] Fedor Krasnov and Alexander Butorin. — "High Spatial Image Classification Problem: Review of Approaches". — In: *International Journal of Open Information Technologies* 7.4 (2019), p. 6—10.

[361] Mars Khasanov and Fedor Krasnov. — "Transactionality of Digital Transformation within an R&D Organization". — In: *International Journal of Open Information Technologies* 7.5 (2019), p. 39—42.

[362] Fedor Krasnov and Mars Khasanov. — "Digital Twin for R&D organization: approaches and methods". — In: *International Journal of Open Information Technologies* 7.6 (2019), p. 62—66.

[363] FV Krasnov, AV Butorin, and AN Sitnikov. — "Obzor podhodov k analizu prostranstvennyh izobrazhenij vysokogo razreshenija dlja primenenija v geofizike". — In: *Cloud of Science* 6.1 (2019), p. 127—143.

[364] FV Krasnov, MM Hasanov, AV Dimentov, and ME Shvarcman. — "Cravnenie soderzhanija kollekciĭ nauchnyh zhurnalov na osnove razrabotannyh tematicheskih modeleĭ i metodiki T4C". — In: *Cloud of science* 6.3 (2019).

[365] A.V. Butorin and F.V. Krasnov. — "METODIKA OCENKI IZMENENIJa DOMINANTNOGO ZNAChENIJa ChASTOTY VOLNOVOGO POLJa VDOL" SEJSMIChESKOJ TRASSY". — In: *Geofizika* 4 (2018), p. 33—39.

[366] Fedor Vladimirovich Krasnov, Alexander Vasilievich Butorin, and Andrey Vital'evich Mikheyenkov. — "Reconstruction of medium reflectivity coefficients based on seismic data through machine learning". — In: *Nauchno-tehnicheskie vedomosti Sankt-Peterburgskogo gosudarstvennogo politehnicheskogo universiteta. Informatika. Telekommunikacii. Upravlenie* 11.1 (2018).

[367] FV Krasnov, MM Hasanov, RM Galeev, and AM Margarit. — "PRINCIPY POSTROENIJa CIFROVOJ PLATFORMY DLJa NAUChNO-TEHNIChESKOGO CENTRA". — In: *Vestnik kibernetiki* 4 (2019), p. 66—73.

[368] F.V. Krasnov. — "TRANZAKCIONNOST" CIFROVOJ TRANSFORMACII V NAUChNOJ ORGANIZACII". — In: *PRONEFT". Professional'no o nefti.* 1 (11) (2019), p. 64—67.

Author's publications in WoS and Scopus editions

[122] Fedor Krasnov, Sofia Dokuka, and Rostislav Yavorskiy. — "The Structure of Organization: The

Coauthorship Network Case". — In: *International Conference on Analysis of Images, Social Networks and Texts*. — Springer. 2016, — P. 100—107.

[139] Fedor Krasnov and Rostislav Yavorskiy. — "Measurement of maturity level of a professional community". — In: *Business Informatics* 23.1 (2013), p. 64—67.

[166] Fedor Krasnov, Sofia Dokuka, Ilya Gorshkov, and Rostislav Yavorskiy. — "Analysis of Strong and Weak Ties in Oil & Gas Professional Community". — In: *CEUR Workshop Proceedings Ser. "Proceedings of International Workshop on Formal Concept Analysis for Knowledge Discovery, FCA4KD 2017"*. — CEUR Workshop Proceedings. 2017, — P. 22—33.

[184] Fedor Krasnov, Evgeniya Vlasova, and Rostislav Yavorskiy. — "Connectivity Analysis of Computer Science Centers based on Scientific Publications Data for Major Russian Cities". — In: *Procedia Computer Science* 31 (2014), p. 892—899.

[185] Fedor Krasnov, Dmitriy Ustalov, and Rostislav Yavorskiy. — "Comparison of online communities on the base of lexical analysis of the news feed". — In: *Proceedings of 2-nd conference on Analysis of Images, Networks and Texts, Yekaterinburg*. — 2013, — P. 254—257.

[186] Fedor Krasnov, Rostislav E Yavorskiy, and Evgeniya Vlasova. — "Indicators of connectivity for urban

scientific communities in Russian cities". — In: *International Conference on Analysis of Images, Social Networks and Texts*. — Springer. 2014, — P. 111—120.

[369] Nikita Klyuchnikov, Fedor Krasnov, and Alexey Zaytsev. — "Data-driven model for the identification of the rock type at a drilling bit". — In: *Journal of Petroleum science and Engineering* (2019).

[370] Fedor Krasnov and Alexander Butorin. — "High-Resolution Seismic Data Deconvolution by A0 Algorithm". — In: *Geosciences* 8.12 (2018), p. 497.

[371] Fedor Krasnov, Nikolay Glavnov, and Alexander Sitnikov. — "A Machine Learning Approach to Enhanced Oil Recovery Prediction". — In: *Lecture Notes in Computer Science*. — Springer International Publishing, 2017, — P. 164—171.

[372] Fedor Krasnov, Rostislav E Yavorskiy, and Evgeniya Vlasova. — "Indicators of Connectivity for Urban Scientific Communities in Russian Cities". — In: *Analysis of Images, Social Networks and Texts*. — Springer, 2014, — P. 111—120.

[373] Fedor Krasnov, Alexander Dimentov, and Mikhail Shvartsman. — "Comparative Analysis of Scientific Papers Collections via Topic Modeling and Co-authorship Networks". — In: *Conference on Artificial Intelligence and Natural Language*. — Springer. 2019, — P. 77—98.

[374] FV Krasnov, ME Shvartsman, AV Dimentov, and AI Sen. — "A Thematic Coherence Study of a Bilingual Corpus of Articles on Oil and Gas Research". — In: *Automatic Documentation and Mathematical Linguistics* 53.3 (2019), p. 138—142.

[375] FV Krasnov, AV Butorin, and AN Sitnikov. — "Avtomatizirovannoe obnaruzhenie geologicheskih ob"ektov v izobrazhenijah sejsmicheskogo polja s primeneniem nejronnyh setej glubokogo obuchenija". — In: *Biznes-informatika* 2 (44) (2018).

[376] FV Krasnov, TV Voznesenskaja, RJe Javorskij, and PV Chesnokova. — "Modelirovanie samoorganizujushhihsja komand v nauchnoj srede". — In: *Biznes-informatika* 13.2 (2019).

[377] Fedor Krasnov, Nikolay Glavnov, and Alexander Sitnikov. — "Application of multidimensional interpolation and random forest regression to enhanced oil recovery modeling". — In: *Proceedings of the 13th Central and Eastern European Software Engineering Conference in Russia on - CEE-SECR '17*. — ACM Press, 2017.

[378] Fedor Krasnov, Alexander Butorin, and Alexander Sitnikov. — "MODERN APPROACHES TO NUMERICAL MODELING OF MICROSEISMIC EVENTS". — In: *GEOPHYSICAL RESEARCH* 20.2 (2019), p. 39—55.

[379] FV Krasnov, ME Shvarcman, and AV Dimentov. — "SRAVNITEL"NYJ ANALIZ KOLLEKCIJ NAUChNYH ZhURNALOV". — In: *Trudy SPIIRAN* 18.3 (2019), p. 766—792.

[380] Fedor Krasnov. — "Topic Classification Through Topic Modeling with Additive Regularization for Collection of Scientific Papers". — In: *Proceedings of the 14th Central and Eastern European Software Engineering Conference Russia*. — ACM. 2018, — P. 5.

[381] Fedor Krasnov and Mars Khasanov. — "Unsupervised Co-Authorship Based Algorithm for Clustering of R&D Trends at Science and Technology Centers in Oil and Gas Industry." — In: *AIST (Supplement), CEUR Workshop Proceedings*. — 2018, — P. 1—12.

[382] Fedor Krasnov and Vladimir Lebedev. — "Clustering of Translation via Topic Modeling". — In: *Journal of Physics: Conference Series*. — Vol. 1405. — 1. — IOP Publishing. 2019, — P. 012008.

[383] F. Krasnov and A. Sen. — "THE NUMBER OF TOPICS OPTIMIZATION: CLUSTERING APPROACH". — In: *CEUR Workshop Proceedings Ser. "MACSPro 2019 - Proceedings of the Modeling and Analysis of Complex Systems and Processes Workshop 2019"*. — 2019, — P. 1—15.

Author's publications in peer-reviewed journals

[384] F.V. Krasnov. — "Chelovek i kommunikacii". — In: *Direktor informacionnoj sluzhby* 11 (2008).

[385] AV Butorin, FV Krasnov, et al. — "Numerical Modeling of Microseismic Events on the Surface". — In: *SPE Russian Petroleum Technology Conference*. — Society of Petroleum Engineers. 2017.

[386] AV Butorin, FV Krasnov, et al. — "Spectral Inversion Methods and its Application for Wave Field Analysis (Russian)". — In: *SPE Russian Petroleum Technology Conference*. — Society of Petroleum Engineers. 2017.

[387] FV Krasnov and AN Sitnikov. — "Ispol'zovanie serializacii dlja hranenija geologo-geofizicheskoj informacii". — In: *Nauchno-teoreticheskij zhurnal* (2015), p. 25.

[388] FV Krasnov. — "Razvitie cherez obshhenie". — In: *Intelligent Enterprise* 9 (2012), p. 18—21.

[389] Fedor Krasnov and Anastasiia Sen. — "The Number of Topics Optimization: Clustering Approach". — In: *Machine Learning and Knowledge Extraction* 1.1 (2019), p. 416—426.

[390] Mars Khasanov, Fedor Krasnov, et al. — "Digital Twin of a Research Organization: Approaches and Methods". — In: *SPE Annual Caspian Technical Conference*. — Society of Petroleum Engineers. 2019.

Author's patents

[1] F.V. Krasnov. — “Programma dlja JeVM «SPO "NauBot"»”. — 2017661666. — 2017.

[2] F.V. Krasnov and A.V. Butorin. — “KSPO "HiRGB"”. — 2017661769. — 2017.

[3] F.V. Krasnov and A.V. Butorin. — “Jekspertnaja Sistema dlja ocenki sejsmicheskih neopredelennostej”. — 2018617693. — 2018.

[4] A.V. Butorin, F.V. Krasnov, and D.G. Murtazin. — “PROGRAMMA "SPECTRA CLUSTER" (SPEKTRA KLASTER) DLJa RASChJoTA KUBA SPEKTRAL”NYH KRIVYH I POSLEDUJuShhEJ EGO KLASTERIZACII”. — 2019664988. — 2019.

Author's publications in proceedings

[170] MM Khasanov, FV Krasnov, BV Belozerov, et al. — “Corporate Wikipedia in Upstream: Bimodal IT Case”. — In: *SPE Annual Technical Conference and Exhibition*. — Society of Petroleum Engineers. 2016.

[391] Fedor Krasnov and Alexander Sergeev. — “Segmentation of IT customers on internal market”. — In: *SEC(R)'09. The 5th Software Engineering Conference (Russia)*. — TEKAMA. 2009.

[392] Rostislav Jeduardovich Javorskij, Fedor Vladimirovich Krasnov, and Dmitrij Ustalov. — “Sravnenie onlajn-soobshhestv na osnove leksicheskogo analiza lenty novostej”. — In: *Doklady vserossijskoj nauchnoj konferencii AIST’2013*. — Nacional’nyj otkrytyj universitet «INTUIT». 2013, — P. 242—245.

[393] Aleksandr Vasil’evich Butorin and Fedor Vladimirovich Krasnov. — “Primenenie metodov spektral’noj inversii”. — In: *Sejsmicheskie tehnologii-2017*. — 2017, — P. 192—195.

[394] AV Butorin and FV Krasnov. — “Approaches to the Analysis of Spectral Decomposition for the Purpose of Detailed Geological Interpretation”. — In: *SPE Russian Petroleum Technology Conference and Exhibition*. — Society of Petroleum Engineers. 2016.

[395] AV Butorin and FV Krasnov. — “Sravnitel’nyj analiz metodov spektral’noj inversii volnovogo polja na primere model’nyh trass”. — In: *Geofizika* 4 (2016), p. 68—76.

[396] V.A. Fagereva, A.V. Butorin, and F.V. Krasnov. — “PRAKTIChESKOE PRIMENENIE REZUL”TATOV SPEKTRAL”NOJ INVERSII”. — In: *V sbornike: Novye idei v naukah o Zemle Materialy XIV Mezhdunarodnoj nauchno-prakticheskoj konferencii: v 7-mi tomah*. — 2019, — P. 220—223.

List of figures

List of tables

Appendix A. Large figures

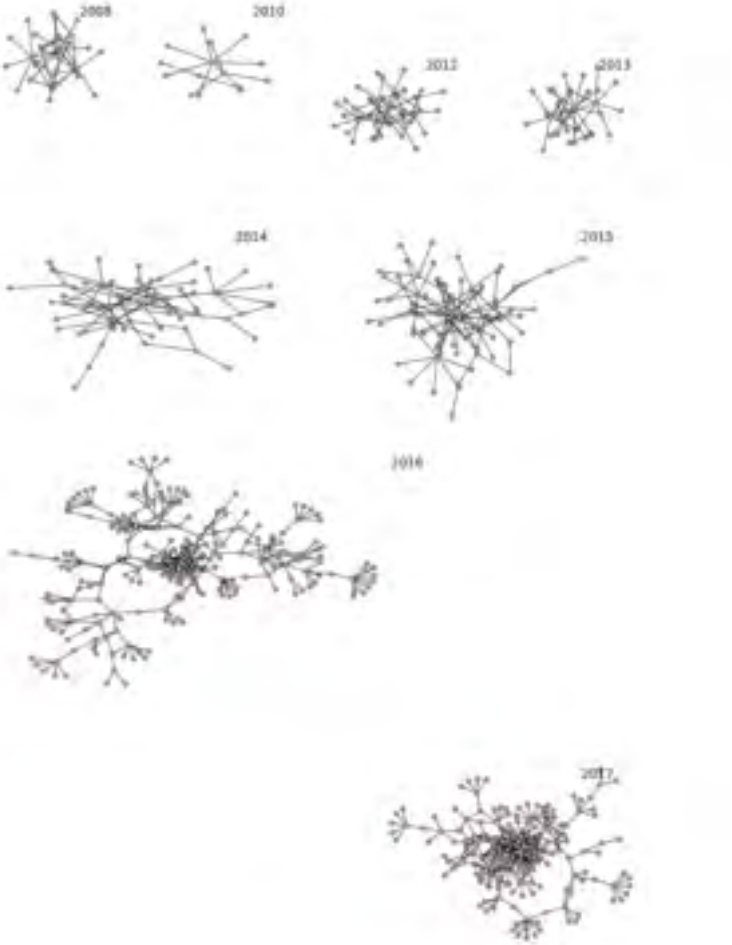


Figure A.1 — The co-authorship development growth dynamic by year graph.

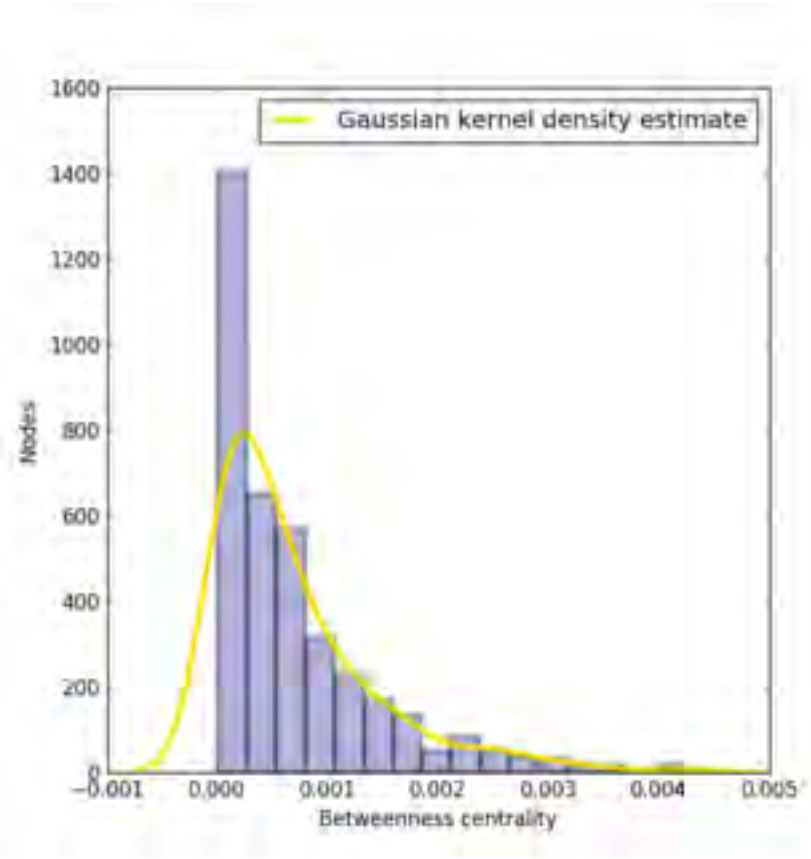


Figure A.2 — Histogram of Betweenness centrality values for the subgraph of the strongest connected component of the co-authorship graph of Gazpromneft STC.

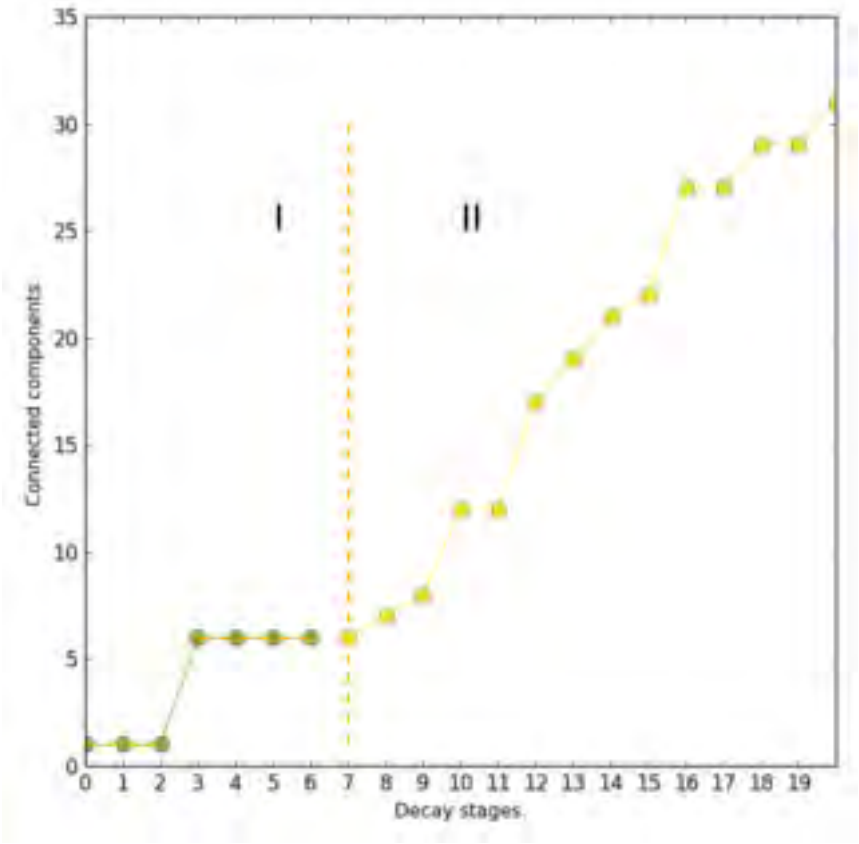


Figure A.3 — Correlation between the connected components number and the number of artificially removed nodes.

Figure A.4 — The cluster of researchers into *Subject 1* extracted through the method of removal of the nodes with the highest values of the Betweenness centrality metrics.

Figure A.5 — Visualization of the strong ties in Oil&Gas professional community.

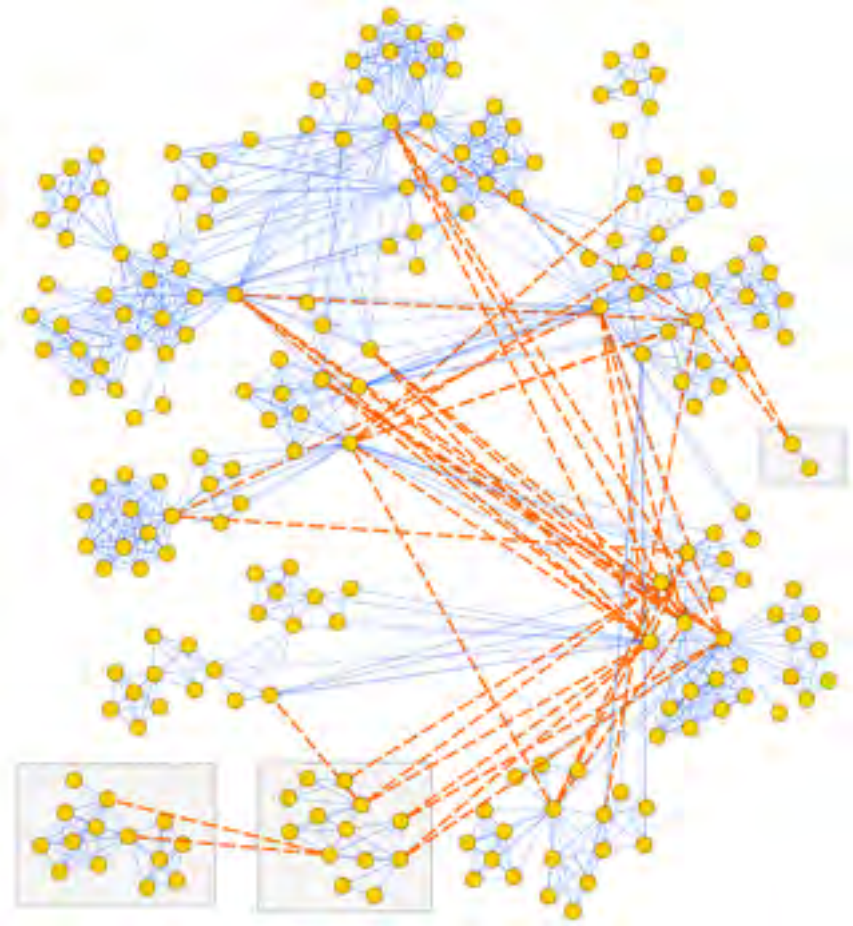


Figure A.6 — Visualization of the largest connected component with the weak ties.

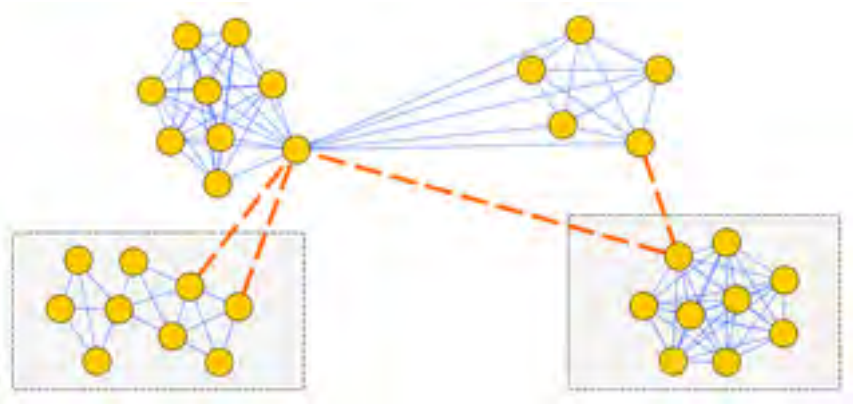


Figure A.7 — Second largest connected component with the weak ties (dashed red). Grey boxes are used to set out previously disconnected fragments, which get bridged with the weak ties.

Figure A.8 — Graph of the new identified weak ties.

List of figures

List of tables

Appendix B. Code listing fragment

Listing of program code for the task of optimizing the number of topics in a topic model.

Listing B.1 Source code fragment

```
# -*- coding: utf-8 -*-
import numpy as np
import artm
from sklearn.metrics.pairwise import cosine_distances
from sklearn.metrics import silhouette_score,
    davies_bouldin_score, calinski_harabaz_score
import os
import sys
import warnings
warnings.filterwarnings("ignore")

wrk = sys.argv[1]

gdim = 100
vecfile = open(wrk+'.model.vec','r')
vtext = vecfile.read().split("\n")
vecfile.close()

vecs = dict()
for l in vtext:
  if len(l) < 1 : continue
  w = l.split()[0]
  v = l.split()[1:]
  vecs[w]=[float(i) for i in v]

text = open(wrk+".vw",'r').read().split("\n")
NR = 10

batch_vectorizer = artm.BatchVectorizer(data_path=wrk+".vw"
    , \
```

```
data_format="vowpal_wabbit", target_folder=wrk, batch_size
    =100)

dictionary = artm.Dictionary()
dictionary.gather(data_path=wrk)

metrics = [] #Все метрики по порядку.
for nr in range(NR):
  txt_ind = np.random.randint(0,len(text)-1,len(text))
  text_new = "\n".join([text[i] for i in txt_ind])
  fh = open(wrk+".vw%d" % nr,'w')
  fh.write(text_new)
  fh.close()

  batch_vectorizer = artm.BatchVectorizer(data_path=wrk+".
    vw%d" % nr, \
    data_format="vowpal_wabbit", target_folder=wrk,
    batch_size=100)

  os.remove(wrk+".vw%d" % nr)

  for T in range(5,20,5) + range(20,40) + range(40,71,5):
    N = 2
    ttopics = ["sbj"+str(i) for i in range(T)]
    ntopics = ["nz"+str(i) for i in range(N)]
    topic_names= ttopics + ntopics
    model_artm = artm.ARTM(num_topics=T+N, topic_names=
    topic_names)
    model_artm.initialize(dictionary)
    #model_artm.cache_theta = True
    #model_artm.cache_phi = True
    model_artm.scores.add(artm.TopicKernelScore(name="
    top_score", topic_names=ttopics,
    probability_mass_threshold=0.5))
    model_artm.fit_offline(batch_vectorizer=
    batch_vectorizer, num_collection_passes=10)
    model_artm.regularizers.add(artm.
    SmoothSparseThetaRegularizer(name='SparseTheta',tau=-1,
    topic_names=ttopics))
```

```
    model_artm.regularizers.add(artm.
    SmoothSparseThetaRegularizer(name='SmoothTheta',tau=1,
    topic_names=ntopics))
    model_artm.regularizers.add(artm.
    SmoothSparsePhiRegularizer(name='SparsePhi',tau=-1,
    topic_names=ttopics))
    model_artm.regularizers.add(artm.
    SmoothSparsePhiRegularizer(name='SmoothPhi',tau=1,
    topic_names=ntopics))
    model_artm.regularizers.add(artm.
    DecorrelatorPhiRegularizer(name='Decorrelation',tau=-1
    e3,topic_names=topic_names))
    model_artm.fit_offline(batch_vectorizer=
    batch_vectorizer, num_collection_passes=10)

    ts = model_artm.score_tracker['top_score']
    tls = ts.last_tokens
    #glove id
    y = []# для силуэта
    X = []
    glove_vec_topic = {}
    for topic_name in  ttopics:
      for w in tls[topic_name]:
        if w.encode('utf8') in vecs:
          gv = vecs[w.encode('utf8')]
          y.append(topic_name)
          X.append(gv)
          if topic_name in glove_vec_topic:
            glove_vec_topic[topic_name].append(gv)
          else:
            glove_vec_topic[topic_name] = [gv]
        else:
          pass #print w

    # centroids
    centroids = np.zeros((T,gdim))
    intra_dists = np.zeros(T)
    inter_topic_dists_list = []
    for tid,topic_name in enumerate(glove_vec_topic):
      centroids[tid] = np.mean(np.array(glove_vec_topic[
    topic_name]),axis=0)
```

```
      #inter topic dists
      glen = len(glove_vec_topic[topic_name])
      intra_dists[tid] = np.average( cosine_distances(
    glove_vec_topic[topic_name], [ centroids[tid] * glen
    ]).ravel())
      #inter_topic_dists += cosine_distances(
    glove_vec_topic[topic_name], [ centroids[tid] * glen
    ]).ravel().tolist() # dist with centroid

  extra_dists = cosine_distances(centroids,centroids)
  #extra_topic_dists = extra_topic_dists[
  extra_topic_dists != 0]

  #inter_topic_dists = np.array(inter_topic_dists_list)
  #inter_topic_dists = inter_topic_dists[
  inter_topic_dists != 0]

  score = (intra_dists[:, None] + intra_dists) /
  extra_dists
  score[score == np.inf] = np.nan
  cdbi = np.mean(np.nanmax(score, axis=1))

  metrics.append ((nr, T,  max(ts.average_contrast), max(
  ts.average_purity), max(ts.average_size),\
    silhouette_score(X,y,metric='cosine',random_state=42)
  , davies_bouldin_score(X,y), calinski_harabaz_score(X,y
  ) , cdbi))

  print metrics[-1]

np.save(wrk+'_metrics.npy',  np.array(metrics))
```