Федеральное государственное автономное образовательное учреждение
высшего образования
«Московский физико-технический институт
(национальный исследовательский университет)»
Физтех-школа Прикладной Математики и Информатики
Кафедра машинного обучения и цифровой гуманитаристики

Направление подготовки / специальность: 09.04.01 Информатика и вычислительная техника
Направленность (профиль) подготовки: Прикладная математика и информатика

МЕТОДЫ ПОСТРОЕНИЯ ХРОНОЛОГИЧЕСКИХ ТЕМАТИЧЕСКИХ КАРТ

(магистерская диссертация)

Студент:
Павлова Елена Денисовна

_____
(подпись студента)

Научный руководитель:
Воронцов Константин Вячеславович,
д-р физ.-мат. наук

_____
(подпись научного руководителя)

Консультант *(при наличии)***:**
_____
_____
(подпись консультанта)

Москва 2026

Содержание

1 Аннотация

В работе рассматривается задача автоматического построения хронологических тематических карт научных областей на основе коллекций научных публикаций с целью анализа тематической структуры и временной динамики развития научного знания. В качестве эталонного примера используется карта наук о сложности Б. Кастеллани, представляющая собой экспертное иерархическое описание области исследований. Работа направлена на исследование возможности автоматического воспроизведения подобных тематических структур для различных научных областей и на разработку методов количественного сравнения автоматически построенных тематических моделей с экспертной картой.

Для проведения исследования был сформирован корпус научных публикаций по тематике науки о сложности, собранный вручную на основе структуры карты Кастеллани, тематических направлений и связанных с ними исследователей. Научная область представляется в виде иерархической тематической структуры, включающей тематические уровни, временные характеристики и связи между исследовательскими направлениями. Для построения тематической карты используются методы тематического моделирования, обработки научных текстов и визуализации тематических структур.

В работе реализован программный конвейер обработки научных публикаций, включающий извлечение текста из PDF-документов, очистку текстов и формирование корпуса для тематического анализа. Для построения тематической модели используется иерархическое тематическое моделирование с несколькими уровнями тем. Дополнительно реализована временная компонента модели, позволяющая учитывать динамику развития научных направлений и строить хронологическую тематическую карту.

Качество полученной тематической структуры оценивается путём сравнения с экспертной картой Кастеллани с использованием предложенной метрики оценки согласованности тематических кластеров. Сравнение выполняется отдельно для различных уровней тематической иерархии и позволяет оценить степень соответствия автоматически выделенных тематических направлений экспертной структуре научной области.

Результатом работы является программный конвейер автоматического построения хронологических тематических карт научных публикаций, включающий этапы формирования корпуса, тематического моделирования, анализа временной структуры и визуализации тематических направлений.

2 Обзор литературы

2.1 Картографирование знаний и научных областей

Задача автоматического картографирования научных областей напрямую относится к области картографирования знаний (science mapping), поскольку этот подход рассматривает научное знание как сложную, структурированную и развивающуюся систему, поддающуюся анализу и визуализации на основе массива научных публикаций.

Картографирование знаний представляет собой подход к анализу и визуализации научных областей, ориентированный на выявление их структуры, динамики и направлений развития [1]. В качестве объекта исследования рассматривается научная область или отдельное исследовательское направление, которое можно описать как развивающийся домен знаний. Такая область включает темы, концепции, ключевые слова и связи между ними. Со временем эти элементы могут изменяться как постепенно, так и существенно.

Работы, посвящённые картографированию научного знания, как правило, опираются на анализ массивов научных публикаций с использованием библиометрических, наукометрических и методов визуальной аналитики. Постоянное увеличение объёма научной литературы существенно затрудняет её детальное ручное изучение, вследствие чего подобные подходы становятся необходимым инструментом для систематизации и навигации в крупных информационных массивах [2]. Применение таких методов позволяет выделять устойчивые тематические направления, фиксировать появление новых исследовательских областей и отслеживать изменения в структуре научного знания. Карты науки используются не только для описания текущего состояния исследуемой области, но и для анализа закономерностей её развития, интерпретация которых основывается на представлениях о динамике и эволюции научных направлений.

Основная цель картографирования знаний заключается в том, чтобы сделать сложные и многомерные области научных исследований обозримыми, поддерживая аналитическое исследование, ориентацию в научной литературе и принятие решений в сфере научной политики и управления исследованиями. Визуальные карты позволяют выявлять связи и закономерности, которые трудно обнаружить при линейном анализе текстов или использовании отдельных показателей, тем самым обеспечивая системное понимание структуры и динамики научных областей.

Карты науки применяются для ответа на широкий круг вопросов: какие основные темы формируют ту или иную научную область, как эти темы связаны между собой, каким образом область развивается во времени и какие основные источники, авторы и центры научной активности формируют данное исследовательское пространство. С их помощью анализируются взаимосвязи между структурными элементами науки на микро- и макроуровнях, факторы возникновения новых научных направлений и формирования междисциплинарных областей, а также более общие механизмы научных изменений [3].

2.2 Классификация и основы построения научных карт

Научные карты строятся в зависимости от того, какие данные используются для их построения и какие отношения между элементами науки при этом моделируются. В общем виде любая характеристика научной деятельности, допускающая представление в виде взаимосвязей между объектами, может служить основой для построения научной карты. В одних случаях в качестве элементов выступают публикации, авторы или журналы, а связи между ними формируются на основе цитирования и производных от него показателей. В других — объектом анализа становятся термины и тематические единицы, связанные между собой через совместное употребление в научных текстах. Отдельный класс карт описывает структуры научного сотрудничества, где узлами являются исследователи, а связи отражают соавторство [4]. Кроме того, существуют карты, основанные на институциональных и редакционных связях между журналами, а также подходы, использующие патентные данные и географическую информацию для анализа пространственного распределения научной активности [5].

Подходы к построению карт научных областей можно разделить на несколько основных групп. Такое разделение связано с типом используемых данных и способом представления научного знания (Таблица 1).

Одним из подходов к классификации карт науки является разделение по типу данных, используемых при их построении. В рамках данного подхода выделяются библиометрические карты [6]. Они формируются на основе формальных связей между научными публикациями, включая цитирование, соавторство, библиографическое сопряжение и совместное использование ключевых слов. Подобные карты позволяют исследовать структуру научных коммуникаций, выявлять наиболее влиятельные публикации и анализировать взаимодействие между исследовательскими сообществами.

Другую группу составляют текстовые карты [7]. В данном случае основное внимание уделяется содержанию научных публикаций. Анализируются тексты статей, аннотации, ключевые слова и другие текстовые представления документов, на основе которых выделяется тематическая структура исследовательской области. Такой подход позволяет исследовать не только связи между публикациями, но и непосредственно содержание научных направлений, а также их тематическую близость.

Отдельную категорию составляют гибридные карты, объединяющие библиометрические и текстовые методы анализа. В подобных моделях одновременно используются как формальные связи между публикациями, так и семантическая информация, извлекаемая из текстов. Это позволяет более полно описывать структуру научной области, сочетая анализ научных взаимодействий с анализом тематического содержания публикаций.

Карты науки также различаются по способу представления структуры исследовательской области. Иерархические карты организуют научное знание в виде нескольких уровней [8]. Верхние уровни обычно содержат крупные исследовательские направления, тогда как нижние уровни включают более специализированные темы, подтемы и отдельные исследовательские области. Подобная структура позволяет анализировать научную область одновременно на разных уровнях детализации.

Сетевые карты представляют научную область в виде графа взаимосвязанных объектов. В таких моделях основное внимание уделяется анализу связей между публикациями, авторами, организациями или тематическими направлениями. Узлы графа соответствуют объектам научной системы, а рёбра отражают различные типы связей между ними. Подобные представления широко используются для исследования структуры научных сообществ и междисциплинарных взаимодействий.

Пространственно-метафорические карты используют геометрические и топографические представления научной области [5]. В подобных моделях исследовательские направления визуализируются как пространство или ландшафт, а расстояние между объектами отражает степень их тематической близости. Такой подход позволяет формировать интуитивно понятные визуальные представления структуры научного знания.

Отдельным направлением классификации является временной аспект картографирования. В этом случае карты подразделяются на статические и динамические. Статические карты отражают структуру научной области в определённый момент времени и позволяют анализировать текущее состояние исследовательских направлений. Динамические или хронологические карты показывают изменение тематической структуры научной области во времени [9]. Они позволяют исследовать процессы возникновения новых тематик, развитие научных направлений и изменение взаимосвязей между ними. Для анализа научных трендов и исследования эволюции научного знания наибольший интерес представляют именно динамические тематические карты.

Основание классификации	Типы карт
Источник и тип данных	• **Библиометрические карты:** на основе цитирований, соавторства, журналов и институтов • **Текстовые карты:** на основе тем, терминов и ключевых слов • **Гибридные карты:** комбинация библиометрических и текстовых данных
Тип представляемой структуры	• **Иерархические карты:** уровни от крупных направлений к темам и подтемам • **Сетевые карты:** графы объектов и их связей • **Пространственно-метафорические карты:** визуализация как ландшафты, поля или топографии
Временной аспект	• **Статические карты:** отражают состояние науки в один момент времени • **Хронологические (динамические) карты:** показывают эволюцию тематических кластеров и развитие исследовательских направлений во времени

Таблица 1: Классификация подходов к построению карт науки

2.3 Карта Кастеллани

Одним из наиболее известных примеров комплексной визуализации научной области является карта наук о сложности Брайана Кастеллани [10] — обобщающая хронологическая тематическая карта, которая сочетает в себе сразу несколько подходов к картографированию науки. Эта карта объединяет несколько ключевых характеристик, описанных выше:

- Иерархическая карта, поскольку организует научное знание в виде пяти основных интеллектуальных традиций, с выделением ключевых тем и методов внутри каждой традиции.
- Текстово-библиометрическая (гибридная) карта, так как опирается на экспертную оценку значимости публикаций и тем, а также на их тематическую принадлежность.
- Хронологическая карта, поскольку показывает, когда каждая тема стала важной областью исследования и как она развивалась во времени.

Следовательно, карта Кастеллани объединяет сразу несколько подходов к классификации научных карт, включая иерархическую организацию структуры, использование комбинированных источников данных, а также учёт временной динамики исследовательских направлений.

На рисунке 1 представлена карта наук о сложности, разработанная Брайаном Кастеллани. В её основе выделены пять ключевых интеллектуальных традиций (intellectual traditions):

1. Dynamical Systems Theory (Теория динамических систем, фиолетовый цвет)
2. Systems Theory (Системная теория, голубой цвет)
3. Complex Systems Theory (Теория сложных систем, светло-жёлтый цвет)
4. Cybernetics (Кибернетика, светло-сиреневый цвет)
5. Artificial Intelligence (Искусственный интеллект, жёлтый цвет)

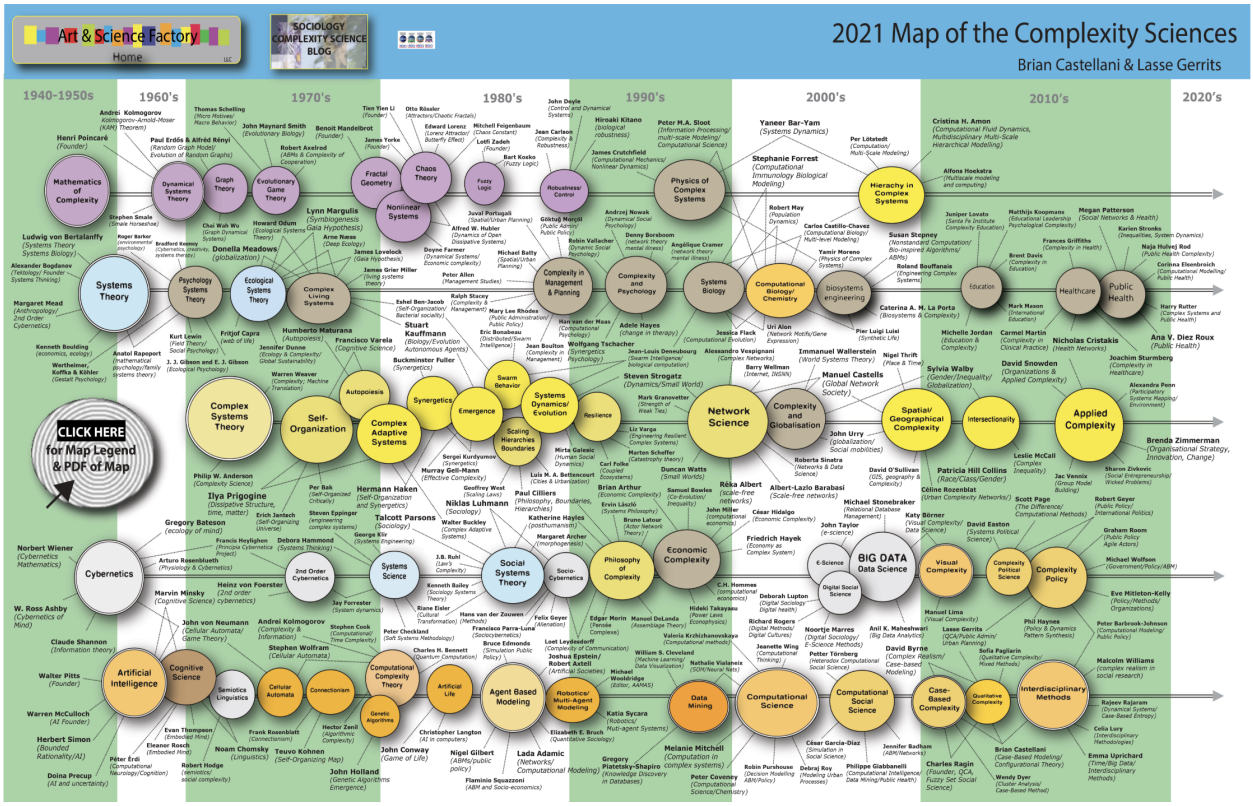

Рис. 1: Карта Кастеллани

Каждая исследовательская традиция на карте представлена собственной цветовой схемой, которая используется для обозначения связанных тематик и применяемых методов исследования. Вокруг пяти базовых традиций группируются ключевые научные направления. Цветовое оформление конкретной темы отражает её доминирующую связь с одной из традиций. Вместе с тем отдельные направления одновременно соотносятся с несколькими

областями, что демонстрирует междисциплинарную природу рассматриваемых исследований.

Одним из центральных элементов карты является временная шкала. Расположение тем вдоль неё соответствует периоду, в который соответствующее направление приобрело заметное значение в научной среде. Такой подход позволяет проследить хронологию формирования и развития отдельных исследовательских областей. Линии, отходящие от тематических узлов, связывают направления с учёными, оказавшими существенное влияние на их становление и дальнейшее развитие. Для отображения дополнительных характеристик и взаимосвязей на карте также используются вспомогательные визуальные элементы.

- Темы, выделенные коричневым цветом, показывают применение содержимого или методов конкретного направления в других областях.
- Двойные круги обозначают пересечение исследовательской традиции с новым научным направлением. Такие элементы показывают появление новых областей исследования и развитие существующих тем.

Карта Кастеллани используется для представления структуры науки о сложности через систему исследовательских традиций и связанных с ними тематических направлений. Помимо этого, карта отражает эволюцию научных тем, демонстрирует междисциплинарные взаимосвязи между различными областями исследований и позволяет учитывать вклад отдельных учёных в формирование и развитие соответствующих направлений.

2.4 Основные подходы построения карт знаний

2.4.1 Библиометрические методы

В рамках библиометрических исследований значительное внимание уделяется методам построения научных карт, основанным на анализе связей между публикациями, авторами и исследовательскими организациями [11]. Использование подобных подходов позволяет изучать внутреннюю структуру научных областей, выявлять устойчивые связи между исследованиями и отслеживать развитие отдельных направлений науки.

Одним из распространённых методов является анализ цитирования, применяемый для оценки научного влияния публикаций и исследователей. В данном случае число цитирований рассматривается как один из индикаторов значимости работы. Co-citation анализ ориентирован на выявление публикаций, которые одновременно цитируются в других исследованиях. Наличие таких совместных цитирований обычно свидетельствует о тематической близости работ или наличии общей теоретической базы.

Подход bibliographic coupling основан на сопоставлении списков литературы различных публикаций. Если работы содержат одинаковые библиографические источники, это интерпретируется как признак их содержательной близости. Данный метод был предложен Кесслером[12] и впоследствии получил широкое распространение в исследованиях, связанных с наукометрическим анализом.

Отдельную группу методов составляют сети соавторства, в которых узлы соответствуют исследователям, а связи формируются на основе совместных публикаций. Анализ подобных структур позволяет исследовать организацию научных сообществ и особенности взаимодействия между авторами.

Кроме того, применяются модели институциональных связей, ориентированные на анализ сотрудничества между научными организациями. В таких подходах учитываются совместные публикации, а также взаимодействия между журналами, например через составы редакционных коллегий. Подобные методы используются для изучения структуры научного сотрудничества и особенностей организации исследовательской деятельности [13].

Для представления библиометрических сетей широко используются методы визуализации, совмещающие построение карт и процедуры кластеризации. Это позволяет одновременно отображать объекты в общем пространстве и группировать их по степени тематической или структурной близости. В результате научная область может быть представлена в виде единой системы взаимосвязанных исследовательских направлений.

2.4.2 Кластеризация и визуализация библиометрических сетей

Библиометрические сети, построенные на основе цитирований, соавторства и совместной встречаемости терминов, обычно имеют сложную структуру. Внутри таких сетей формируются группы тесно связанных публикаций и авторов, которые соответствуют отдельным тематическим направлениям научной области [14]. Для выявления подобных структур применяются методы кластеризации. Наиболее распространёнными являются алгоритмы оптимизации модульности [15], методы поиска сообществ [16] и подходы VOS [17].

Кластеризация позволяет выделять группы публикаций, авторов или терминов с высокой внутренней связностью. Такие группы обычно соответствуют отдельным исследовательским направлениям. Методы визуализации используются для представления этой структуры в виде карты, где расстояние между узлами отражает степень тематической близости объектов.

Методы кластеризации и визуализации тесно связаны между собой. Качественная визуальная карта позволяет выделять отдельные тематические кластеры и анализировать структуру научной области в целом [18].

Одним из наиболее известных и подробно разработанных подходов является унифицированный метод построения и кластеризации библиометрических сетей, предложенный Лудо Уолтманом, Нисом Яном ван Экком и Э. Нойонсом [19]. В данной работе авторы показывают, что задачи визуализации и кластеризации могут быть описаны в рамках единого математического подхода. Такой подход обеспечивает согласованность между расположением узлов на карте и их распределением по тематическим кластерам. Различие между картографированием и кластеризацией определяется способом задания расстояния между узлами. Сильные ассоциативные связи между объектами приводят к их сближению на карте или включению в один кластер. Слабые связи, напротив, приводят к разделению объектов. За счёт этого геометрическая структура карты согласуется со структурой тематических кластеров.

Для сети из n узлов вводится мера ассоциации между узлами i и j

$$s_{ij} = \frac{2mc_{ij}}{c_i c_j} \tag{1}$$

где

c_{ij} обозначает количество связей между узлами i и j,

$c_i = \sum_j c_{ij}$ обозначает общее число связей узла i,

$m = \frac{1}{2}\sum_i c_i$ обозначает общее число связей в сети.

Данная величина отражает относительную силу связи между объектами с учётом их общей активности.

Картографирование и кластеризация формулируются как задача минимизации функции

$$V(x_1, \ldots, x_n) = \sum_{i<j} s_{ij} d_{ij}^2 - \sum_{i<j} d_{ij} \tag{2}$$

где d_{ij} обозначает расстояние между узлами i и j.

Первая часть функции описывает притяжение между сильно связанными узлами. Вторая часть задаёт отталкивание между объектами и предотвращает схлопывание карты.

При построении карты расстояние определяется как евклидово

$$d_{ij} = \|x_i - x_j\| \tag{3}$$

Для кластеризации расстояние задаётся в дискретном виде

$$d_{ij} = \begin{cases} 0, & \text{если } x_i = x_j, \\ \dfrac{1}{\gamma}, & \text{если } x_i \neq x_j. \end{cases} \tag{4}$$

В случае кластеризации минимизация функции V эквивалентна максимизации обобщённой модульности, связанной с классической модульностью Ньюмана Гирвана и дополненной параметром разрешения γ.

2.4.3 Тематическое моделирование научных текстов

Тематическое моделирование (topic modelling) представляет собой направление в области машинного обучения и статистического анализа текстов, ориентированное на выявление скрытых тематических закономерностей в крупных коллекциях документов. В отличие от методов, основанных на ручной разметке данных, тематические модели автоматически выделяют группы взаимосвязанных слов, совместное появление которых позволяет описывать содержательные темы корпуса. Подобные подходы широко используются при анализе научной литературы для определения ключевых исследовательских направлений, выявления тематических тенденций и изучения их изменения во времени [20, 21].

Одной из наиболее известных моделей тематического анализа является Latent Dirichlet

Allocation (LDA). В рамках данной модели каждый документ рассматривается как комбинация нескольких скрытых тем, а каждая тема описывается вероятностным распределением терминов [22]. Благодаря своей интерпретируемости и относительной простоте LDA остаётся одной из наиболее распространённых моделей тематического моделирования.

Несмотря на широкое применение, модель LDA обладает рядом ограничений. В основе модели лежит представление текста в виде «мешка слов», вследствие чего не учитывается порядок слов и сложные семантические зависимости между ними. Кроме того, число тем задаётся заранее, что требует дополнительной настройки модели. Отдельной проблемой является снижение эффективности при работе с большими словарями и корпусами, содержащими быстро изменяющуюся лексику [23].

Для преодоления указанных ограничений были разработаны более сложные методы тематического анализа.

- **Иерархические тематические модели**. Данный класс моделей используется для описания многоуровневой структуры тематик, в которой крупные исследовательские направления располагаются на верхних уровнях иерархии, а более специализированные темы и подтемы формируются на нижних уровнях. Такой подход позволяет учитывать взаимосвязи между темами в рамках единой структуры и анализировать научную область на разных уровнях детализации. В современных исследованиях для построения подобных моделей применяются как вероятностные методы, так и нейросетевые архитектуры [24, 25].

- **Динамические тематические модели (DTM)**. Модели данного типа предназначены для анализа изменения тематических структур во времени. Они позволяют исследовать процессы появления новых тем, развития исследовательских направлений и изменения структуры научных областей в различных временных интервалах. Подобные методы активно используются при анализе научных публикаций и исследовании эволюции научного знания [26, 27].

- **Нейросетевые динамические модели**. Данные методы представляют собой развитие классических динамических тематических моделей. В таких подходах используются эмбеддинги слов и документов, позволяющие учитывать контекстные и семантические связи между терминами. Это повышает согласованность тематических представлений и позволяет лучше учитывать смысловую структуру текста. Одним из примеров подобных методов является модель ANTM [28], основанная на использовании контекстных признаков и методов сопоставления тем между временными интервалами.

Помимо LDA, в задачах тематического моделирования применяются и другие подходы. Одним из наиболее известных методов является неотрицательная матричная факторизация (NMF) [29]. Подобные модели часто позволяют получать более интерпретируемые тематические структуры за счёт факторизации матрицы терминов и документов. Также широкое распространение получили методы, основанные на нейросетевых эмбеддингах, учитывающих контекстные и семантические связи между словами [30, 31].

Современные исследования всё чаще ориентированы на использование методов глубокого обучения. В подобных моделях применяются вариационные автоэнкодеры, графовые нейронные сети и трансформерные архитектуры. Такие подходы позволяют учитывать сложные зависимости между словами и документами, а также моделировать семантическую близость терминов. Модели семейства ProdLDA и Neural Topic Model демонстрируют устойчивость к шуму и подходят для анализа крупных текстовых коллекций [32].

Отдельное направление составляют гибридные методы, объединяющие вероятностные тематические модели и нейросетевые эмбеддинги. Подобные подходы сочетают интерпретируемость классических тематических моделей с возможностями глубокого семантического анализа. К данной группе относятся BERTopic, Top2Vec и Contextualized Topic Models. Использование подобных методов позволяет формировать более точные тематические представления и исследовать взаимосвязи между темами [32].

Отдельную сложность представляет задача оценки качества тематических моделей. Для этого применяются метрики когерентности тем, меры семантической связности терминов и методы визуального анализа результатов. Современные исследования также рассматривают автоматическое формирование названий тем и использование контекстуальных эмбеддингов для оценки согласованности тематических структур [30].

Таким образом, современные методы тематического моделирования включают широкий спектр подходов: от классических вероятностных моделей LDA и NMF до гибридных и нейросетевых архитектур. Выбор конкретного метода определяется характеристиками корпуса, задачами исследования, требованиями к интерпретируемости результатов и необходимостью анализа временной динамики тематических структур.

2.5 Автоматизация построения научных карт

Автоматизированное построение научных карт представляет собой направление исследований, ориентированное на применение вычислительных и аналитических методов для сокращения объёма ручной работы при сборе, обработке и визуализации научных данных. Актуальность подобных подходов обусловлена быстрым увеличением количества научных публикаций и сопутствующих метаданных, вследствие чего традиционные методы анализа оказываются недостаточно эффективными и плохо масштабируются при работе с крупными информационными массивами.

2.5.1 Методы автоматического картографирования науки

Подходы к автоматическому картографированию науки включают как полностью автоматизированные методы, так и гибридные решения, сочетающие вычислительный анализ с экспертной оценкой результатов. В большинстве случаев автоматизация охватывает этапы сбора научных публикаций, обработки метаданных, построения сетей цитирования и соавторства, тематической кластеризации и последующей визуализации полученных структур в виде научных карт.

Для решения подобных задач применяются специализированные программные средства и библиотеки, среди которых широко используются VOSviewer, CiteSpace, Bibliometrix и Sci² Tool [33]. Эти инструменты ориентированы на работу с крупными массивами научных публикаций и позволяют автоматически выявлять основные структурные компоненты исследуемой области. Например, система VOSviewer поддерживает построение карт на основе анализа цитирования, соавторства и совместной встречаемости терминов. Кроме того, она предоставляет средства визуализации тематических кластеров и взаимосвязей между ними [34].

Автоматическое построение научных карт обычно реализуется в виде последовательных конвейеров обработки данных. Такие конвейеры включают этапы загрузки публикаций из внешних источников, нормализации метаданных, извлечения текстовых и сетевых признаков, кластеризации объектов и визуализации результатов. В качестве источников данных часто используются базы Web of Science, Scopus и Dimensions. Например, пакет Bibliometrix для языка R позволяет автоматизировать обработку крупных массивов библиографических данных, рассчитывать сетевые характеристики и строить временные тематические зависимости [35, 36].

В последние годы методы искусственного интеллекта и машинного обучения всё активнее применяются в задачах автоматического картографирования науки. Современные исследования показывают, что алгоритмы ИИ позволяют автоматизировать поиск релевантных публикаций, выполнять дисамбигуацию авторов, анализировать развитие тематических направлений и оценивать влияние научных работ [37]. Использование подобных подходов существенно расширяет возможности классического библиометрического анализа и позволяет эффективно работать с большими массивами научных данных, включая междисциплинарные коллекции публикаций.

Помимо традиционных программных систем, в последние годы начали активно развиваться многоагентные подходы к автоматизации анализа научной информации. Подобные системы способны автоматически выполнять поиск научных публикаций, строить сетевые структуры взаимосвязей, генерировать программный код и формировать аналитические отчёты. Одним из примеров подобных решений является AI-augmented bibliometric framework, преобразующий запросы на естественном языке в автоматизированные сценарии анализа научных данных [38]. Использование таких систем позволяет значительно сократить объём ручной работы при анализе научных корпусов.

Автоматизация анализа научных публикаций также способствует более широкому использованию методов обработки естественного языка (NLP) для исследования содержания научных текстов. К подобным методам относятся извлечение ключевых терминов, тематическое моделирование, анализ контекстов цитирования и семантическое сопоставление тематик. Такие подходы позволяют учитывать не только формальные связи между публикациями, но и смысловые взаимосвязи между текстами. Методы, основанные на текстовых эмбеддингах и глубоких нейронных сетях, во многих случаях обеспечивают более точную тематическую кластеризацию и позволяют выявлять скрытые структуры в научных корпусах.

Несмотря на преимущества автоматизации, подобные системы имеют ряд ограничений [37]. Одной из основных проблем остаётся качество исходных данных. На практике часто встречаются неполные метаданные, ошибки в именах авторов, неоднородные описания публикаций и дублирующиеся записи. Дополнительные сложности возникают на этапе интерпретации результатов, поскольку автоматические алгоритмы могут формировать шумовые связи и требовать дополнительной настройки параметров. Кроме того, современные методы искусственного интеллекта и машинного обучения требуют значительных вычислительных ресурсов и специализированных знаний в области анализа данных. Это создаёт дополнительные требования к исследовательской инфраструктуре и подготовке специалистов.

3 Формирование и анализ корпуса научных публикаций

3.1 Анализ структуры карты Кастеллани

Карта наук о сложности Брайана Кастеллани [39] представляет собой не только визуализацию научной области, но и сложную структурированную систему организации знаний, объединяющую тематические, временные и интеллектуальные связи внутри исследовательского направления. В контексте настоящей работы карта рассматривается прежде всего как объект последующей формализации и автоматической обработки, что требует анализа её внутренней структуры с точки зрения представления данных, иерархии тематик и взаимосвязей между элементами. В отличие от большинства современных методов science mapping, основанных на автоматическом анализе библиометрических данных или тематическом моделировании текстов, карта Кастеллани была сформирована экспертным способом и отражает авторскую интерпретацию структуры наук о сложности. При этом карта обладает рядом характеристик, которые делают её особенно значимой для исследования автоматических методов построения хронологических тематических карт. Она одновременно содержит тематическую иерархию, временную структуру, связи между исследовательскими направлениями и указание на вклад отдельных авторов в развитие научных тем. Таким образом, карта представляет собой многослойную структуру знаний, включающую несколько типов сущностей и отношений между ними.

С точки зрения организации знаний карта построена по принципу тематической декомпозиции научной области. В её основе лежит разделение науки о сложности на крупные интеллектуальные традиции, внутри которых выделяются более специализированные тематические направления. Каждая традиция формирует отдельный тематический кластер, объединяющий взаимосвязанные концепции, методы и исследовательские области. При этом карта не является строгой древовидной иерархией: многие темы обладают междисциплинарным характером и могут быть связаны сразу с несколькими интеллектуальными направлениями. Анализ структуры карты показывает наличие нескольких уровней тематической организации. Верхний уровень представлен крупными интеллектуальными традициями, отражающими фундаментальные исследовательские парадигмы в области complexity science. Эти традиции выполняют роль наиболее общих категорий, внутри которых группируются

тематические области более низкого уровня. Следующий уровень включает отдельные исследовательские направления, методы и концептуальные области, связанные с конкретными аспектами изучения сложных систем. На более детальном уровне располагаются специализированные темы, прикладные области и междисциплинарные направления, возникающие на пересечении нескольких интеллектуальных традиций.

Карта обладает выраженной иерархической структурой, в которой наблюдается переход от общих исследовательских парадигм к более узким и специализированным тематическим областям. При этом сама иерархия носит не строго формальный, а скорее семантический характер. Связи между уровнями определяются не фиксированными правилами классификации, а экспертной интерпретацией тематической близости и исторического развития исследовательских направлений. Это существенно усложняет задачу автоматической реконструкции подобной структуры на основе текстовых данных. Отдельный интерес представляет способ отображения междисциплинарных связей. В отличие от традиционных иерархических классификаторов, где каждая тема принадлежит только одной категории, карта Кастеллани допускает существование пересекающихся тематических областей. Некоторые направления располагаются на стыке нескольких интеллектуальных традиций, отражая перенос методов и концепций между различными исследовательскими школами. Наличие подобных связей является одной из ключевых особенностей карты, поскольку именно междисциплинарность рассматривается как фундаментальное свойство наук о сложности.

Для дальнейшей автоматизированной обработки данных важную роль играет временная организация карты. Временная шкала используется для отображения динамики развития научной области: тематические направления размещаются в соответствии с периодами их наибольшей активности и значимости для исследовательского сообщества. Благодаря этому карта отражает не только структуру предметной области, но и изменения, происходящие в ней с течением времени. При этом временная привязка относится не к отдельным публикациям, а к тематическим направлениям и исследовательским областям. Такой подход сближает структуру карты с задачами динамического тематического моделирования, ориентированного на анализ эволюции тем во времени.

Дополнительный уровень представления информации формируется за счёт связей между авторами и тематическими направлениями. Исследователи соотносятся с теми областями, развитие которых связано с их научной деятельностью и вкладом. В результате карта отражает не только тематическую организацию научной области, но и особенности её интеллектуального развития. В данном случае авторы рассматриваются не как элементы классической сети соавторства, а как участники формирования и развития отдельных направлений исследований. Подобный подход позволяет дополнительно связать тематическую структуру области с историей развития научного сообщества.

С формальной точки зрения карта Кастеллани может быть описана как многослойный граф, включающий несколько типов сущностей. К таким сущностям относятся интеллектуальные традиции, исследовательские направления, специализированные темы, временные интервалы и авторы. Между ними формируются иерархические, тематические, временные

и интеллектуальные связи. Подобная организация отличается от традиционных библиометрических карт, которые, как правило, моделируют только один тип отношений, например цитирование публикаций или тематическую близость документов.

Несмотря на высокую аналитическую ценность, использование карты в вычислительных методах затруднено. Карта существует только в виде визуального представления и не содержит формализованного машиночитаемого описания. Отсутствуют открытые данные, описывающие иерархию тем, временные связи, принадлежность тем к интеллектуальным традициям и связи между авторами и исследовательскими направлениями.

По этой причине потребовалась ручная формализация структуры карты. Основной задачей данного этапа стало преобразование визуального экспертного представления в структурированный формат, пригодный для автоматической обработки. В качестве промежуточного формата была выбрана Excel структура, позволяющая явно задавать сущности карты и связи между ними.

В процессе формализации были выделены основные элементы карты: интеллектуальные традиции верхнего уровня, тематические направления, специализированные темы, авторы и временные периоды. Для каждой темы фиксировались её положение в иерархии, принадлежность к определённой интеллектуальной традиции, наличие междисциплинарных связей и временные характеристики. Дополнительно в структуру были включены авторы, ассоциированные с соответствующими тематическими направлениями. Отдельное внимание уделялось описанию иерархических отношений между темами. Поскольку карта не содержит строго формализованной древовидной структуры, при ручной обработке потребовалась интерпретация визуального расположения элементов и их смысловых связей. В результате была построена многоуровневая структура, включающая темы верхнего уровня, подтемы и специализированные направления исследований. Формализованная Excel-структура фактически стала промежуточным представлением экспертной карты и использовалась далее как основа для построения корпуса научных публикаций. На её основе выполнялся поиск релевантных научных работ, связанных с конкретными авторами и тематическими направлениями, представленными на карте. Кроме того, формализованная структура использовалась как эталонная модель при сравнении автоматически построенной тематической карты с экспертной картой Кастеллани.

3.2 Сбор корпуса научных публикаций

После проведения структурного анализа карты Кастеллани и её формализации следующим этапом исследования стало формирование корпуса научных публикаций, предназначенного для последующего тематического моделирования и построения хронологической тематической карты. Основной задачей данного этапа являлось создание репрезентативной коллекции научных текстов, отражающей ключевые направления исследований в области наук о сложности и соответствующей тематической структуре, выделенной на основе экспертной карты.

В отличие от большинства существующих подходов, использующих автоматический

сбор публикаций по заранее заданным ключевым словам или библиометрическим запросам, в рамках настоящей работы корпус формировался вручную. Такой подход был обусловлен спецификой исследуемой предметной области и необходимостью сохранить соответствие между экспертной структурой карты Кастеллани и последующим корпусом публикаций. В качестве основы для формирования корпуса использовались авторы с карты Кастеллани. В процессе сбора выполнялся поиск научных работ авторов, представленных на карте Кастеллани.

Поиск научных материалов выполнялся вручную с использованием открытых научных источников и специализированных платформ распространения научных публикаций. В качестве основных источников использовались Google Scholar, ResearchGate, arXiv, Semantic Scholar, открытые архивы научных публикаций, сайты научных журналов и материалы конференций. Дополнительно применялись электронные библиотеки университетов и открытые издательские платформы, предоставляющие доступ к полным текстам научных работ.

Ручной сбор корпуса являлся важной частью исследования. Автоматическое формирование корпуса на основе ключевых слов или библиометрических метаданных могло привести к включению большого количества нерелевантных публикаций, содержащих термины, связанные с науками о сложности, но не относящихся к интеллектуальным традициям, представленными Кастеллани в его карте. Кроме того, область наук о сложности характеризуется высокой междисциплинарностью и вариативностью терминологии, что существенно усложняет автоматический подбор публикаций только на основе поисковых запросов.

В процессе формирования корпуса каждая публикация проверялась ручную с точки зрения тематической релевантности. Основным критерием для включения публикаци в итоговый корпус являлась принадлежность работы к исследовательским направлениям, связанным с науками о сложности и представленным в структуре карты Кастеллани. Предпочтение отдавалось публикациям с полным текстом, доступным в открытом доступе, что было необходимо для последующего извлечения текстового содержимого и построения тематической модели. В корпус включались научные статьи, материалы конференций, главы книг, монографии, внесшие значимый вклад в развитие рассматриваемых направлений.

В результате был сформирован корпус, включающий 419 научных публикаций, связанных с науками о сложности. Все публикации были сохранены в формате PDF и объединены в единую коллекцию документов для последующей автоматической обработки. Полученный корпус использовался на следующих этапах исследования, включая извлечение текстов, предобработку документов, построение тематической модели и анализ временной динамики тематических направлений.

3.3 Структура собранного датасета

Сформированный в рамках исследования датасет представляет собой структурированный корпус научных публикаций, предназначенный для последующего тематического моделирования и построения хронологических тематических карт. Корпус был подготовлен на основе формализованной структуры карты Кастеллани и включает публикации, относящиеся к различным направлениям исследований в области complexity science, systems science,

network science, computational social science и смежных дисциплин.

Структура датасета организована в виде набора взаимосвязанных метаданных и текстовых представлений документов. Для каждой публикации были сохранены уникальный идентификатор документа (doc_id), название публикации (title), авторы (authors), год публикации (year), тематические метки различных уровней и пути к PDF-файлам и извлечённым текстовым представлениям. Дополнительно для документов сохранялись иерархические тематические уровни, соответствующие структуре экспертной карты, включая принадлежность к одному из основных тематических направлений. Подобная организация корпуса позволила использовать единое структурированное представление как для этапов preprocessing и тематического моделирования, так и для последующего анализа временной динамики научных направлений.

Основные статистические характеристики корпуса приведены в таблице 2.

Таблица 2: Общая статистика корпуса научных публикаций

Параметр	**Значение**
Общее количество документов	419
Формат хранения	PDF
Наличие текстовых представлений	Да
Количество основных тематических направлений	5
Наличие иерархической разметки	Да
Временной диапазон публикаций	1920–2025

Временная структура корпуса

Временная структура корпуса характеризуется выраженной неравномерностью распределения публикаций по годам. Распределение документов по годам представлено на рисунке 2.

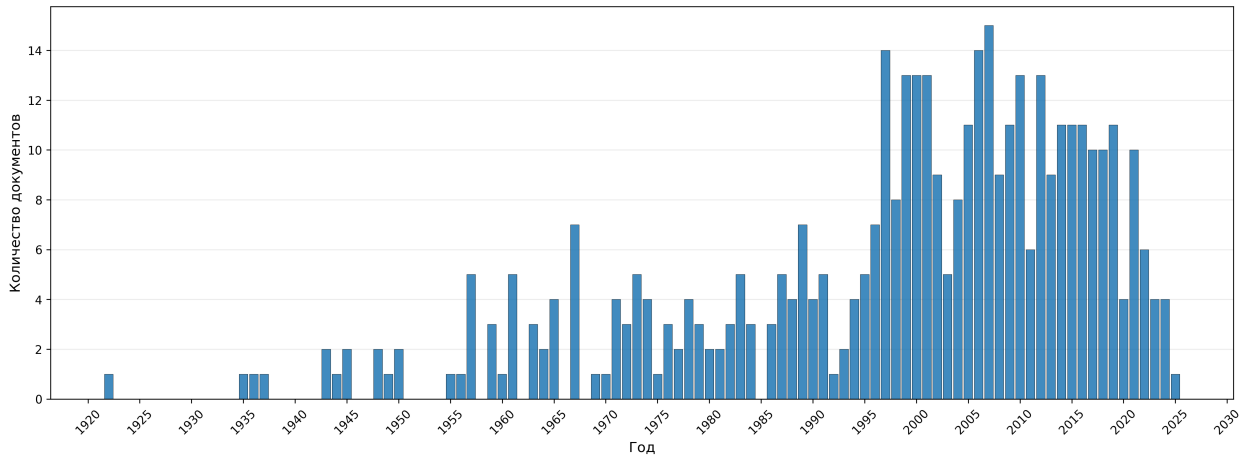

Рис. 2: Распределение документов по годам

Распределение публикаций по годам характеризуется постепенным увеличением числа доступных документов начиная со второй половины XX века. Наиболее высокая концентрация публикаций наблюдается в интервале 1995–2015 годов. Подобная структура корпуса

связана не только с развитием исследований в области complexity science, но и с особенностями доступности научных материалов в открытых цифровых источниках. Современные публикации значительно чаще представлены в полнотекстовом формате в электронных библиотеках, научных архивах и open-access репозиториях, тогда как более ранние работы нередко существуют только в виде печатных изданий, низкокачественных сканов или недоступны в открытом доступе. В результате корпус содержит существенно больше современных публикаций, пригодных для автоматического извлечения текста и последующего тематического моделирования.

Дополнительное представление temporal-структуры корпуса приведено на рисунке 2, где документы агрегированы по десятилетиям.

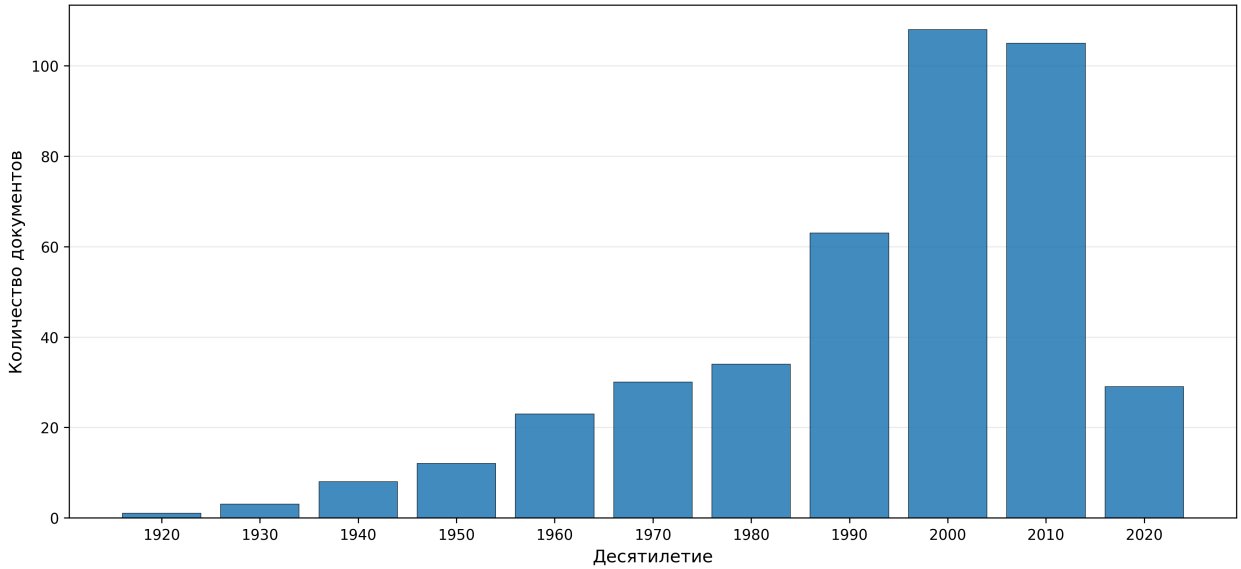

Рис. 3: Распределение документов по десятилетиям

Тематическая структура корпуса

Тематическая структура корпуса организована иерархически и соответствует экспертной структуре карты Кастеллани. Распределение документов по пяти основным тематическим направлениям приведено на рисунке 4.

Наиболее представленными направлениями являются *Artificial Intelligence* и *Complex Systems Theory*, включающие 137 и 115 публикаций соответственно. Существенная доля корпуса также относится к направлениям *Systems Science* и *Cybernetics*.

Более детализированное распределение публикаций по темам второго уровня представлено на рисунке 5.

Среди наиболее представленных тематик второго уровня выделяются *Complexity and Psychology*, *Network Science*, *Computational Biology/Chemistry*, *Interdisciplinary Methods* и *Social Systems Theory*. При этом распределение документов по темам второго уровня характеризуется заметной неравномерностью. Для части тематик было найдено лишь ограниченное количество публикаций, тогда как отдельные направления представлены существенно более полно. Подобная неоднородность обусловлена как различной степенью представленности исследовательских областей в открытых научных источниках, так и особенностями самой

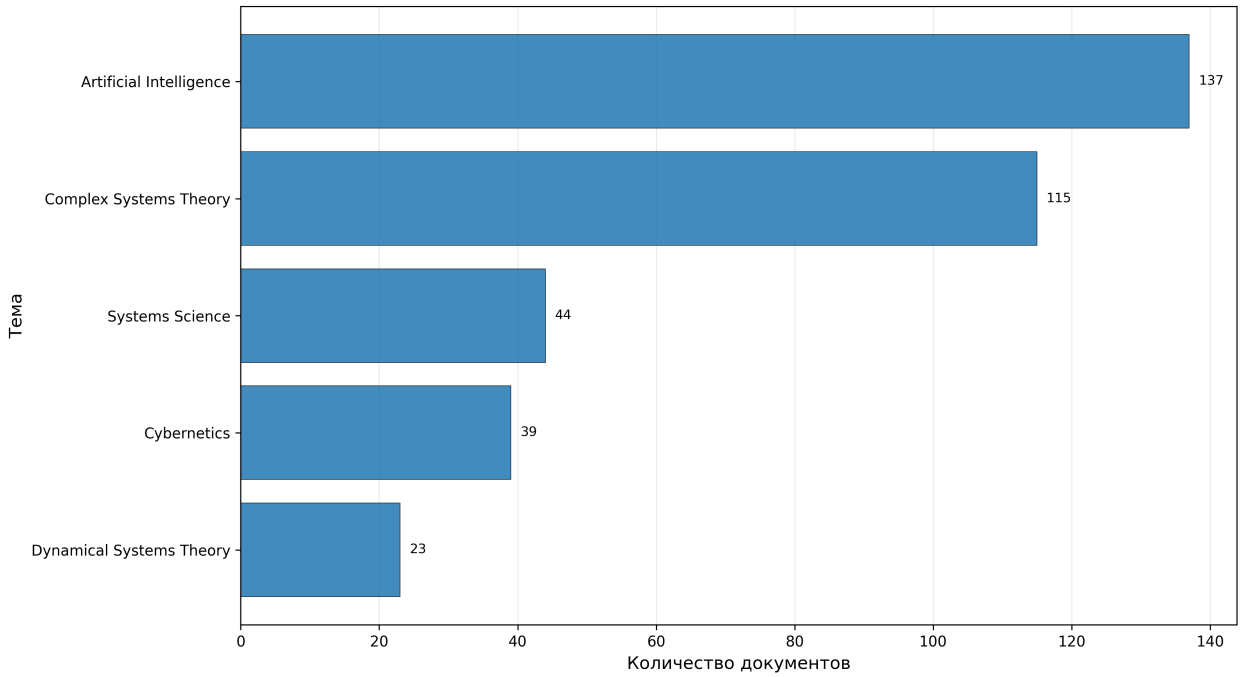

Рис. 4: Распределение документов по основным темам

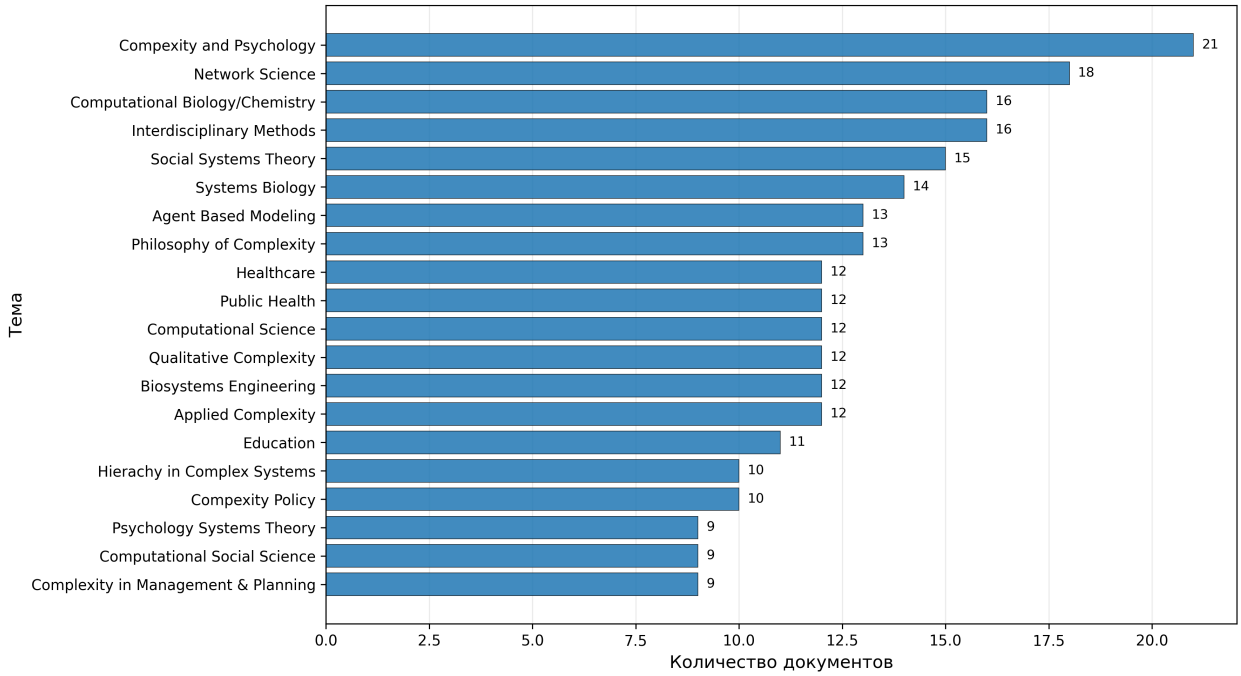

Рис. 5: Распределение документов по темам второго уровня

структуры карты Кастеллани, в которой отдельные направления обладают различной степенью детализации. Наличие тем с небольшим количеством документов является важной особенностью корпуса и создаёт дополнительные ограничения для последующего тематического моделирования, поскольку устойчивость автоматически выделяемых тематических структур напрямую зависит от объёма доступных текстовых данных внутри каждой тематической области.

На третьем уровне иерархии тематическая структура становится существенно более

специализированной и начинает отражать отдельных исследователей и исследовательские школы. Соответствующее распределение приведено на рисунке 6.

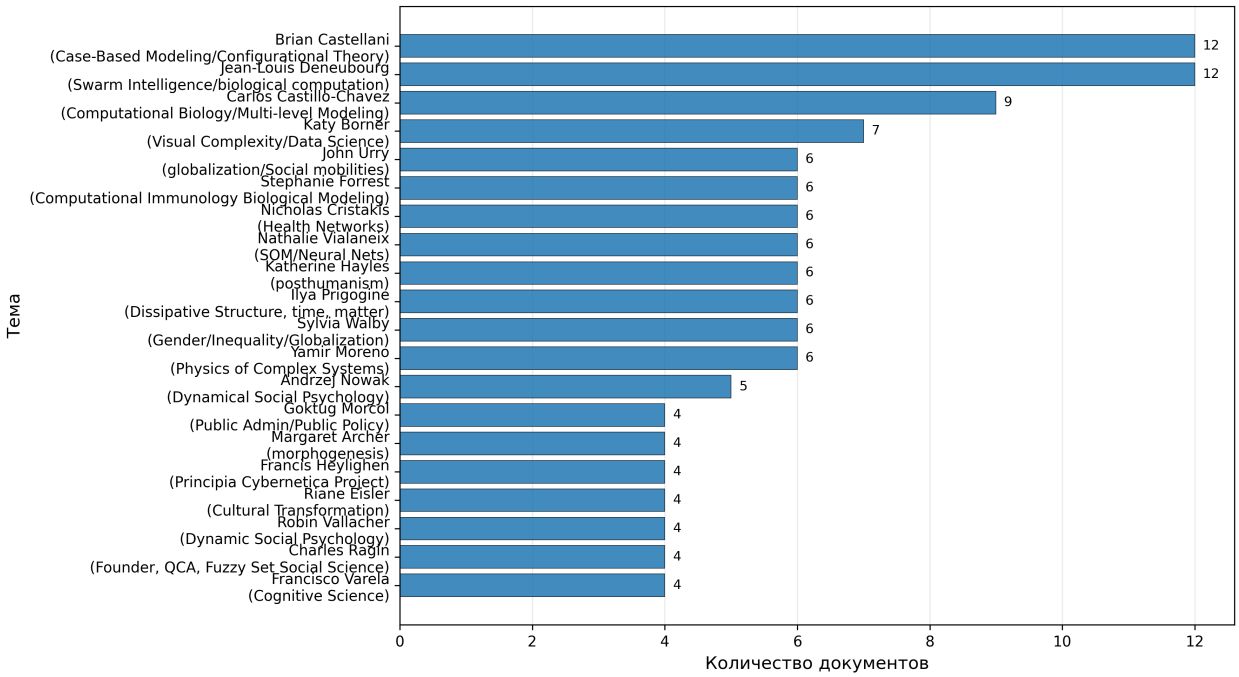

Рис. 6: Распределение документов по темам третьего уровня

Темы третьего уровня включают как специализированные исследовательские направления, так и отдельных исследователей, связанных с определёнными интеллектуальными традициями.

Структура PDF-корпуса

Дополнительной характеристикой корпуса является структура PDF-документов, представленная на рисунке 7.

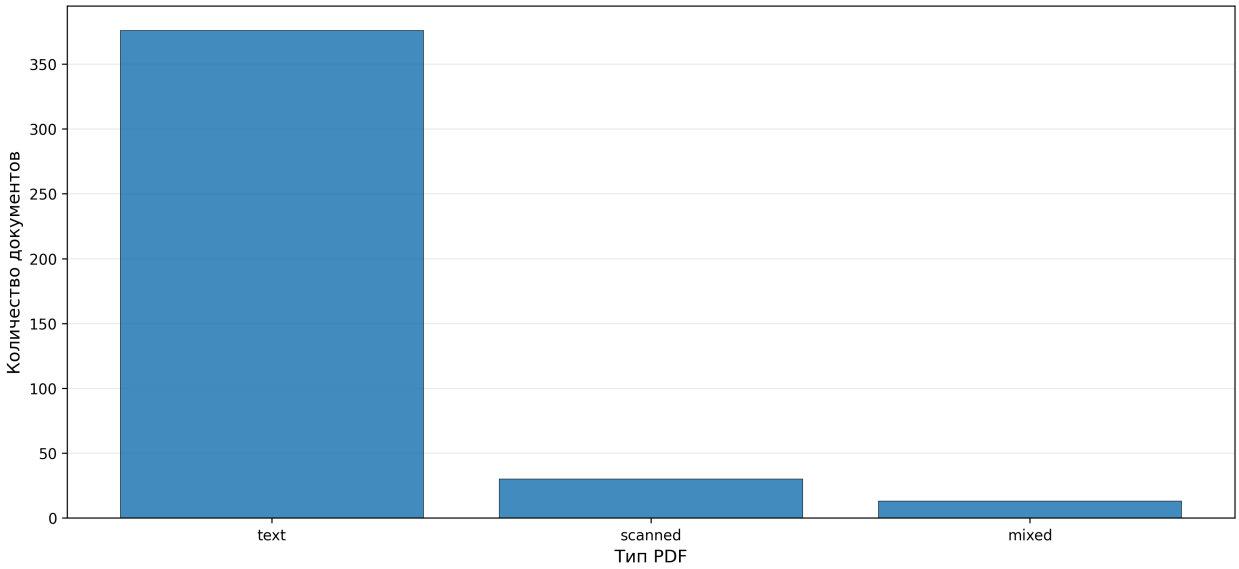

Рис. 7: Типы PDF-документов в корпусе

Основную часть корпуса составляют text-based PDF-документы, содержащие встро-

енный текстовый слой и пригодные для автоматического извлечения текста без применения OCR-подходов. Одновременно в корпусе присутствуют scanned и mixed PDF-документы, содержащие либо сканированные изображения страниц, либо смешанные текстово-графические представления. Наличие подобных документов обусловило необходимость разработки отдельного preprocessing pipeline для обработки различных типов PDF-файлов и нормализации текстовых представлений.

Размер документов и текстовая структура

Размеры документов в корпусе характеризуются выраженной неоднородностью. Распределение документов по группам длины представлено на рисунке 8.

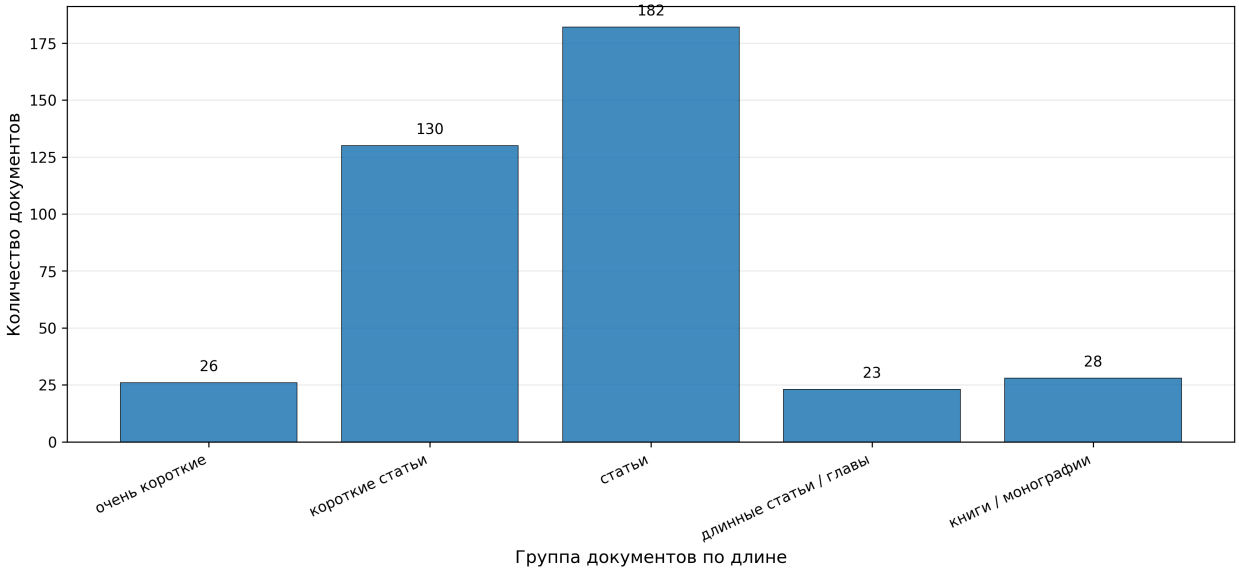

Рис. 8: Распределение документов по группам длины

Наибольшую часть корпуса составляют статьи средней длины и короткие научные статьи. Одновременно в датасете присутствуют как очень короткие публикации, так и крупные монографии и главы книг. Подобное распределение отражает разнообразие типов научных материалов, включённых в корпус.

Более детальное распределение размеров документов по количеству токенов приведено на рисунке 9.

Гистограмма распределения длины документов характеризуется наличием выраженного длинного хвоста. Большая часть публикаций содержит сравнительно ограниченное количество токенов, однако в корпусе присутствует небольшое число крупных документов, объём которых значительно превышает средние значения. Подобная неоднородность является важной особенностью корпуса с точки зрения тематического моделирования, поскольку крупные документы способны оказывать непропорционально высокий вклад в формирование итоговой тематической структуры. По этой причине на последующих этапах обработки применялись методы нормализации и балансировки текстовых представлений.

Статистика корпуса и междисциплинарность

Сводная статистика корпуса подтверждает его выраженный междисциплинарный характер. В датасете представлены публикации, относящиеся к *computational*, *social*, *biological*

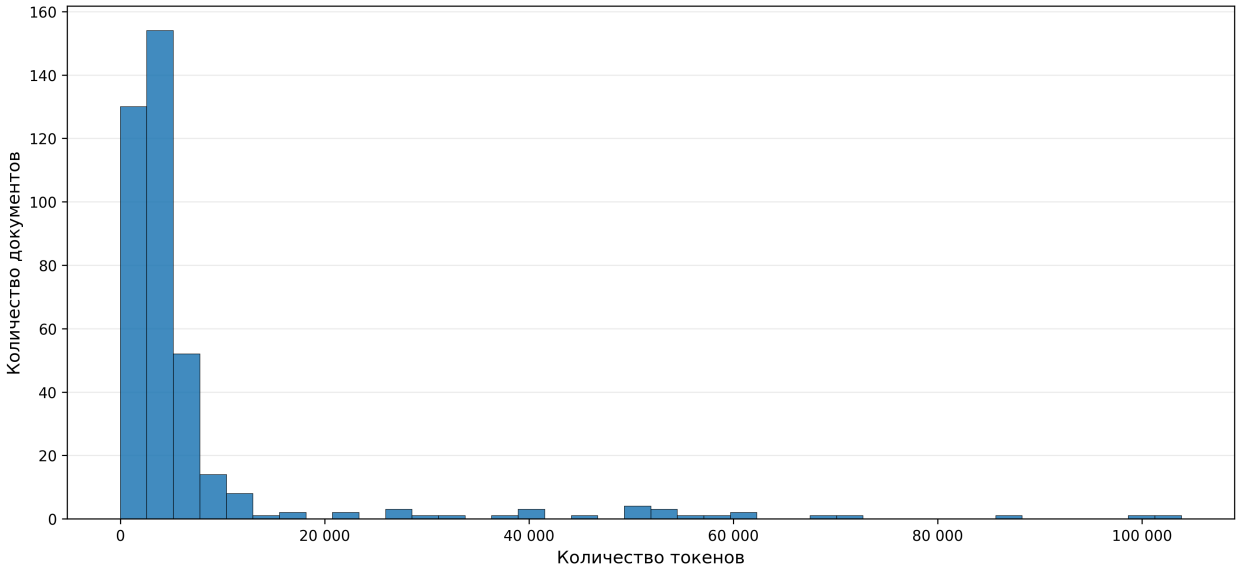

Рис. 9: Распределение документов по группам длины

и *systems-oriented* направлениям исследований. Значительная часть документов связана сразу с несколькими тематическими направлениями, что соответствует особенностям complexity science как междисциплинарной области.

Дополнительно в корпусе были выделены публикации, относящиеся сразу к нескольким тематическим направлениям. Подобные документы отражают междисциплинарный характер исследований в области complexity science и позволяют учитывать пересечение различных научных тематик при дальнейшем анализе корпуса. Наличие междисциплинарных публикаций является важным элементом структуры датасета, поскольку позволяет исследовать не только отдельные тематические области, но и связи между ними.

В результате был сформирован структурированный корпус научных публикаций, содержащий тематически разнообразные материалы по различным направлениям науки о сложности. Полученный корпус использовался на последующих этапах исследования для предобработки текстов, построения иерархической тематической модели, анализа временной структуры научных направлений и формирования автоматизированной хронологической тематической карты.

3.4 Проблемы и ограничения корпуса

Сформированный корпус научных публикаций имеет ряд ограничений, которые связаны со спецификой исходных собранных данных и процессом подготовки этих документов. Одним из основных ограничений стало неоднородное качество PDF файлов: часть публикаций содержала артефакты извлечения текста, включая некорректные символы, ошибки кодировки, разрывы слов, элементы служебной информации и другие шумовые фрагменты, возникающие при обработке PDF документов. Таким образом, качество итогового текста во многом зависело от типа исходного файла и способа его цифрового представления.

Дополнительным фактором являлось наличие scanned PDF-документов, не содержа-

щих встроенного текстового слоя. В отдельных случаях документы содержали смешанный текстовый слой, включающий как корректно извлечённый текст, так и OCR-фрагменты с пониженным качеством распознавания. Данный фактор мог влиять на качество итогового текстового корпуса и последующее тематическое моделирование.

Корпус также характеризуется выраженной неоднородностью структуры документов. В датасет включались научные статьи, публикации с конференций, главы книг и монографии, существенно различающиеся по объёму, стилю написания и уровню детализации материала. Наличие документов различной длины создаёт дополнительную вариативность текстовых представлений и может влиять на распределение тематических признаков внутри корпуса.

Одним из ограничений является неравномерное распределение публикаций по тематическим направлениям. Наиболее полно представлены направления *Artificial Intelligence* и *Complex Systems Theory*, тогда как отдельные специализированные области содержат ограниченное число документов. Подобный тематический дисбаланс является естественным следствием как структуры экспертной карты Кастеллани, так и различной доступности научных публикаций в открытых источниках.

Следует отметить, что сформированный корпус не является полным отражением всей области *complexity science*. Датасет формировался на основе структуры карты Кастеллани и ориентирован прежде всего на представленные в ней интеллектуальные традиции и исследовательские направления. Дополнительно на состав корпуса влияла доступность публикаций в открытом доступе и наличие машиночитаемых PDF-версий документов. Часть работ, представленных на экспертной карте, отсутствовала в пригодном для автоматической обработки формате либо была недоступна для извлечения текста.

Несмотря на перечисленные ограничения, сформированный корпус обладает достаточным объёмом, временным охватом и тематическим разнообразием для проведения тематического моделирования и последующего построения хронологической тематической карты исследуемой научной области.

4 Предобработка текстов и построение корпуса

Архитектура конвейера обработки

После формирования корпуса научных публикаций следующим этапом исследования являлась подготовка текстовых данных для тематического моделирования. На данном этапе был реализован пайплайн предобработки, ориентированный на обработку научных PDF-документов и формирование стандартизированного корпуса в формате Vowpal Wabbit для последующего использования в BigARTM.

Предобработка была реализована в виде последовательности отдельных Python-скриптов, выполняющих независимые этапы обработки данных. Подобная архитектура позволила разделить этапы извлечения текста, очистки документов, токенизации и построения корпуса, а также упростила воспроизводимость pipeline и контроль промежуточных результатов. Общая структура preprocessing pipeline представлена на рисунке 10.

Рис. 10: Процесс предобработки данных

На первом этапе выполнялась классификация PDF-документов по типу представления текста. В рамках реализованного pipeline выделялись три категории документов: text PDF, scanned PDF и mixed PDF. Классификация выполнялась на основе анализа количества извлекаемого текста и наличия изображений внутри страниц документа.

В ходе предварительного анализа было установлено, что сканированные PDF существенно ухудшают качество итогового корпуса вследствие большого количества OCR-артефактов. Ошибки распознавания текста приводили к появлению шумовых токенов, искажённых терминов и фрагментов слов, негативно влияющих на устойчивость тематического моделирования. По этой причине в финальной версии пайплайна предобработки было принято решение сосредоточиться преимущественно на text PDF-документах, содержащих встроенный текстовый слой. Документы категории смешанных PDF использовались ограниченно и проходили дополнительную очистку.

Извлечение текста осуществлялось с использованием библиотеки PyMuPDF (fitz). Для каждого PDF-документа выполнялось последовательное извлечение текстового содержимого страниц. Полученные тексты сохранялись в виде отдельных .txt файлов, что позволяло разделить этап извлечения текста и последующую NLP-обработку. Одновременно с извлечением текста формировалась metadata-структура корпуса. Для каждого документа сохранялись идентификатор документа (doc_id), название, авторы, год публикации, путь к PDF-файлу, путь к извлечённому тексту, тип PDF и принадлежность к тематическим уровням карты Кастеллани.

После извлечения текстов выполнялась очистка документов от PDF-шумов, служебных элементов и артефактов научных публикаций. Далее производилась токенизация и лемматизация текста с использованием библиотеки spaCy. На основе обработанных текстов дополнительно извлекались фразовые токены, представляющие собой устойчивые многословные научные выражения. Финальным этапом процесса предобработки являлось построение корпуса в формате Vowpal Wabbit. Для каждого документа формировалась строка, содержащая несколько модальностей: заголовок статьи, основной текст и специальные anchor-токены. Итоговый корпус использовался для обучения иерархической тематической модели BigARTM.

Таким образом, preprocessing pipeline представлял собой последовательный workflow подготовки научного корпуса, ориентированный на снижение уровня шума, нормализацию терминологии и повышение качества последующего тематического моделирования.

4.1 Очистка научных текстов

Извлечённые из PDF документов тексты содержали большое количество шумовых элементов, связанных со структурой научных публикаций и особенностями PDF формата. По этой причине одним из ключевых этапов предобработки являлась очистка текста и нормализация содержимого документов. Очистка была ориентирована на научные PDF документы и учитывала типичные артефакты публикаций, включая DOI, URL, email адреса, колонтитулы, номера страниц, переносы слов и PDF ligatures.

Одним из этапов предварительной обработки корпуса являлась коррекция *ligatures*, возникающих при извлечении текста из PDF файлов научных публикаций. Подобные типографские символы формировались в процессе PDF рендеринга и в дальнейшем могли приводить к ошибкам при токенизации и лемматизации текста. В ходе обработки наиболее распространённые ligature конструкции заменялись стандартными символьными последовательностями, что позволило сократить количество некорректно сформированных токенов и повысить качество текстовой обработки.

Дополнительно выполнялась нормализация структуры текста, направленная на устранение артефактов, возникающих при извлечении содержимого из PDF-документов. В процессе обработки восстанавливались слова, искусственно разделённые переносами строк, а также удалялись лишние символы перевода строки. Подобная обработка позволяла сохранять целостность научных терминов и уменьшать количество ошибочных токенов, попадающих в итоговый корпус.

Важным этапом очистки являлось удаление URL-адресов, DOI-идентификаторов и адресов электронной почты. Подобные элементы, как правило, не содержат тематически значимой информации, однако формируют большое количество уникальных шумовых токенов, негативно влияющих на качество словаря и результаты тематического моделирования. Кроме того, из текстов удалялись различные служебные элементы, возникающие при автоматическом извлечении текста из PDF-файлов, включая номера страниц, колонтитулы и другие технические артефакты. Исключение подобных фрагментов позволяло сократить объём шу-

мовых данных и повысить качество последующего тематического анализа корпуса.

Также на этапе предобработки выполнялось удаление разделов библиографии и списков литературы. Научные публикации содержат значительные объёмы цитируемых источников, которые могут искажать тематическое распределение документов. При обнаружении библиографического раздела текст документа обрезался начиная с соответствующего фрагмента, если данный раздел располагался во второй половине публикации.

В результате выполненной предобработки удалось существенно снизить уровень шума в корпусе и устранить значительную часть технических артефактов, связанных с обработкой PDF документов. Пример обработки текста приведён ниже.

Исходный фрагмент текста представлен на рисунке 11.

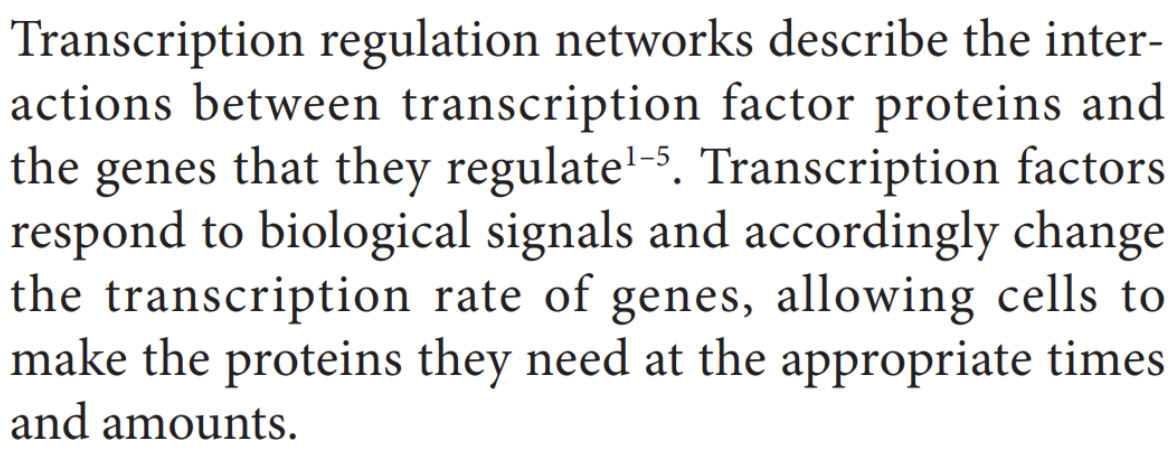
Transcription regulation networks describe the interactions between transcription factor proteins and the genes that they regulate[1–5]. Transcription factors respond to biological signals and accordingly change the transcription rate of genes, allowing cells to make the proteins they need at the appropriate times and amounts.

Рис. 11: Пример исходного текста

После предобработки:

transcription regulation network interaction transcription factor protein gene transcription factor respond biological signal change transcription rate gene allow cell protein need appropriate time

Токенизация и лемматизация

После очистки текстов выполнялась токенизация и лемматизация документов. Для реализации данного этапа использовалась библиотека spaCy и модель en_core_web_sm — компактная предобученная англоязычная NLP-модель, включающая средства токенизации, лемматизации, POS-tagging и синтаксического анализа. Модель обеспечивает достаточное качество обработки научных текстов при относительно низких вычислительных затратах, что является важным при обработке крупного корпуса PDF-документов. Токенизация выполнялась с использованием встроенного механизма spaCy. Для каждого токена вычислялась лемма, после чего выполнялась серия фильтраций. В финальную версию корпуса включались только токены частей речи: существительное, прилагательное, глагол. Использование POS-фильтрации позволяло исключить значительную часть функциональных слов, не несущих тематической нагрузки. Существительные формировали основную предметную терминологию корпуса, прилагательные позволяли сохранять специфику научных описаний, а глаголы отражали процессы и динамические аспекты исследуемых систем. Дополнительно удалялись

слишком короткие токены, числовые значения, punctuation tokens и шумовые фрагменты.

Одним из важных элементов предобработки являлось использование научных стоп-слов. Помимо стандартного набора стоп слов spaCy, был сформирован специализированный список научных служебных слов, в том числе:

```
[
"paper", "article", "study", "research",
"figure", "table", "section", "journal",
"conference", "introduction", "conclusion"
]
```

Научные публикации содержат большое количество общенаучной лексики, практически не связанной с конкретными тематическими направлениями. Подобные слова встречаются в большинстве документов корпуса и создают высокочастотный шум. Их удаление позволило повысить тематическую концентрацию корпуса.

После лемматизации происходила нормализация терминологии. Различные грамматические формы слов приводились к единому представлению, что снижало размер словаря и улучшало устойчивость тематических распределений.

Извлечение ключевых фраз

Одной из особенностей научных текстов является широкое использование устойчивых многословных терминов. Многие ключевые концепции complexity science представлены не отдельными словами, а словосочетаниями, например, *complex adaptive systems*. Использование только unigram-токенов приводило к потере значительной части смысловой структуры научных терминов. По этой причине в процесс предобработки был включен этап извлечения фразовых токенов. Максимальная длина фразовых токенов ограничивалась четырьмя словами, подобное ограничение позволяло избежать появления слишком длинных и нестабильных выражений.

Формирование VW-корпуса

После завершения предобработки выполнялось построение итогового корпуса в формате Vowpal Wabbit. Данный формат использовался в качестве входного представления для библиотеки BigARTM. Каждый документ представлялся в виде строки следующего вида:

```
doc_id |title ... |text ... |anchor ...
```

В рамках preprocessing pipeline использовались три модальности: title, text, anchor. Модальность title содержала токены заголовка, основной объём информации содержался в модальности text, включавшей токены основного текста документа и дополнительно использовались специальные anchor tokens. Anchor tokens использовались для мягкого тематического якорения части документов относительно основных направлений карты Кастеллани. Для каждого крупного тематического направления выбиралось несколько документов (3-4), после чего соответствующий anchor token добавлялся в VW-представление документа.

Итоговый конвейер обработки данных позволил сформировать стандартизированный корпус научных публикаций, пригодный для построения иерархической тематической модели BigARTM. Снижение уровня шума, нормализация терминологии и использование phrase

tokens существенно улучшили качество тематической структуры модели и интерпретируемость итоговых тематических кластеров.

5 Построение иерархической тематической модели

5.1 Теоретические основы тематического моделирования

Тематическое моделирование представляет собой совокупность вероятностных методов анализа текстовых данных, предназначенных для выявления скрытых тематических закономерностей в коллекциях документов. В основе подобных подходов лежит предположение о том, что каждый документ содержит несколько тем в различных пропорциях, а каждая тема может быть описана вероятностным распределением терминов.

Для анализа научных публикаций тематическое моделирование имеет особое значение, поскольку позволяет автоматически выделять исследовательские направления, выявлять скрытые связи между различными областями знаний и исследовать структуру научных корпусов без ручной классификации документов. В рамках данной работы методы тематического моделирования используются как основной инструмент построения хронологической тематической карты научной области.

Одним из ранних вероятностных подходов к тематическому моделированию является метод Probabilistic Latent Semantic Analysis (PLSA) [40]. В данной модели предполагается существование скрытых тем, определяющих вероятностное распределение слов в документах. Вероятность появления слова w в документе d описывается следующим выражением

$$p(w|d) = \sum_{t \in T} p(w|t)p(t|d) \tag{5}$$

где

- $p(w|t)$ обозначает распределение слов внутри темы;
- $p(t|d)$ обозначает распределение тем внутри документа.

Модель PLSA показала высокую эффективность при анализе текстовых коллекций различного типа. Однако данный подход обладает рядом ограничений. Одной из основных проблем является отсутствие полноценной генеративной вероятностной модели документов. Кроме того, модель демонстрирует недостаточную устойчивость при обработке новых документов, не использовавшихся в процессе обучения.

Развитием вероятностного тематического моделирования стала модель Latent Dirichlet Allocation (LDA) [22]. В данной модели используются априорные распределения Дирихле для тематических распределений документов и слов. Это позволяет повысить устойчивость модели и улучшить качество генерализации.

В модели LDA каждый документ описывается распределением тем

$$\theta_d \sim Dir(\alpha) \tag{6}$$

а каждая тема описывается распределением слов

$$\phi_t \sim Dir(\beta) \tag{7}$$

где

- θ_d обозначает тематическое распределение документа;
- ϕ_t обозначает распределение слов внутри темы;
- α и β являются параметрами распределений Дирихле, определяющими характер распределения тем и слов.

LDA стала одним из наиболее распространённых методов тематического моделирования научных корпусов. При этом модель имеет ряд ограничений, особенно при анализе сложных научных коллекций. В частности, LDA слабо контролирует интерпретируемость тем, не позволяет гибко управлять тематической структурой и затрудняет построение устойчивых иерархий.

Для решения этих ограничений была предложена концепция Additive Regularization of Topic Models (ARTM) [41]. В отличие от LDA, ARTM не использует фиксированную вероятностную генеративную схему. Тематическое моделирование формулируется как задача оптимизации с использованием набора регуляризаторов. Оптимизируемый функционал модели имеет следующий вид

$$L(\Phi, \Theta) = \sum_{d,w} n_{dw} \ln \sum_{t} \phi_{wt}\theta_{td} + R(\Phi, \Theta) \tag{8}$$

где

- (Φ)

  обозначает матрицу распределений слов по темам

- (Θ)

  обозначает матрицу распределений тем по документам

- $(R(\Phi, \Theta))$

  обозначает набор регуляризаторов

Основным преимуществом ARTM является возможность гибкого управления свойствами модели с помощью регуляризации. Это позволяет повышать разреженность тематических распределений, уменьшать корреляцию между темами и улучшать интерпретируемость

результатов. Для научных корпусов подобные свойства особенно важны, поскольку тексты содержат большое количество пересекающихся терминов и междисциплинарных направлений.

Регуляризованные тематические модели позволяют адаптировать модель под особенности конкретной задачи анализа текстов. В данной работе регуляризация использовалась для формирования интерпретируемой иерархической структуры тем, близкой по организации к экспертной карте Кастеллани.

Для построения тематической модели в настоящем исследовании использовалась библиотека BigARTM, реализующая подход ARTM и предоставляющая средства для построения иерархических, мультимодальных и semi-supervised тематических моделей.

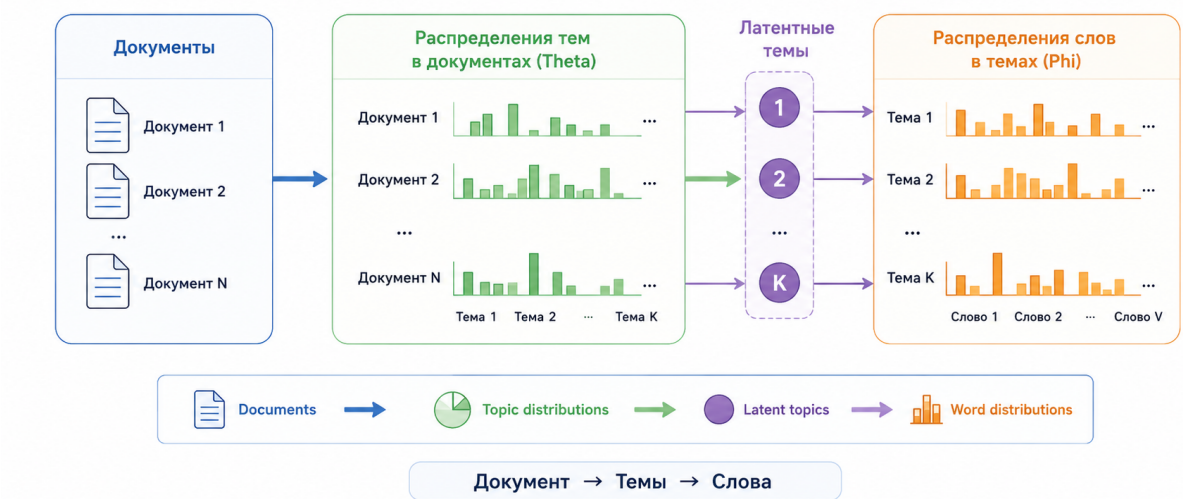

Рис. 12: Общая схема тематического моделирования научного корпуса

5.2 Модель ARTM и библиотека BigARTM

В рамках настоящей работы для построения тематической модели использовалась библиотека BigARTM[42], реализующая подход Additive Regularization of Topic Models. Выбор данной библиотеки обусловлен необходимостью построения интерпретируемой иерархической тематической структуры научного корпуса с возможностью использования регуляризации и мультимодальных представлений документов.

Основой модели ARTM являются две матрицы:

- матрица Φ;
- матрица Θ.

Матрица Φ описывает распределение слов по темам:

$$\Phi = \phi_{wt} \tag{9}$$

где:

- $\phi_{wt} = p(w|t)$ — вероятность слова (w) в теме (t).

Каждый столбец матрицы

$$(\Phi)$$

соответствует отдельной теме и представляет распределение терминов внутри неё. Анализ наиболее вероятных слов позволяет определить содержательное значение темы.

Матрица

$$(\Theta)$$

описывает распределение тем по документам

$$\Theta = \theta_{td} \tag{10}$$

где

- $\theta_{td} = p(t|d)$ обозначает вероятность темы (t) в документе (d)

Матрица

$$(\Theta)$$

использовалась для определения тематической принадлежности документов, построения временных распределений тем, вычисления тематических весов и формирования хронологической структуры карты.

В процессе обучения модели применялись регуляризаторы разреженности и декорреляции тем.

Регуляризатор разреженности использовался для формирования более компактных тематических распределений

$$R_{sparse}(\Phi, \Theta) \tag{11}$$

Использование разреженности позволяло уменьшить количество малозначимых слов внутри тем и повысить интерпретируемость результатов.

Дополнительно использовался регуляризатор декорреляции

$$R_{decor}(\Phi) \tag{12}$$

Данный регуляризатор уменьшал пересечение тем по словарю и способствовал формированию более независимых тематических направлений. Для научных корпусов это особенно важно, поскольку многие термины одновременно используются в нескольких исследовательских областях.

Одним из ключевых преимуществ BigARTM является поддержка мультимодальных представлений документов. В рамках настоящей работы использовались следующие модальности:

- `title`;

- `text`;
- `anchor`.

Использование отдельных модальностей позволило по-разному учитывать вклад заголовков, основного текста и anchor-токенов при обучении модели.

Дополнительным преимуществом BigARTM является возможность построения иерархических тематических моделей. Данная функциональность использовалась для формирования структуры тем, близкой к организации экспертной карты Кастеллани.

Важной особенностью библиотеки также является поддержка semi-supervised тематического моделирования. Использование weak supervision и anchor tokens позволило частично управлять разделением тем и повысить устойчивость модели на научном корпусе.

BigARTM хорошо подходит для анализа научных публикаций благодаря:

- поддержке крупных словарей;
- возможности использования регуляризации;
- поддержке иерархических структур;
- гибкой настройке модальностей;
- удобной работе с корпусами в формате Vowpal Wabbit.

В рамках настоящего исследования BigARTM выступала как основной инструмент формирования тематической структуры научного корпуса, используемой далее для построения хронологической тематической карты.

5.3 Реализация модели

Для построения тематической структуры научного корпуса в рамках настоящей работы использовалась иерархическая модель hARTM, реализованная средствами библиотеки BigARTM Весь код предоставлен в открытом доступе на GitHub[43]. Итоговая конфигурация модели была получена после серии экспериментальных запусков, направленных на подбор числа тем, параметров регуляризации и структуры иерархии.

Обучение модели выполнялось последовательно для трёх уровней тематической иерархии:

- уровень крупных тематических направлений (`level0`);
- уровень специализированных научных областей (`level1`);
- уровень детализированных исследовательских тематик (`level2`).

Конфигурация уровней модели представлена в таблице 3.

Level	Num Topics	Passes
level0	8	25
level1	62	35
level2	160	45

Таблица 3: Конфигурация уровней иерархической модели

Увеличение количества тем по уровням иерархии связано с переходом от крупных исследовательских направлений к более специализированным научным тематикам. Верхний уровень модели описывает обобщённую структуру корпуса и соответствует основным направлениям complexity science. Нижние уровни содержат существенно более детализированные тематические кластеры, отражающие отдельные исследовательские области, методы и интеллектуальные традиции.

Количество проходов обучения `passes` увеличивалось при переходе к более глубоким уровням иерархии. Для нижних уровней требовалось большее число итераций, поскольку специализированные темы обладают более высокой семантической плотностью и требуют более точной настройки распределений токенов.

В процессе обучения использовались регуляризаторы разреженности и декорреляции. В реализации применялись `SmoothSparsePhiRegularizer`, `SmoothSparseThetaRegularizer` и `DecorrelatorPhiRegularizer`.

Регуляризация разреженности для матрицы

$$(\Phi)$$

задавалась следующим выражением

$$R_\Phi = \tau_\Phi \sum_{w,t} \phi_{wt} \ln \phi_{wt} \tag{13}$$

Для матрицы

$$(\Theta)$$

использовалась аналогичная регуляризация

$$R_\Theta = \tau_\Theta \sum_{d,t} \theta_{td} \ln \theta_{td} \tag{14}$$

Применение регуляризации разреженности позволяло сократить количество малозначимых терминов внутри тем, повысить интерпретируемость тематических распределений и уменьшить влияние шумовых токенов. Кроме того, использование данного подхода способствовало формированию более специализированных тематических кластеров.

Для снижения пересечения тем по общему словарю использовался регуляризатор декорреляции:

$$R_{decor} = \tau_{decor} \sum_{t_i \neq t_j} \sum_{w} \phi_{wt_i} \phi_{wt_j} \tag{15}$$

Использование данного регуляризатора было направлено на уменьшение корреляции между темами и формирование более независимых тематических структур. Для корпуса научных публикаций это являлось особенно важным, поскольку значительная часть терминов используется одновременно в нескольких исследовательских областях.

Проведённые эксперименты показали, что отсутствие регуляризации декорреляции приводит к усилению смешивания тем и увеличению количества пересекающихся тематических кластеров.

Для различных модальностей корпуса задавались отдельные весовые коэффициенты. В рамках модели использовались модальности `text`, `title` и `anchor`.

Основной модальностью являлась `text`, содержащая полный текст научных публикаций. Для неё использовался наибольший вес при обучении модели, поскольку именно данная модальность определяла основную тематическую структуру корпуса.

Модальность `title` применялась с повышенным коэффициентом относительно стандартной текстовой модальности. Заголовки научных публикаций обладают высокой тематической информативностью и, как правило, содержат ключевые термины исследовательской области. Повышенный вес данной модальности позволял повысить устойчивость тематических распределений верхнего уровня.

Модальность `anchor` использовалась для реализации подхода weak supervision и содержала специальные anchor-токены. Вклад данной модальности был существенно ниже по сравнению с основной текстовой модальностью. Она применялась в качестве механизма мягкого тематического разделения. Основная функция anchor-токенов заключалась в стабилизации крупных тематических направлений и уменьшении смешивания близких тем.

Первоначально также использовалась author modality. Однако серия экспериментальных запусков показала, что включение авторов в словарь модели приводит к смещению тематической структуры. Вместо тематических направлений модель начинала кластеризовать документы по исследовательским школам и отдельным авторам. Это приводило к снижению тематической интерпретируемости и ухудшению качества.

По этой причине модальность авторов была исключена из финальной версии модели. Информация об авторах использовалась позднее отдельно при построении итоговой хронологической тематической карты.

Обучение модели выполнялось последовательно для каждого уровня иерархии. После завершения обучения сохранялись:

- матрицы Φ;
- матрицы Θ;
- top tokens;
- topic-year distributions;

- hierarchical topic relations.

Полученная модель (рисунок 13) использовалась как основа для построения хронологической тематической карты научного корпуса.

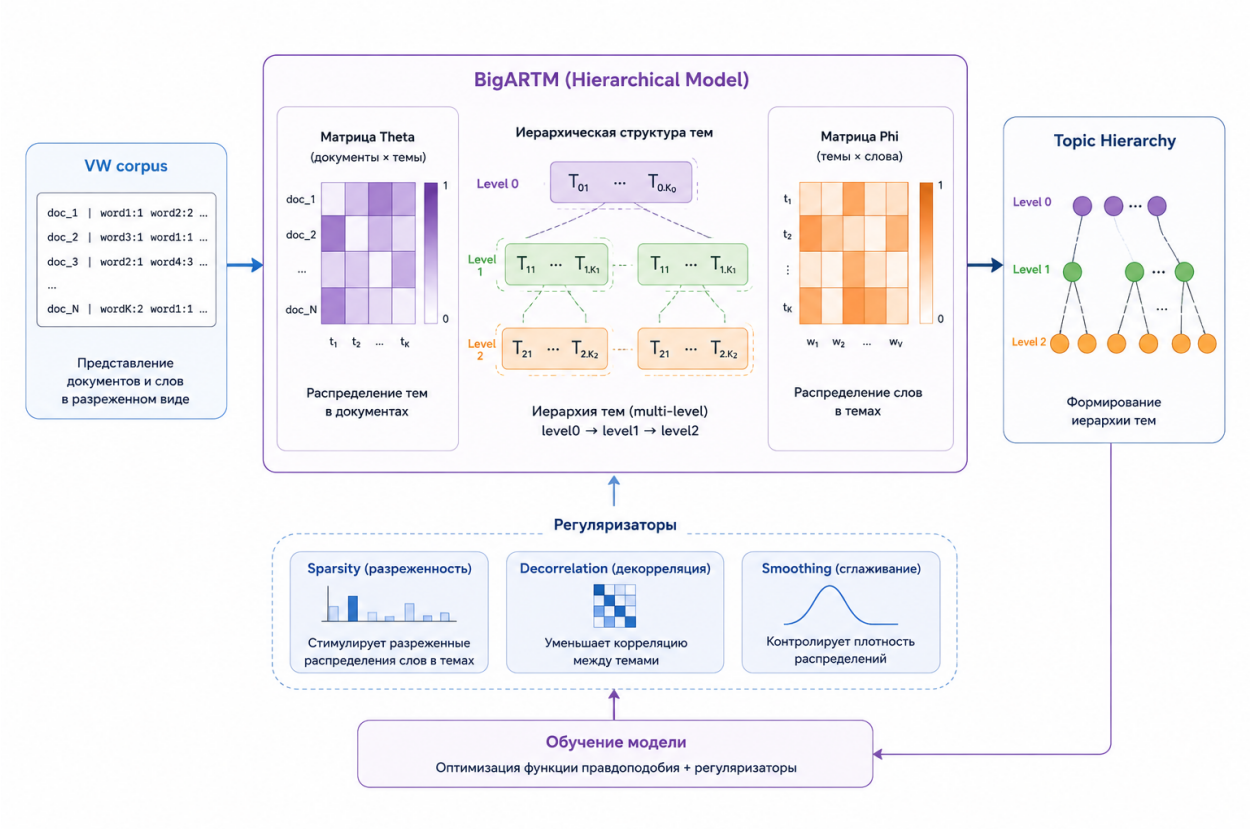

Рис. 13: Архитектура иерархической модели BigARTM

5.4 Формирование topic hierarchy

Одной из ключевых задач настоящего исследования являлось построение иерархической тематической структуры, отражающей организацию научной области complexity science. Для решения данной задачи использовался hierarchical ARTM pipeline, реализующий многоуровневую структуру тематических представлений.

Верхний уровень модели (`level0`) описывал наиболее крупные исследовательские направления корпуса. Темы уровня `level1` представляли более специализированные научные области, связанные с отдельными исследовательскими направлениями. На уровне `level2` формировались наиболее детализированные тематические кластеры, отражающие конкретные исследовательские школы, методы и интеллектуальные традиции.

Связи между уровнями иерархии формировались на основе распределений токенов внутри тем. Каждая тема описывалась topic vector, формируемым из распределения слов матрицы Φ.

Для оценки близости тем использовалась cosine similarity:

$$\text{cosine}(A, B) = \frac{A \cdot B}{|A||B|} \tag{16}$$

где:

- A и B — topic vectors;
- $A \cdot B$ — скалярное произведение векторов;
- $|A|$ и $|B|$ — нормы векторов.

Метрика косинусной близости применялась для решения нескольких задач:

- определения степени близости между темами;
- построения связей типа «родитель–дочерняя тема»;
- объединения тематически близких кластеров;
- формирования иерархической структуры графа.

Темы, обладающие высокой косинусной близостью, объединялись в единую иерархическую систему. Темы верхнего уровня интерпретировались как более общие исследовательские направления, тогда как темы нижних уровней использовались для детализации соответствующих областей.

Построение иерархии осуществлялось по принципу перехода от обобщённых тематических направлений к более специализированным темам. Верхние уровни иерархии содержали компактные и менее детализированные тематические структуры, в то время как нижние уровни включали специализированные тематические кластеры. С увеличением глубины иерархии возрастала степень детализации тематической модели.

Дополнительно выполнялась кластеризация тем на основе близости их тематических векторов. Темы со схожими распределениями терминов объединялись в локальные тематические группы.

Использование данного подхода позволяло выделять крупные исследовательские направления, анализировать внутреннюю структуру научной области, обнаруживать специализированные тематические сообщества и формировать hierarchical topic graph, структурно близкий к экспертной карте Кастеллани.

Для визуализации иерархии использовались граф тем *topic graph*, связи между родительскими и дочерними темами *parent child edges*, сеть тематической близости *topic similarity network* и хронологическое расположение тем *chronological topic layout*.

При построении итоговой иерархии использовались наиболее вероятные токены тем, распределения документов из матрицы

$$(\Theta)$$

, тематическая близость тем и временные распределения тематических направлений.

Дополнительно вычислялись временные распределения тем по годам. Для каждой темы определялись взвешенный год темы, период максимальной активности, временной интервал активного развития и динамика изменения тематического направления во времени.

Временная информация использовалась позднее при построении хронологической тематической карты и размещении тем на временной оси.

Итоговая иерархическая структура тем представляла собой многоуровневый тематический граф, отражающий:

- структуру научной области;
- взаимосвязи исследовательских направлений;
- тематическую специализацию;
- временную динамику научных тем.

6 Хронологическое моделирование и визуализация

6.1 Построение временной компоненты

После построения иерархической тематической модели следующим этапом являлось формирование временной компоненты тематической карты. Основной задачей данного этапа было определение временного распределения тематик и построение связи между тематической структурой корпуса и хронологией публикаций.

Временная компонента строилась на основе матрицы тематических распределений документов Θ и метаданных корпуса, содержащих годы публикации документов. Для каждого документа корпуса был известен год публикации y_d, а также тематическое распределение:

$$\theta_{td} = p(t|d) \tag{17}$$

где:

- θ_{td} — вес темы t в документе d;
- y_d — год публикации документа.

На основе полученных распределений выполнялось построение матрицы “тема год”, отражающей интенсивность присутствия каждой темы в различные временные периоды. Для каждой темы производилось агрегирование тематических весов документов по годам

$$T_{ty} = \sum_{d:y_d=y} \theta_{td} \tag{18}$$

где

- T_{ty} обозначает суммарный тематический вес темы t в году y;
- θ_{td} обозначает вес темы в документе;
- y_d обозначает год публикации документа.

Такое агрегирование позволило получить временное распределение каждой темы внутри корпуса научных публикаций. В отличие от использования количества документов, применение Θ весов позволяло учитывать степень тематической принадлежности документа. Один документ мог одновременно участвовать в формировании нескольких тем с различной интенсивностью.

Для построения временной структуры использовались только документы, обладающие ненулевыми весами, дополнительно выполнялась фильтрация слабых тематических связей. Документы с крайне малыми значениями θ_{td} не оказывали существенного влияния на временную структуру темы и исключались из временной агрегации.

В результате формировалась временная матрица распределения тематик, отражающая периоды появления тем, изменение интенсивности тем во времени, рост и снижение исследовательской активности, распределение тематических направлений по десятилетиям.

Дополнительно выполнялась нормализация временных распределений. Поскольку количество публикаций существенно увеличивалось после 1990 х годов, прямое использование абсолютных тематических весов приводило к доминированию современных публикаций. Для устранения данного эффекта использовалась относительная временная агрегация тематических весов.

На основе временной матрицы анализировалась динамика тематик. Для каждой темы определялись

- период появления темы;
- периоды максимальной активности;
- продолжительность существования темы;
- устойчивость тематического направления во времени.

Полученная временная структура использовалась далее при построении хронологической тематической карты и размещении тематических узлов относительно временной оси.

6.2 Определение временных характеристик тем

После построения временных распределений тем выполнялось вычисление временных характеристик тематических направлений. Основной задачей данного этапа являлось определение репрезентативного периода существования темы и формирование устойчивой временной координаты для последующей визуализации.

Для вычисления центрального периода активности темы использовалась взвешенная медиана по годам публикации документов. По сравнению со средним значением взвешенная

медиана обладает большей устойчивостью к выбросам и отдельным документам, удалённым от основного временного периода темы.

Для темы t использовался набор документов

$$D_t = d_1, d_2, ..., d_n \tag{19}$$

Каждому документу соответствовали год публикации y_d и тематический вес θ_{td}.

Взвешенная медиана определялась как минимальное значение года m, для которого выполнялись следующие условия

$$\sum_{y_d<m} \theta_{td} \leq \frac{1}{2} \sum_{d} \theta_{td} \tag{20}$$

и

$$\sum_{y_d>m} \theta_{td} \leq \frac{1}{2} \sum_{d} \theta_{td} \tag{21}$$

Полученное значение использовалось как репрезентативная временная характеристика темы.

Использование взвешенной медианы было связано с особенностями научного корпуса. Для части тематических направлений в данных присутствовали единичные ранние публикации, существенно удалённые от основного периода развития темы. Использование среднего значения в подобных случаях приводило к смещению временной координаты и ухудшению качества хронологической визуализации.

Взвешенная медиана позволяла учитывать тематические веса документов, снижать влияние временных выбросов и более корректно определять центральный период существования темы. Дополнительным преимуществом метода являлась его устойчивость при асимметричном распределении публикаций по временной шкале.

Для каждой темы дополнительно вычислялся год максимальной активности

$$y_t^{peak} = \arg\max_{y} T_{ty} \tag{22}$$

где T_{ty} обозначает суммарный тематический вес темы в году y.

Год максимальной активности характеризовал период наибольшего распространения соответствующего исследовательского направления внутри корпуса публикаций.

Для определения временного интервала активности темы использовался диапазон лет, внутри которого концентрировалась основная часть тематического веса. При вычислении учитывались преимущественно документы с высокими значениями θ, что позволяло исключить слабые тематические соответствия.

Дополнительно исследовалась временная динамика тематических направлений. Для части тем наблюдалось постепенное развитие на протяжении длительного периода времени. Для других направлений фиксировались резкие изменения публикационной активности, смещение исследовательского интереса между различными временными интервалами и по-

явление новых междисциплинарных областей.

Наиболее выраженный рост количества тематических направлений наблюдался после 1990 х годов, что связано с развитием вычислительных методов, сетевого анализа и междисциплинарных исследований сложных систем.

Полученные временные характеристики использовались при построении хронологической тематической карты и определяли расположение тематических узлов относительно временной оси.

6.3 Добавление авторов в тематическую карту

После формирования временной структуры тематической модели выполнялся этап добавления авторов в итоговую тематическую карту. Основной целью данного этапа являлось повышение интерпретируемости тематических направлений и связывание выявленных тем с конкретными исследовательскими школами и представителями научного сообщества.

Авторы не использовались в качестве самостоятельной модальности тематической модели и не участвовали в процессе обучения BigARTM. Формирование тематической структуры осуществлялось исключительно на основе текстового содержимого научных публикаций. Такой подход позволял избежать искусственного влияния авторских связей на распределение тем и сохранял независимость тематической организации модели.

Информация об авторах добавлялась только на этапе визуализации результатов. Для каждой темы определялись наиболее релевантные документы на основе theta-весов:

$$\theta_{td} = p(t|d) \tag{23}$$

Для дальнейшего анализа использовались документы с максимальными значениями θ_{td}, наиболее полно характеризующие содержание соответствующей темы. На основе выбранных публикаций из метаданных корпуса извлекалась информация об авторах.

Для каждой темы формировалось множество наиболее характерных авторов:

$$A_t = a_1, a_2, ..., a_k \tag{24}$$

где:

- A_t — множество авторов, ассоциированных с темой;
- a_i — автор публикации, тематически связанной с рассматриваемой темой.

При построении карты учитывались только наиболее часто встречающиеся авторы среди документов, обладающих высокой тематической релевантностью. Такая фильтрация позволяла уменьшить влияние случайных связей и повысить устойчивость интерпретации результатов.

Связи типа “автор–тема” формировались на основе тематической близости публикаций внутри корпуса. При этом данные о соавторстве, цитировании и другие внешние библиометрические зависимости в модели не использовались. Такой подход позволял избежать

дополнительного смещения тематической структуры и сохранять тематические кластеры, сформированные непосредственно на основе содержимого научных текстов.

Добавление авторов на этапе визуализации повышало интерпретируемость итоговой тематической карты. Тематические направления связывались с конкретными исследователями и исследовательскими школами, что упрощало анализ полученных тематических групп и позволяло сопоставлять автоматически выделенные направления с экспертной структурой карты Кастеллани. Кроме того, использование авторских связей улучшало читаемость итоговой визуализации и делало структуру карты более информативной.

Информация об авторах также использовалась при формировании подписей тематических узлов. Для части крупных тематических направлений в подписи дополнительно включались наиболее характерные токены темы, имена исследователей, наиболее связанных с тематикой, а также временные характеристики соответствующего направления исследований.

Использование данного подхода позволяло одновременно отображать тематическое содержание научной области, исследовательские направления и интеллектуальные связи между ними.

6.4 Визуализация тематической карты

После формирования иерархической тематической структуры выполнялся этап визуализации тематической карты. На данной стадии строилась статическая иерархическая карта без учёта временной компоненты. Основной целью визуализации являлось представление структуры научной области и отображение связей между различными уровнями тематической иерархии.

Тематическая карта задавалась в виде графовой структуры:

$$G = (V, E) \tag{25}$$

где:

- V — множество тематических узлов;
- E — множество связей между узлами графа.

Вершинами графа выступали темы различных уровней иерархической модели. Рёбра между узлами отражали отношения типа parent–child между уровнями тематической иерархии.

Дополнительно использовались связи тематической близости, вычисляемые на основе косинусной меры сходства тематических распределений:

$$sim(t_i, t_j) = \frac{\phi_i \cdot \phi_j}{|\phi_i||\phi_j|} \tag{26}$$

где:

- ϕ_i — вектор токенов темы t_i;
- ϕ_j — вектор токенов темы t_j.

Для построения графовой структуры (рисунок 14) использовалась библиотека NetworkX. После формирования граф экспортировался в формат GEXF и дополнительно обрабатывался в среде Gephi.

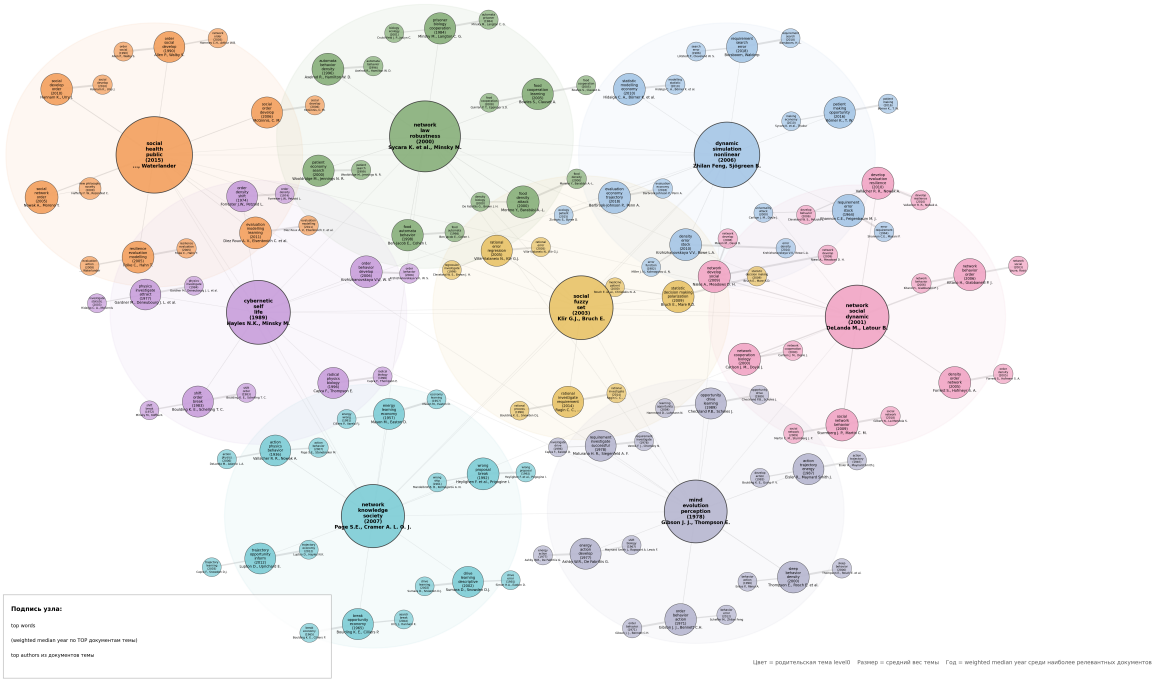

Рис. 14: Иерархическая тематическая карта

При построении визуализации применялись алгоритмы размещения узлов, основанные на силовых моделях графа. Узлы, соответствующие тематически близким направлениям, автоматически располагались на небольшом расстоянии друг от друга, формируя локальные тематические кластеры.

Для отображения принадлежности тем к крупным исследовательским направлениям использовалось цветовое кодирование. Дочерние темы наследовали цвет соответствующей родительской темы верхнего уровня. Такой подход позволял визуально выделять:

- крупные исследовательские направления;
- тематически связанные группы;
- междисциплинарные связи между различными областями.

Размер тематических узлов определялся средним тематическим весом темы внутри корпуса документов. Более крупные узлы соответствовали темам с высокими значениями theta-весов и большим количеством связанных публикаций.

Связи между уровнями графа отражали иерархическую организацию тематической модели:

- верхние уровни соответствовали обобщённым исследовательским направлениям;
- нижние уровни содержали специализированные тематические области;
- дочерние узлы уточняли содержание родительских тем.

Дополнительно на карте отображались подписи ключевых тематических направлений и авторов. Для наиболее значимых тематических узлов визуализировались:

- ключевые токены темы;
- наиболее характерные авторы;
- основные исследовательские направления.

Полученная визуализация позволяла отобразить:

- структуру исследуемой научной области;
- взаимосвязи между тематическими направлениями;
- степень тематической специализации;
- иерархическую организацию научных тем.

6.5 Построение хронологической тематической карты

После построения иерархической тематической структуры выполнялось добавление временной компоненты и формирование хронологической тематической карты. Основной задачей данного этапа являлось отображение эволюции научных направлений во времени.

Для каждой темы использовался репрезентативный год, вычисленный на основе взвешенной медианы. Полученное значение преобразовывалось в координату по оси X:

$$x_t = \alpha(y_t - y_{min}) \tag{27}$$

где:

- x_t — координата темы по временной оси;
- y_t — репрезентативный год темы;
- y_{min} — минимальный год корпуса;

- α — коэффициент масштабирования.

Подобное преобразование обеспечивало линейное отображение тем относительно временной шкалы.

Темы автоматически распределялись по десятилетиям. Более ранние темы располагались в левой части карты, а современные тематические направления — в правой части. На рисунке 15 представлена иоговая визуализация.

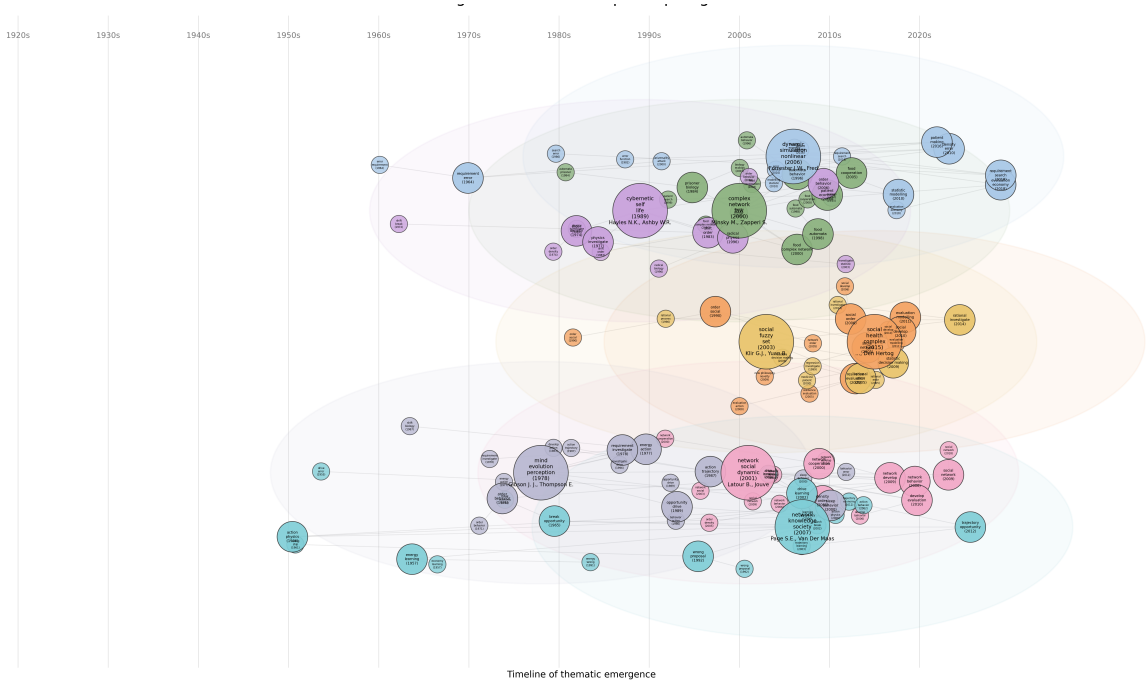

Рис. 15: Хронологическая тематическая карта

Визуализация позволяла проследить temporal dynamics исследуемой научной области. Наиболее ранние тематические направления были связаны с:

- кибернетикой;
- системной теорией;
- математическими моделями динамических систем.

После 1990 х годов наблюдался заметный рост числа тематических направлений. Наиболее активно развивались области сетевого анализа, computational social science, complex systems и междисциплинарные вычислительные методы.

Современные темы формировали плотные тематические группы с высокой степенью междисциплинарности и большим количеством взаимосвязей между исследовательскими направлениями.

Цветовое кодирование использовалось для сохранения принадлежности тематик к направлениям верхнего уровня. Размер узлов отражал значимость темы внутри корпуса публикаций.

Визуализация дополнительно включала связи между уровнями иерархии, тематические группы, отношения тематической близости и динамику развития исследовательских направлений.

Итоговая карта представляла собой многоуровневый хронологический тематический граф, отражающий структуру научной области, развитие тематик во времени, взаимосвязи исследовательских направлений и появление новых тематических областей.

Полученная визуализация позволяла анализировать периоды возникновения тематик, развитие междисциплинарных направлений, эволюцию scientific complexity landscape и temporal dynamics тематической структуры корпуса.

7 Сравнение полученной модели с картой Кастеллани

7.1 Постановка задачи сравнения

Одной из целей данного исследования являлась оценка соответствия автоматически построенной тематической структуры экспертной карте Кастеллани. Поскольку карта Кастеллани представляет собой вручную сформированную структуру научной области complexity science, в рамках работы она рассматривалась как эталонная экспертная модель.

На этапе формирования датасета для каждой публикации задавалась экспертная тематическая принадлежность по нескольким уровням иерархии

- `castellani_level0`;
- `castellani_level1`;
- `castellani_level2`.

В процессе обучения иерархической модели BigARTM формировались собственные тематические кластеры документов. Таким образом, для каждого документа присутствовали два варианта тематической принадлежности

- экспертная тематическая разметка;
- тематическая принадлежность, автоматически полученная в результате работы модели.

Основная задача сравнения заключалась в оценке степени согласованности между двумя вариантами тематического разбиения

- экспертной тематической структурой;
- автоматически построенной тематической моделью.

Сравнение выполнялось отдельно для каждого уровня иерархии. Такой подход позволял оценить, насколько иерархическая модель BigARTM способна воспроизводить крупные исследовательские направления, специализированные тематические области и более детализированные научные тематики.

7.2 Методика оценки качества

Для оценки качества тематической модели использовалась BCubed метрика [44], предназначенная для сравнения двух кластеризаций. В рамках данной работы BCubed применялась для оценки согласованности

- экспертных тематических кластеров;
- тематических кластеров, полученных моделью BigARTM.

BCubed метрика является одной из наиболее устойчивых метрик сравнения кластеризаций, поскольку учитывает принадлежность каждого отдельного документа и корректно работает при различии размеров кластеров. Для научных корпусов подобное свойство особенно важно, поскольку тематические направления представлены неравномерным количеством публикаций.

В рамках реализации для каждого документа определялись:

- истинный кластер;
- предсказанный кластер.

Истинный кластер определялся на основе экспертной разметки Кастеллани. Предсказанный кластер определялся как тема с максимальным значением в матрице Θ:

$$t_d = \arg\max_t \theta_{td} \tag{28}$$

где:

- θ_{td} — вероятность темы t в документе d;
- t_d — итоговая тема документа.

Для каждого документа вычислялась BCubed precision:

$$Precision(i) = \frac{|C_{pred}(i) \cap C_{true}(i)|}{|C_{pred}(i)|} \tag{29}$$

где:

- $C_{pred}(i)$ — множество документов, попавших в тот же кластер модели;
- $C_{true}(i)$ — множество документов с той же экспертной тематикой.

Аналогично вычислялась BCubed recall:

$$Recall(i) = \frac{|C_{pred}(i) \cap C_{true}(i)|}{|C_{true}(i)|} \tag{30}$$

Итоговые значения precision и recall вычислялись как среднее по всем документам корпуса.

Финальная BCubed F1-мера определялась следующим образом:

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{31}$$

Использование метрики BCubed особенно удобно для тематических моделей, поскольку она оценивает согласованность кластеризации документов, а не совпадение названий тем или наборов наиболее вероятных токенов.

7.3 Реализация BCubed-сравнения

Сравнение тематической модели с экспертной структурой Кастеллани выполнялось в виде отдельного pipeline анализа результатов тематического моделирования.

В качестве входных данных использовались матрицы

$$(\Theta)$$

, экспертная разметка документов и тематические уровни Кастеллани.

Сравнение проводилось отдельно для трёх уровней тематической структуры. Для каждого уровня использовались соответствующие экспертные метки. На первом этапе загружались матрицы

$$(\Theta)$$

, полученные после обучения иерархической модели BigARTM. Далее для каждого документа определялась тема с максимальным тематическим весом.

В результате каждому документу сопоставлялись экспертная тема и автоматически определённая тема модели.

Сравнение выполнялось только для документов, присутствующих одновременно в экспертной разметке и в матрицах Θ. После сопоставления документов вычислялись:

- BCubed precision;
- BCubed recall;
- BCubed F1.

Дополнительно для каждого уровня строились таблицы соответствия реальных и предсказанных кластеров (*true/pred clusters*), матрицы ошибок (*confusion matrices*) и таблицы посравнительного анализа документов (*per-document comparison tables*). Подобный подход позволял анализировать устойчивость тематических кластеров, степень смешивания тематик, соответствие экспертной иерархии и качество автоматического разделения научных направлений.

7.4 Результаты сравнения

Итоговые результаты BCubed сравнения представлены в таблице 4.

Таблица 4: Результаты BCubed сравнения

Уровень	Precision	Recall	F1
level0	0.219	0.234	0.226
level1	0.310	0.306	0.308
level2	0.206	0.675	0.316

Дополнительная статистика кластеризации представлена в таблице 5.

Таблица 5: Статистика тематических кластеров

Уровень	Документы	Expert clusters	Predicted clusters
level0	377	9	8
level1	319	53	53
level2	377	185	55

Для нулевого уровня модель показала умеренное качество соответствия экспертной структуре. На данном уровне тематические направления обладают высокой степенью междисциплинарности и содержат большое количество пересекающихся терминов. Это усложняет их автоматическое разделение.

Наиболее устойчивые результаты были получены на первом уровне. Для данного уровня наблюдается наиболее сбалансированное соотношение precision и recall

- precision = 0.310;
- recall = 0.306;
- F1 = 0.308.

Полученные результаты показывают, что модель наиболее устойчиво воспроизводит специализированные тематические области среднего уровня детализации.

Для второго уровня наблюдается наиболее высокое значение recall

$$Recall = 0.675 \tag{32}$$

При этом значение precision заметно ниже

$$Precision = 0.206 \tag{33}$$

Подобное поведение связано с высокой степенью детализации экспертной структуры Кастеллани. Экспертная разметка включает 185 специализированных тематических кластеров, тогда как модель формирует 55 автоматически выделенных групп. В результате несколько экспертных тем объединяются моделью в более крупные тематические области.

Дополнительно были построены таблицы соответствия между экспертными и автоматически полученными темами. Анализ матриц ошибок показал, что модель наиболее устойчиво выделяет направления network science, social systems, computational complexity и systems theory.

Наибольшее смешивание наблюдалось между interdisciplinary topics, social complexity, computational social science и systems oriented направлениями.

7.5 Анализ результатов

Полученные результаты показывают, что иерархическая модель BigARTM способна воспроизводить значительную часть экспертной тематической структуры научного корпуса.

Наиболее устойчивые результаты были получены на среднем уровне иерархии. На данном уровне тематические направления обладают достаточной степенью специализации для устойчивого автоматического разделения. При этом они ещё не характеризуются чрезмерной детализацией, присутствующей в экспертной карте.

Для верхнего уровня иерархии качество кластеризации оказывается ниже, что связано с высокой междисциплинарностью крупных исследовательских направлений. Темы верхнего уровня в области complexity science содержат большое количество общих терминов, одновременно используемых в различных научных областях.

Наиболее сложным для автоматического моделирования оказался нижний уровень иерархии. Несмотря на высокие значения recall, показатель precision остаётся сравнительно низким. Это связано с тем, что специализированные научные направления обладают высокой семантической близостью. Дополнительную сложность создаёт использование схожей терминологии различными интеллектуальными традициями, а также более высокая степень детализации экспертной карты Кастеллани по сравнению с автоматически построенной моделью.

Отдельной особенностью корпуса является его выраженный междисциплинарный характер. Значительная часть публикаций одновременно относится к нескольким тематическим направлениям, тогда как BCubed сравнение использует single label clustering на основе правила

$$\arg\max(\Theta) \tag{34}$$

Подобное упрощение ограничивает возможность полного учёта смешанных тематических распределений документов.

Следует учитывать и ограничения самой метрики BCubed. Данная метрика ориентирована на оценку согласованности кластеризации документов, соответствия тематических групп и структуры распределения документов по темам.

При этом BCubed не учитывает семантическое качество top tokens, временную организацию тематической карты, особенности визуальной структуры и корректность иерархического расположения графа.

Дополнительные ограничения связаны с самой экспертной разметкой Кастеллани. Экспертная структура формировалась вручную и содержит пересекающиеся тематические направления, неоднозначные границы между темами и различную степень детализации исследовательских областей.

Несмотря на перечисленные ограничения, результаты BCubed сравнения показывают, что hierarchical BigARTM воспроизводит значительную часть структуры экспертной карты Кастеллани. Автоматически выделенные тематические направления демонстрируют согласованность с экспертной организацией научного корпуса.

Построенная hierarchical topic model позволяет автоматически формировать тематическую структуру научной области и использовать её при построении хронологической тематической карты. Полученная temporal topic map демонстрирует согласованность с экспертной тематической структурой и предоставляет возможность анализа динамики развития исследовательских направлений в области complexity science.

8 Заключение

В рамках настоящей работы были исследованы методы автоматического построения хронологических тематических карт научных публикаций на основе тематического моделирования и анализа временной структуры научного корпуса. На защиту выносятся следующие результаты:

В рамках настоящей работы были исследованы методы автоматического построения хронологических тематических карт научных публикаций. На защиту выносятся следующие результаты:

- датасет научных публикаций по тематике науки о сложности;
- метод автоматического построения хронологической тематической карты, отличающийся использованием иерархического тематического моделирования и анализа временной структуры публикаций;
- методы визуализации тематической карты, обеспечивающие совместное отображение тематических и временных связей;
- методика сравнения автоматически построенной тематической структуры с экспертной тематической картой.

В ходе исследования был выполнен анализ структуры карты Кастеллани как экспертной тематической модели области complexity science. На основе проведённого анализа была разработана формализованная структура тематической карты, включающая иерархию тематик, связи между направлениями исследований, временные характеристики и тематическую принадлежность публикаций.

Для проведения тематического моделирования был сформирован структурированный корпус научных публикаций, содержащий научные статьи, главы книг и монографии, относящиеся к исследовательским направлениям complexity science. Корпус был подготовлен в формате, пригодном для автоматического тематического анализа и дальнейшего построения тематической карты.

В работе был реализован пайплайн предобработки корпуса, включающий:

- извлечение текстов из PDF-документов;
- очистку научных текстов;
- токенизацию и лемматизацию;
- удаление шумовых фрагментов;
- построение корпуса в формате Vowpal Wabbit.

Дополнительно была реализована система выделения ключевых фраз и тематических anchor-токенов, использовавшихся для стабилизации тематической структуры модели.

На основе подготовленного корпуса была построена иерархическая тематическая модель с использованием библиотеки BigARTM. Реализованная модель включала несколько уровней тематической иерархии и позволяла выделять как крупные научные направления, так и специализированные исследовательские тематики. Для повышения интерпретируемости тем использовались регуляризаторы разреженности и декорреляции.

Одним из ключевых результатов работы стала реализация временной компоненты тематической модели. Для тематических кластеров были вычислены временные характеристики, включая репрезентативный год и распределения тем по годам. Это позволило интегрировать тематическую и временную структуру научного корпуса и построить хронологическую тематическую карту.

В работе была реализована система визуализации тематической структуры, включающая иерархический тематический граф, временное расположение тем, тематические связи и визуализацию авторов вместе с исследовательскими направлениями. Полученная карта позволила отобразить структуру научной области, взаимосвязи тематик, развитие исследовательских направлений во времени и формирование специализированных тематических областей.

Для количественной оценки качества тематической модели было выполнено сравнение с экспертной структурой Кастеллани с использованием метрики кластеризации BCubed. Результаты показали, что иерархическая модель BigARTM воспроизводит значительную часть экспертной тематической структуры и позволяет автоматически выделять устойчивые тематические направления научного корпуса.

Поставленная цель исследования была достигнута. В рамках работы был разработан программный пайплайн автоматического построения хронологических тематических карт на основе научных публикаций. Предложенный подход ориентирован не на воспроизведение конкретной карты Кастеллани, а на создание универсального метода анализа научных корпусов и автоматического формирования тематических карт для различных предметных областей.

Научная новизна работы заключается в объединении иерархического тематического моделирования и анализа временной структуры научных публикаций в рамках единого пайплайна построения тематических карт. В работе реализовано автоматическое формирование

временного отражения, интеграция тематической и временной структуры корпуса, а также использование взвешанного временного отражения для временного позиционирования тематических кластеров.

Дополнительно в работе реализован semi-supervised подход с использованием тематических anchor-токенов, позволяющий частично управлять тематической структурой модели без жёсткой ручной разметки документов.

Несмотря на полученные результаты, проведённое исследование имеет ряд ограничений. Одним из них является неоднородность исходного корпуса научных публикаций. Качество PDF-документов существенно различалось, а часть текстов содержала OCR-артефакты, ошибки распознавания и шумовые фрагменты, возникающие в процессе автоматического извлечения текста. Дополнительную сложность создавал тематический дисбаланс корпуса и неравномерная представленность отдельных исследовательских направлений.

Следует учитывать и ограничения экспертной структуры карты Кастеллани. В области complexity science тематические границы часто являются пересекающимися, а отдельные публикации могут одновременно относиться к нескольким исследовательским направлениям. Особенно это характерно для междисциплинарных работ, объединяющих методы и подходы из различных научных областей. Подобные особенности усложняют задачу автоматического тематического моделирования и затрудняют количественное сравнение автоматически полученных тематических структур с экспертной разметкой.

Дальнейшее развитие работы может быть связано с расширением корпуса научных публикаций и использованием более крупных текстовых коллекций. Дополнительный интерес представляет применение динамических тематических моделей, позволяющих более точно анализировать изменение тематических направлений во времени. Также перспективным направлением является реализация автоматического обновления тематической карты при появлении новых публикаций и изменении структуры научной области.

Перспективным направлением дальнейших исследований является интеграция методов семантического анализа текстов и моделей векторных представлений документов. Использование подобных подходов может повысить качество тематической кластеризации, улучшить устойчивость тематических связей и более точно учитывать семантическую близость научных публикаций.

Отдельное направление развития связано с совершенствованием методов визуализации тематических карт. В частности, интерес представляет построение интерактивных хронологических тематических графов, позволяющих исследовать развитие научных направлений и анализировать изменение структуры тематических связей во времени.

Таким образом, результаты проведённого исследования показывают, что методы иерархического тематического моделирования позволяют автоматически формировать хронологические тематические карты научных публикаций и использовать их для анализа структуры, взаимосвязей и динамики развития сложных междисциплинарных научных областей.

Список литературы

[1] Chaomei Chen. Science mapping: a systematic review of the literature. *Journal of data and information science*, 2(2), 2017.

[2] Chaomei Chen. Mapping science. In *Mapping Scientific Frontiers: The quest for knowledge visualization*, pages 259–320. Springer, 2013.

[3] Eugenio Petrovich. Science mapping. *ISKO Encyclopedia of Knowledge Organization*, 2020.

[4] Luis de Marcos, Manuel Goyanes, and Adrián Domínguez-Díaz. Mapping science through editorial board interlocking: connections and distance between fields of knowledge and institutional affiliations. *Scientometrics*, 129(6):3385–3406, 2024.

[5] Loet Leydesdorff. Various methods for the mapping of science. *Scientometrics*, 11(5-6):295–324, 1987.

[6] Vilker Zucolotto Pessin, Luciana Harue Yamane, and Renato Ribeiro Siman. Smart bibliometrics: an integrated method of science mapping and bibliometric analysis. *Scientometrics*, 127(6):3695–3718, 2022.

[7] Fidelia Ibekwe-SanJuan. Information science in the web era: A term-based approach to domain mapping. *Proceedings of the American Society for Information Science and Technology*, 46(1):1–23, 2009.

[8] Gergely Palla, Gergely Tibély, Enys Mones, Péter Pollner, and Tamás Vicsek. Hierarchical networks of scientific journals. *Palgrave Communications*, 1(1):1–9, 2015.

[9] Eugenio Petrovich et al. Science mapping and science maps. *Knowledge Organization*, 48(7-8):535–562, 2021.

[10] Brian Castellani. Brian castellani on the complexity sciences. *Theory, Culture & Society*, 2014.

[11] Naveen Donthu, Satish Kumar, Debmalya Mukherjee, Nitesh Pandey, and Weng Marc Lim. How to conduct a bibliometric analysis: An overview and guidelines. *Journal of business research*, 133:285–296, 2021.

[12] Maxwell Mirton Kessler. Bibliographic coupling between scientific papers. *American documentation*, 14(1):10–25, 1963.

[13] Nees Jan Van Eck and Ludo Waltman. Visualizing bibliometric networks. In *Measuring scholarly impact: Methods and practice*, pages 285–320. Springer, 2014.

[14] Mark EJ Newman. The structure of scientific collaboration networks. *Proceedings of the national academy of sciences*, 98(2):404–409, 2001.

[15] Mark EJ Newman. Modularity and community structure in networks. *Proceedings of the national academy of sciences*, 103(23):8577–8582, 2006.

[16] Santo Fortunato. Community detection in graphs. *Physics reports*, 486(3-5):75–174, 2010.

[17] Nees Van Eck and Ludo Waltman. Software survey: Vosviewer, a computer program for bibliometric mapping. *scientometrics*, 84(2):523–538, 2010.

[18] José A Moral-Muñoz, Enrique Herrera-Viedma, Antonio Santisteban-Espejo, and Manuel J Cobo. Software tools for conducting bibliometric analysis in science: An up-to-date review. *Profesional de la Información*, 29(1), 2020.

[19] Ludo Waltman, Nees Jan Van Eck, and Ed CM Noyons. A unified approach to mapping and clustering of bibliometric networks. *Journal of informetrics*, 4(4):629–635, 2010.

[20] Ike Vayansky and Sathish AP Kumar. A review of topic modeling methods. *Information Systems*, 94:101582, 2020.

[21] Aly Abdelrazek, Yomna Eid, Eman Gawish, Walaa Medhat, and Ahmed Hassan. Topic modeling algorithms and applications: A survey. *Information Systems*, 112:102131, 2023.

[22] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[23] Simeng Qin, Mingli Zhang, Haiju Hu, and Gang Li. A joint-training topic model for social media texts. *Humanities and Social Sciences Communications*, 12(1):1–16, 2025.

[24] Pengbo Mao, Hegang Chen, Yanghui Rao, Haoran Xie, and Fu Lee Wang. Contrastive learning for hierarchical topic modeling. *Natural Language Processing Journal*, 6:100058, 2024.

[25] Delvin Ce Zhang, Menglin Yang, Xiaobao Wu, Jiasheng Zhang, and Hady W. Lauw. Hierarchical graph topic modeling with topic tree-based transformer. *arXiv preprint*, 2025.

[26] NH Ghassemi Marius Kloft Sophie Fellenz CK James, M Nagda. Evaluating dynamic topic models. In *Proceedings of ACL*, 2024.

[27] H Al-Sayeh KU Sattler N Onah, N Steinmetz. Experimental evaluation of dynamic topic modeling algorithms. *arXiv preprint*, 2025.

[28] Hamed Rahimi, Hubert Naacke, Camelia Constantin, and Bernd Amann. Antm: An aligned neural topic model for exploring evolving topics. *arXiv preprint*, 2023.

[29] Daniel D Lee and H Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788–791, 1999.

[30] Aliya Nugumanova, Almas Alzhanov, Aiganym Mansurova, Kamilla Rakhymbek, and Yerzhan Baiburin. Semantic non-negative matrix factorization for term extraction. *Big Data and Cognitive Computing*, 7:72, 2024.

[31] Jianyu Wang and Xiao-Lei Zhang. Deep nmf topic modeling. *Neurocomputing*, 482:230–245, 2024.

[32] He Zhao, Dinh Phung, Viet Huynh, Yuan Jin, Lan Du, and Wray Buntine. Topic modelling meets deep neural networks: A survey. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, 2021.

[33] R Klimas V Osinska. Mapping science: tools for bibliometric and altmetric studies. *Information Research*, 2021.

[34] Humberto Arruda, Edison Renato Silva, Marcus Lessa, Domício Proença Jr, and Roberto Bartholo. Vosviewer and bibliometrix. *Journal of the Medical Library Association: JMLA*, 110(3):392, 2022.

[35] Hamid Derviş. Bibliometric analysis using bibliometrix an r package. *Journal of scientometric research*, 8(3):156–160, 2019.

[36] Mohd Zaidi Bin Amiruddin, Achmad Samsudin, Andi Suhandi, Bayram Coştu, and Binar Kurnia Prahani. Scientific mapping and trend of conceptual change: A bibliometric analysis. *Social Sciences & Humanities Open*, 11:101208, 2025.

[37] Hamid Reza Saeidnia, Elaheh Hosseini, Shadi Abdoli, and Marcel Ausloos. Unleashing the power of ai. a systematic review of cutting-edge techniques in ai-enhanced scientometrics, webometrics, and bibliometrics. *arXiv preprint*, 2024.

[38] Adela Bara and Simona-Vasilica Oprea. Ai-augmented bibliometric framework: A paradigm shift with agentic ai for dynamic, snippet-based research analysis. *arXiv preprint*, 2025.

[39] Brian Castellani. Complexity map. `https://art-sciencefactory.com/complexity-map_feb09.html`, 2009. [Электронный ресурс]. Дата обращения: 05.01.2026.

[40] Thomas Hofmann. Probabilistic latent semantic analysis. *arXiv preprint arXiv:1301.6705*, 2013.

[41] Konstantin Vorontsov and Anna Potapenko. Additive regularization of topic models. *Machine Learning*, 101(1):303–323, 2015.

[42] 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*, pages 370–381. Springer, 2015.

[43] Elena Pavlova. bigartm-thesis-pipeline. `https://github.com/elena-pavlova-d/bigartm-thesis-pipeline`, 2026. GitHub repository.

[44] Enrique Amigó, Julio Gonzalo, Javier Artiles, and Felisa Verdejo. A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Information retrieval*, 12(4):461–486, 2009.