Пермский филиал федерального государственного автономного
образовательного учреждения высшего образования
Национальный исследовательский университет
«Высшая школа экономики»

Факультет социально-экономических и компьютерных наук

Коркодинов Данил Александрович

РАЗРАБОТКА ПРОГРАММНЫХ СРЕДСТВ ДЛЯ КЛАСТЕРИЗАЦИИ СООБЩЕСТВ В СОЦИАЛЬНЫХ МЕДИА

Выпускная квалификационная работа

студента образовательной программы «Программная инженерия»
по направлению подготовки 09.03.04 Программная инженерия

Руководитель
к.ф-м.н, доцент, доцент
кафедры информационных
технологий в бизнесе

_____

Е. Б. Замятина

Пермь, 2025 год

Аннотация

Выпускная квалификационная работа выполнена студентом 4 курса направления подготовки 09.03.04 «Программная инженерия» по теме: «Разработка программных средств для кластеризации сообществ в социальных медиа».

Руководитель: Замятина Елена Борисовна, к.ф-м.н, доцент, доцент кафедры информационных технологий в бизнесе.

Объём работы: 60 страниц.

Работа состоит из трёх глав. В первой главе рассмотрены теоретические аспекты анализа социальных сетей, приведены определения ключевых понятий, охарактеризованы основные типы социальных платформ, описаны критерии выделения онлайн-сообществ и рассмотрены методы кластеризации на основе графовой структуры и содержания. Также приведён обзор алгоритмов и критериев оценки качества кластеризации.

Во второй главе изложен процесс проектирования и реализации программной системы. Описана архитектура решения, приведены обоснования выбора технологий и реализованные модули: сбора данных с помощью VK API, очистки и нормализации информации, построения графов и запуска кластеризации. Проведён анализ пользовательского интерфейса и визуализация полученных результатов.

Третья глава посвящена тестированию и экспериментальной оценке эффективности системы. Проведено сравнение различных подходов к кластеризации, выявлены наиболее влиятельные пользователи и проанализирована эмоциональная окраска публикаций в сообществах. Представлены выводы о работоспособности предложенного решения и предложены направления дальнейшего развития.

Результаты работы могут быть применены в аналитических системах мониторинга социальных медиа, маркетинговых исследованиях и при построении рекомендательных сервисов.

Оглавление

Введение

Развитие цифровых технологий и широкое распространение социальных сетей сильно изменили способы взаимодействия людей и обмена информацией. Миллионы пользователей ежедневно публикуют контент, комментируют записи, вступают в сообщества и устанавливают социальные связи. Эти данные представляют значительный интерес как для научных исследований, так и для прикладных задач в бизнесе, маркетинге, социологии и других областях.

Одной из важнейших характеристик социальной сети является её внутренняя структура — то, как пользователи группируются в сообщества, взаимодействуют друг с другом, образуют локальные группы интересов. Выявление таких сообществ позволяет более глубоко понять поведение пользователей, выделить ключевых участников и исследовать информационные потоки внутри сети.

Особый интерес в этом контексте представляет социальная сеть ВКонтакте (VK) — крупнейшая платформа в русскоязычном сегменте интернета. Благодаря своей массовости и разнообразию форм взаимодействия между пользователями, VK предоставляет богатую почву для анализа структуры социальных связей.

Однако из-за объёма данных, высокой динамичности и неявной структуры сообществ возникает необходимость в разработке специализированных программных средств, способных автоматически анализировать и выявлять эти скрытые группировки. Такие инструменты могут быть полезны для построения рекомендательных систем, таргетинга, мониторинга общественного мнения и других задач.

Объектом моего исследования являются пользователи социальной сети VK и их взаимодействия, предметом – программные средства и методы, используемые для анализа и выделения сообществ в социальных сетях.

Целью работы является создание программного средства для кластеризации сообществ пользователей в социальной сети ВКонтакте.

Для достижения цели были разработаны и поставлены следующие задачи:

1. Проанализировать существующие методы кластеризации и возможности их применения для выявления сообществ в социальных сетях.
2. Проанализировать особенности структуры социальных сетей и понятие сообщества.
3. Определить критерии, по которым можно выделять сообщества в социальной сети.
4. Реализовать систему сбора и предварительную обработку данных из VK.
5. Разработать приложение, позволяющего проводить кластеризацию пользователей социальных сетей.
6. Оценить работоспособность и эффективность разработанного приложения на практике.

Практическая значимость данной работы заключается в возможности применения разработанного программного средства для прикладного анализа пользовательских сообществ в социальной сети ВКонтакте. Выделение скрытых групп позволяет более точно сегментировать аудиторию, выявлять лидеров мнений, анализировать интересы пользователей и повышать эффективность взаимодействия с целевыми группами. Такие возможности особенно актуальны для маркетинга, социологических исследований, информационной аналитики и построения рекомендательных систем, где важно учитывать структуру и характер связей внутри социальной сети.

Глава 1 Теоретические основы кластеризации сообществ в социальных медиа

В данной главе рассматриваются ключевые теоретические аспекты, необходимые для понимания процесса кластеризации сообществ в социальных медиа. Приводятся определения социальных сетей и онлайн-сообществ, а также их роль в современном цифровом обществе. Освещаются основные подходы к анализу структуры социальных графов, включая методы кластеризации и алгоритмы, используемые для выявления сообществ в графовых данных. Отдельное внимание уделяется применению методов машинного обучения для решения задач кластеризации, а также рассмотрению метрик, позволяющих оценить качество полученных кластеров. Глава служит теоретическим фундаментом для последующего проектирования и реализации программной системы.

1.1 Понятие социальных сетей и онлайн-сообществ

В современном мире социальные сети играют ключевую роль как в повседневной жизни людей, так и в функционировании общества в целом. Они служат не только средством для мгновенного обмена информацией, но и инструментом формирования общественного мнения, влияя на восприятие событий через алгоритмы персонализации контента. Кроме того, социальные платформы активно используются для установления и поддержания межличностных и профессиональных связей, объединяя людей по интересам, профессиям и ценностям. Выполняя одновременно коммуникативные, информационные, организационные и аналитические функции, социальные сети становятся важнейшим элементом современной цифровой инфраструктуры, охватывающим практически все сферы социальной жизни. По состоянию на январь 2025 года количество пользователей социальных сетей во всем мире достигло 5,24 миллиарда человек, что составляет 63,9% от всей численности планеты [1]. Это свидетельствует о значительном росте популярности данных платформ и их интеграции в повседневную жизнь людей.

В Российской Федерации наиболее популярной социальной сетью является VKontakte. По прогнозам, количество пользователей VK в России увеличится с 73,79

миллиона в 2024 году до 78,59 миллиона в 2028 году, что представляет собой рост на 6,51% [2]. Такая стабильная динамика роста объясняется глубокой интеграцией VK в цифровую экосистему страны: платформа активно используется не только для личного общения, но и в образовательных, коммерческих и информационных целях. Здесь размещаются официальные страницы государственных учреждений, ведётся бизнес-активность, распространяются новости и организуются мероприятия. Кроме того, VK развивает собственную медиаплатформу, сервисы видео и музыкального контента, что способствует удержанию пользователей и расширению аудитории. Всё это делает VK одним из ключевых каналов цифровой коммуникации в России, с огромным влиянием на социальные процессы и повседневную жизнь граждан.

Прежде всего необходимо разобраться, что лежит в основе понятия «социальные сети». Концепция социальных сетей рассматривается в контексте различных научных подходов, отражающих разнообразие взаимосвязей и влияний между людьми. Различные исследователи и академические круги предлагают разнообразные определения данному понятию, подчёркивая его сложность и многогранность.

Так, российские исследователи Э.Н. Забарная и И.В. Куриленко определяют социальные сети как сообщества в виртуальной среде, объединяющие людей на основе общих интересов или целей [3]. Они подчёркивают, что такие сообщества формируются по принципу самоорганизации и характеризуются горизонтальными связями между участниками.

В социологических теориях социальные сети рассматриваются как структуры, состоящие из множества агентов и определённых связей между ними [4]. Это понимание акцентирует внимание на анализе моделей взаимодействия между акторами для понимания социальной структуры общества.

Психологические исследования фокусируются на влиянии социальных сетей на поведение и эмоциональное состояние пользователей. Например, исследования показывают, что взаимодействие в социальных сетях может вызывать различные эмоциональные переживания, влияющие на общее качество жизни и психологическое благополучие пользователей [5].

Таким образом, анализ различных научных подходов к определению социальных сетей демонстрирует, что это явление обладает высокой степенью комплексности и многогранности. Социальные сети представляют собой не просто совокупность взаимосвязей между людьми, но и динамически развивающиеся структуры, в которых переплетаются личные, профессиональные, культурные и информационные взаимодействия. Исследователи подчёркивают, что характер этих связей способен оказывать существенное влияние на поведение индивидов, их восприятие социальной реальности, эмоциональное состояние и процессы принятия решений. Более того, социальные сети играют ключевую роль в формировании и трансляции норм, ценностей и идеологических установок, выступая одновременно как механизм социализации, канал коммуникации и инструмент управления вниманием массовой аудитории. В условиях цифровой трансформации общества социальные сети становятся важнейшим элементом современной информационной инфраструктуры, активно влияющим на социокультурные, политические и экономические процессы, включая образование, маркетинг, государственное управление, здравоохранение и многое другое. Их исследование требует междисциплинарного подхода и глубокого понимания как технических, так и гуманитарных аспектов цифрового взаимодействия.

Для глубокого понимания феномена социальных сетей важно учитывать их многообразие, обусловленное различиями в функциональности, целевой аудитории, механизмах взаимодействия и сфере применения. Социальные сети не являются однородным явлением — напротив, они представляют собой широкий спектр платформ, каждая из которых выполняет свои уникальные задачи и ориентирована на определённые виды активности. В связи с этим в научной и прикладной литературе принято выделять несколько основных типов социальных сетей, различающихся по назначению и структуре взаимодействия между пользователями.

1. Коммуникативные социальные сети — направлены на личное общение между пользователями. Яркие примеры: VKontakte, WhatsApp, Telegram. Эти платформы предлагают инструменты для обмена сообщениями, публикации постов, создания групп и сообществ, а также мультимедийного контента. Они выполняют

важные функции в поддержании социальных связей, как в частной, так и в профессиональной сферах.

2. Контент-ориентированные сети — сосредоточены на создании, распространении и потреблении мультимедийного контента. К ним относятся такие платформы, как YouTube, TikTok где основное внимание уделяется видеороликам, фото и сторис. Эти сети играют огромную роль в цифровом маркетинге, инфлюенсер-культуре и формировании массовых трендов.

3. Профессиональные социальные сети — ориентированы на деловое общение, поиск работы, рекрутинг и формирование профессионального имиджа. Примером является LinkedIn, где пользователи создают резюме, обмениваются опытом и налаживают деловые контакты. Такие платформы особенно актуальны в контексте цифровой экономики и удалённой занятости.

4. Тематические и нишевые сети — объединяют пользователей по интересам, хобби или профессиональной сфере. Примеры включают GitHub для разработчиков, Goodreads для любителей книг или ResearchGate для учёных. Они способствуют обмену знаниями, совместной работе над проектами и развитию сообществ по интересам.

5. Социальные сети с элементами геймификации — предоставляют пользователям возможность взаимодействовать в рамках игрового или полуигрового процесса. Примеры: Twitch, Steam Community, Discord. Они часто сочетают в себе функции общения, стриминга и управления сообществами.

Каждая из этих категорий обладает своими особенностями и логикой взаимодействия, что требует специфического подхода к анализу данных, моделей поведения пользователей и методов кластеризации. В зависимости от целей исследования и разрабатываемых программных решений важно учитывать тип социальной сети, её структуру, открытость API, формат представления данных и другие технологические аспекты.

Среди множества социальных платформ, действующих в российском цифровом пространстве, особое место занимает ВКонтакте (VK) — крупнейшая и наиболее популярная социальная сеть в России и странах СНГ. Основанная в 2006

году, она изначально задумывалась как аналог западных платформ, таких как Facebook, но быстро развилась в самостоятельную экосистему с уникальным функционалом и культурной спецификой. На сегодняшний день VK насчитывает более 73 миллионов активных пользователей в России [2], обеспечивая широкие возможности для коммуникации, потребления контента и объединения пользователей в сообщества по интересам.

VKontakte предоставляет пользователям богатый набор инструментов: от личных сообщений и публичных постов до интегрированных музыкальных и видеосервисов. Одной из ключевых особенностей платформы является наличие групп и сообществ — пользовательских объединений, которые могут быть посвящены любой теме: от мемов и новостей до науки, бизнеса и образования. Эти сообщества формируют своеобразную внутреннюю структуру ВКонтакте, которая особенно интересна для анализа: группы активно взаимодействуют между собой, имеют пересекающуюся аудиторию, отличаются по степени вовлеченности и активности.

Кроме того, VK предлагает развитую API-инфраструктуру, предоставляющую доступ к данным о пользователях, группах, комментариях, репостах и других активностях. Это делает платформу удобной для автоматизированного сбора и анализа данных в исследовательских и инженерных целях. Благодаря этому VK активно используется в научных и прикладных исследованиях по анализу социальных графов, распространению информации, выявлению сообществ и поведению пользователей в цифровой среде.

Отдельного внимания заслуживает графовая структура VK: каждый пользователь связан с другими пользователями через отношения дружбы, подписки или совместного участия в группах. Эти связи формируют сложный граф, который можно анализировать с помощью алгоритмов машинного обучения и методов сетевого анализа. Такая структура идеально подходит для задач кластеризации, позволяющей выявить сообщества внутри социальной сети, оценить их плотность, активность и взаимодействие между собой.

Также следует отметить значительное влияние VK на информационное пространство: через платформу распространяются новости, культурные тренды,

маркетинговые кампании и социальные инициативы. В связи с этим VK становится не только объектом пользовательской активности, но и полем для социологических, маркетинговых и политических исследований.

Сообщества во ВКонтакте представляют собой одну из центральных структурных единиц платформы, отражающую интересы, цели и модели взаимодействия миллионов пользователей. Они могут выполнять различные функции: от развлекательных и информационных до образовательных, коммерческих и активистских. Благодаря гибкости инструментов, предоставляемых платформой, пользователи создают сообщества, варьирующиеся по масштабу, тематике, формату публикаций и модели вовлечения аудитории. Для корректного анализа и кластеризации таких объединений необходимо понимать, какие типы сообществ существуют, чем они различаются и какие параметры могут быть использованы для их классификации.

Во ВКонтакте существует множество сообществ, ориентированных на определённые географические области. Они различаются по охвату аудитории и характеру публикуемого контента в зависимости от локализации. В этом контексте можно выделить следующие уровни территориальной привязки:

1. Локальные – посвящены конкретному городу, району, учебному заведению или организации (например, «Типичный Екатеринбург», «СПбГУ — абитуриенты»)
2. Региональные – охватывают широкие территории (например, «Новости Красноярского края»)
3. Федеральные/глобальные – не имеют географической привязки и ориентированы на широкую аудиторию по всей стране или за её пределами (например, «Лентач», «Наука и жизнь»).

Сообщества используют различные способы подачи информации, ориентируясь на потребности своей аудитории. Это могут быть текстовые публикации, визуальные материалы, интерактивные элементы и другие медиаформаты. Основные форматы взаимодействия включают в себя:

1. Новостные сообщества – публикуют актуальные события, новости, аналитику
2. Развлекательные сообщества – ориентированы на юмор, мемы, развлекательные видео
3. Образовательные – делятся полезными материалами, лекциями, книгами
4. Тематические группы – фокусируются на конкретной теме: программирование, фитнес, книги, музыка и т. д.
5. Магазины и бренды – коммерческие страницы, продвигающие товары и услуги

Некоторые группы ориентированы на активное вовлечение участников в коммуникацию, в то время как другие предполагают пассивное потребление контента. Такая разница определяет степень участия и тип обратной связи. В зависимости от модели взаимодействия можно выделить следующие типы:

1. Открытые дискуссионные площадки – с активными комментариями, опросами, прямыми эфирами
2. Односторонние трансляторы – публикация контента без активного обсуждения
3. Интерактивные сообщества – с элементами геймификации, вовлекающие пользователей в конкурсы, челленджи, марафоны и т. д.

Количество подписчиков, активность участников и степень доверия к публикуемому контенту позволяют оценить влияние сообщества. Это влияет на масштаб воздействия на аудиторию и потенциал для распространения информации. В этом контексте сообщества можно классифицировать следующим образом:

1. Микросообщества – с аудиторией до 10 тысяч человек
2. Средние сообщества – от 10 до 100 тысяч человек
3. Крупные сообщества – свыше 100 тысяч человек
4. Лидеры мнений – аккаунты, обладающие высокой степенью доверия аудитории и способные влиять на поведение, решения и взгляды подписчиков

Каждое сообщество создается с определённой задачей, которая определяет тематику, структуру и характер публикаций. Цель формирует контентную стратегию и формат общения с подписчиками. Основные цели, по которым можно классифицировать сообщества, включают:

1. Коммерческие – продвижение товаров, услуг, брендов
2. Социальные – обсуждение общественно значимых тем
3. Информационные – передача новостей, статей, экспертных мнений
4. Творческие и фанатские – объединяют по интересам в области музыки, кино, искусства
5. Образовательные и научно-популярные – предоставляют знания, проводят лекции, курсы, обучающие марафоны

Анализ критериев, по которым классифицируются сообщества в социальных сетях, играет важную роль в понимании их структуры и функционирования. Такие параметры, как уровень активности, тематика, географическая привязка и тип взаимодействия, позволяют глубже изучать поведение участников и внутреннюю динамику сообществ. Это, в свою очередь, открывает возможности для оценки их влияния на социокультурные процессы и разработки эффективных стратегий управления в цифровой среде.

На основании проведённого анализа можно сделать вывод, что сообщества в социальной сети ВКонтакте представляют собой сложные и разнородные структуры, формирующие цифровое пространство с высокой степенью динамики и вовлеченности. Их разнообразие проявляется в географической направленности, тематике, форматах взаимодействия, уровне влияния и целевых установках. Такое многообразие обусловлено гибкой архитектурой самой платформы и широким спектром пользовательских интересов, что делает VK одним из наиболее подходящих объектов для исследования сообществ в контексте цифровой социологии и анализа больших данных. Понимание ключевых критериев, по которым можно классифицировать сообщества, позволяет не только структурировать информационное пространство, но и выявлять скрытые закономерности в поведении пользователей, степени их вовлечённости и характере информационных потоков. Это особенно важно для построения систем автоматической кластеризации,

выявления целевых аудиторий, прогнозирования информационного влияния и разработки цифровых стратегий взаимодействия с пользователями. В условиях растущего влияния социальных сетей на общественные, культурные и экономические процессы, системный подход к анализу структуры и функций сообществ становится необходимым инструментом для эффективной навигации и управления в цифровой среде.

1.2 Обзор задач и подходов к анализу сообществ

Анализ сообществ в социальных сетях представляет собой одно из приоритетных направлений как в академических исследованиях, так и в прикладных разработках в сферах маркетинга, безопасности, политического анализа и социологии. Сообщество в данном контексте понимается как группа пользователей, между которыми установлены более плотные и устойчивые связи по сравнению с остальной частью сети. Выявление и изучение таких структур позволяет глубже понять механизмы распространения информации, формирования мнений и организации коммуникации в цифровой среде.

В прикладных задачах анализ сообществ находит применение, например, в маркетинге — при сегментации аудитории, персонализации рекламных стратегий и оценке вовлечённости потребителей [6]. В политико-аналитической и социальной сферах он используется для мониторинга общественных настроений, выявления признаков социальной напряжённости и анализа репутационных рисков. Инструменты мониторинга, такие как «Медиалогия», позволяют отслеживать дискурсы и повестки в реальном времени, что критически важно для государственных и коммерческих структур.

Сообщество в социальной сети может формироваться на основе самых разных факторов — от общих интересов и тематик до географической близости или высокой частоты взаимодействий. Структурно они представляют собой кластеры в графе связей, где вершины — это пользователи, а рёбра — связи между ними. Одной из ключевых характеристик таких структур является плотность связей: чем выше частота взаимодействий между участниками, тем сильнее вовлечённость и сплочённость группы. Немаловажным фактором также является семантическая близость — общность интересов, лексики, подписок и тематик. Уровень активности

пользователей, выражающийся в количестве публикаций, комментариев и реакций, позволяет оценить устойчивость и динамику сообществ как самостоятельных единиц цифровой среды.

Выявление и анализ таких образований предполагает решение ряда исследовательских и прикладных задач. Одной из базовых задач является автоматическое выделение сообществ, или кластеризация. Она направлена на разбиение сети на группы с высокой внутренней связанностью, что отражает реальную структуру взаимодействий между пользователями. Решение этой задачи, как правило, базируется на методах анализа графов, которые позволяют идентифицировать локально плотные области в структуре сети.

После выделения сообществ важно изучить их внутренние характеристики: плотность связей, централизацию, активность участников. Это позволяет оценить, насколько устойчивы и иерархичны данные структуры, а также насколько они подвержены внешним воздействиям. Дополнительным направлением является исследование динамики сообществ во времени — их роста, слияния, распада, миграции участников. Такие процессы могут быть вызваны как изменением интересов пользователей, так и внешними информационными событиями.

Для комплексного понимания природы сообществ необходим также анализ тематической направленности и семантической структуры коммуникации. Изучение контента, публикуемого участниками, позволяет выявлять преобладающие темы, тональность обсуждений и потенциальные области интереса. Эти характеристики важны при определении функциональной роли сообщества и его влияния на повестку дня в цифровом пространстве.

Значимым аспектом анализа является также оценка уровня вовлечённости — то есть активности пользователей внутри группы. Этот параметр позволяет судить о жизнеспособности сообщества и его значимости для участников. Активные, устойчивые сообщества играют важную роль в распространении информации и формировании сетевых трендов. Ещё одной актуальной задачей является прогнозирование поведения участников на основе накопленных данных: например, вероятности отклика на контент, ухода из сообщества или переключения внимания на другие тематики.

Разнообразие задач анализа сообществ обусловливает потребность в использовании различных аналитических подходов. Классически их делят на три группы: структурные, контентные и гибридные. Структурные подходы базируются на анализе графа взаимодействий между пользователями, не учитывая при этом содержательное наполнение коммуникации. Они хорошо подходят для задач выявления сообществ и построения их структуры, особенно при наличии больших объёмов данных о связях. Контентные подходы, напротив, опираются на анализ текстов и другого медиаконтента. С их помощью можно исследовать интересы, тематику и эмоциональный фон внутри сообщества. Гибридные методы сочетают преимущества обоих подходов, обеспечивая более точный и глубокий анализ за счёт интеграции структурной и семантической информации.

На современном этапе особое внимание уделяется разработке методов, способных эффективно работать с большими графами и неполными данными. Среди них можно выделить графовые нейронные сети (GNN), способные одновременно учитывать структуру сети и свойства узлов. Несмотря на высокую эффективность, такие модели требуют значительных вычислительных ресурсов и большого объёма размеченных данных, что ограничивает их практическое применение в отдельных случаях.

В рамках данной работы основное внимание уделяется применению методов бикластеризации для выявления сообществ в социальной сети ВКонтакте. В отличие от традиционных методов кластеризации, которые группируют объекты по схожести признаков в одном измерении, бикластеризация позволяет одновременно анализировать два измерения — объекты и признаки. Это особенно актуально для анализа социальных сетей, где важна не только структура связей между пользователями, но и содержание их активности. Метод спектральной бикластеризации хорошо зарекомендовал себя в задачах с разреженными и высокоразмерными данными, поскольку устойчив к шуму и способен выявлять скрытые зависимости между группами пользователей и тематическими характеристиками.

Выбор структурного подхода с применением бикластеризации обусловлен особенностями доступных во ВКонтакте данных. Платформа предоставляет

широкий спектр информации о взаимодействиях пользователей — включая членство в группах, подписки и активность в публикациях. Эти данные позволяют построить граф, в котором пользователи и сообщества представлены как связанные сущности. Бикластеризация даёт возможность выделять подмножества пользователей, схожих между собой по модели взаимодействия с контентом или группами, что повышает точность выявления тематически устойчивых сообществ и делает анализ более интерпретируемым.

Таким образом, в результате анализа были обоснованы методологические основы подхода, применяемого в данной работе. Он сочетает формальную точность структурного анализа с гибкостью бикластеризации, что позволяет эффективно решать задачи выявления и изучения сообществ в рамках социальной сети ВКонтакте. Предложенный подход послужит основой для построения программной системы кластеризации, которая будет реализована и протестирована в рамках последующих этапов проекта.

1.3 Алгоритмические основы кластеризации графов в задачах анализа сообществ

Кластеризация графов является одним из базовых методов при анализе структуры социальных сетей. Она предполагает разбиение множества узлов графа на такие группы, в пределах которых связи между участниками более плотные, чем между элементами разных групп. В контексте анализа социальных сетей задача кластеризации позволяет выделять сообщества пользователей, схожих по модели взаимодействий, интересам и другим признакам. Это, в свою очередь, даёт возможность глубже понять внутреннюю организацию цифрового общества, характер распространения информации, наличие локальных центров влияния и степень связанности отдельных сегментов сети.

Алгоритмы кластеризации графов активно используются в исследованиях, связанных с социальной сегментацией, моделированием поведения пользователей, определением информационных барьеров и маршрутов передачи контента. Их эффективность и применимость напрямую зависят от структуры исследуемой сети: объёма данных, степени разреженности, наличия перекрывающихся кластеров,

уровня зашумлённости и других факторов. В этой связи особое значение приобретает правильный выбор подхода, который должен учитывать как особенности входных данных, так и цели анализа.

Одним из широко распространённых методов является оптимизация модульности, на которой основан алгоритм Louvain. Он демонстрирует высокую скорость и способность масштабироваться на графы с миллионами узлов. Алгоритм итеративно объединяет узлы и кластеры, максимизируя значение модульности — показателя, отражающего плотность внутренних связей по отношению к внешним. Такие методы позволяют эффективно решать задачи разбиения графа без необходимости предварительного задания количества кластеров и с высокой устойчивостью к шуму, что делает их особенно полезными при работе с большими объёмами данных из социальных сетей.

Наряду с классическими методами всё большее распространение получают эвристические подходы, вдохновлённые природными процессами. Биоинспирированные алгоритмы, в числе которых генетические алгоритмы, алгоритмы муравьиной колонии и алгоритмы роя частиц, показывают высокую гибкость и адаптивность при решении задач кластеризации в условиях неопределённости. Генетические алгоритмы, имитирующие эволюционные механизмы, работают с популяцией решений, где каждое представляет собой возможное разбиение графа. Алгоритм муравьиной колонии моделирует коллективное поведение агентов, которые перемещаются по графу, оставляя следы в виде феромонов и постепенно выявляя области наибольшей плотности связей. Алгоритм роя частиц основывается на коллективной оптимизации и позволяет приближённо находить глобальные экстремумы, исследуя пространство возможных разбиений на основе исторической информации о траекториях движения частиц. Преимуществом таких методов является способность находить скрытые и неочевидные кластеры, устойчиво работать в условиях отсутствия полной информации о структуре сети и эффективно обходить локальные минимумы в пространстве решений.

Развитие методов машинного обучения в последние годы привело к появлению новых направлений в области кластеризации графов. В частности,

получили широкое распространение графовые нейронные сети, которые позволяют моделировать сложные зависимости в графовых структурах и одновременно учитывать как структуру сети, так и свойства её узлов. Такие модели, как Graph Convolutional Networks и Graph Attention Networks, демонстрируют высокую точность в задачах классификации и прогнозирования, однако требуют большого объёма размеченных данных и значительных вычислительных ресурсов, что ограничивает их применение в рамках проектов, нацеленных на универсальность и практическую реализуемость без предварительной подготовки обучающей выборки.

В данной работе приоритет был отдан структурному подходу с использованием биоинспирированных алгоритмов кластеризации. Это обусловлено, с одной стороны, высокой информативностью графа связей пользователей ВКонтакте, включающего данные о дружбе, подписках, совместном участии в сообществах, а с другой — необходимостью работы в условиях частичной и потенциально зашумлённой информации. Биоинспирированные алгоритмы не требуют жёстких предпосылок о распределении данных, устойчивы к неполному графу и позволяют выявлять как плотные, так и разреженные сообщества. Их применение обеспечит необходимую гибкость при построении программного решения, способного адаптироваться к реальной структуре социальной сети.

Таким образом, алгоритмическая база кластеризации графов охватывает широкий спектр методов — от классических, основанных на спектральных преобразованиях и модульности, до современных эвристических подходов и нейросетевых моделей. В контексте данной работы использование биоинспирированных алгоритмов позволяет эффективно решать задачу выявления сообществ в графовой модели социальной сети, обеспечивая баланс между точностью, адаптивностью и вычислительной устойчивостью.

1.4 Критерии оценивания качества кластеризации

Оценка качества кластеризации является неотъемлемой частью процесса анализа графов и выявления сообществ. Без количественной и качественной оценки результатов невозможно судить об эффективности используемого алгоритма, корректности выделения кластеров и пригодности модели для дальнейшего

практического применения. Особенно актуальна проблема выбора метрик в задачах, где отсутствует априорная информация о реальной структуре данных, как это часто бывает при анализе социальных сетей.

Критерии оценки качества кластеризации можно условно разделить на внутренние и внешние. Внутренние метрики основываются исключительно на свойствах полученных кластеров и структуры графа, не требуя заранее известных меток или эталонного разбиения. Внешние, напротив, предполагают наличие истинной разметки и используются, как правило, в экспериментальных условиях или при наличии экспертной информации.

Одним из наиболее популярных и широко применяемых внутренних критериев в анализе социальных сетей является модульность. Она измеряет разницу между долей рёбер, находящихся внутри кластеров, и ожидаемым значением этой доли в случайной модели графа с аналогичным распределением степеней. Значения модульности варьируются от –1 до 1, где более высокие значения указывают на наличие выраженной кластерной структуры. Модульность особенно чувствительна к плотности внутренних связей и подходит для оценки качества структурных разбиений графов. Однако при работе с крупными и разреженными сетями она может демонстрировать эффект смещения в сторону укрупнения кластеров, что следует учитывать при интерпретации результатов.

Для анализа компактности и разделимости кластеров могут применяться коэффициенты плотности и внутрикластерной связанности. Эти метрики позволяют оценить степень сплочённости внутри сообществ и слабость связей между ними, что особенно важно при анализе перекрывающихся или разнородных кластеров. Дополнительную информацию может дать использование коэффициента силуэта, адаптированного к графовым структурам, который отражает среднюю разницу между расстоянием до собственного кластера и ближайшего чужого.

В случае, если доступны метки истинной принадлежности к сообществам (например, при симуляции данных или наличии эталонной структуры), применяются внешние метрики, такие как индекс Рэнда, скорректированный индекс Рэнда (Adjusted Rand Index, ARI), индекс Джаккара и нормализованная взаимная информация (Normalized Mutual Information, NMI). Эти показатели позволяют

количественно сравнивать полученное разбиение с эталоном и использовать их при сравнении различных алгоритмов или параметров.

Наконец, в прикладных задачах всё большую роль играют интерпретируемые метрики, отражающие не только структурную обоснованность кластеров, но и их содержательное наполнение. Например, может оцениваться тематическая однородность кластеров при анализе текстов, активность пользователей внутри сообщества, стабильность кластеров во времени и другие прикладные характеристики.

В рамках данной работы основной упор будет сделан на использование модульности как базовой метрики оценки структурного качества кластеризации. Дополнительно планируется учитывать плотность связей, распределение размеров кластеров и устойчивость результатов при многократном запуске алгоритма с различными начальными условиями. Комплексный подход к оценке позволит не только выбрать наиболее подходящий метод разбиения, но и провести содержательный анализ полученной структуры социальной сети ВКонтакте.

Глава 2 Проектирование и разработка программной системы

Анализ, проведённый в первой главе, позволил определить основные задачи и подходы к кластеризации сообществ в социальных сетях, а также обосновать выбор структурного графового представления и применение биоинспирированных алгоритмов. На основе полученных теоретических данных в данной главе осуществляется переход к практической части — проектированию и реализации программной системы, предназначенной для автоматизированного сбора данных из социальной сети ВКонтакте, построения графа взаимодействий между пользователями и последующей кластеризации выявленных сообществ.

Основная цель данной главы — описание архитектуры, компонентов и логики функционирования программного комплекса, обеспечивающего полный цикл обработки данных: от их извлечения и хранения до анализа и визуализации результатов. Особое внимание будет уделено обоснованию применяемых технологий, построению модели хранения информации, выбору параметров алгоритмов кластеризации, а также пользовательскому интерфейсу, обеспечивающему доступ к полученным данным и результатам анализа.

Структура главы соответствует этапам жизненного цикла разработки: от постановки задачи и определения функциональных требований до реализации отдельных модулей и их интеграции в единую систему. Такой подход позволяет не только обеспечить полноту описания, но и проследить соответствие архитектурных решений целям и задачам проекта.

2.1 Постановка задачи и определение требований

Результаты теоретического анализа, представленные в первой главе, позволили сформулировать общую концепцию программного решения, направленного на автоматизированное выявление сообществ пользователей в социальной сети ВКонтакте. Учитывая характер задачи, её постановка предполагает последовательную реализацию нескольких ключевых этапов: сбор и предобработка данных, построение графовой модели социальной сети, применение алгоритма кластеризации и представление результатов в удобной форме.

Основной целью разрабатываемой системы является автоматизация процесса анализа структуры социальной сети путём выделения кластеров пользователей с учётом плотности их взаимодействий. Результатом работы системы должно быть разбиение пользователей на сообщества, интерпретируемые как группы с высокой степенью внутрекластерной связанности. Такая информация может быть использована в задачах сегментации аудитории, анализа поведения пользователей, выявления ключевых центров влияния и построения моделей информационного распространения.

На основании этой цели формулируются следующие функциональные требования к системе:

1. Сбор данных: система должна обеспечивать автоматический доступ к данным ВКонтакте через официальное API, включая информацию о друзьях, подписках, участии в сообществах и другой активности, отражающей связи между пользователями.

2. Предварительная обработка данных: необходимо реализовать механизмы фильтрации, нормализации и очистки данных для формирования корректной графовой модели. Система должна обеспечивать устойчивую работу с неполными или разреженными данными, а также исключать дублирование информации.

3. Построение графа: данные, полученные на предыдущем этапе, должны быть трансформированы в графовую структуру, где вершины соответствуют пользователям, а рёбра — их связям. Возможность задания параметров формирования рёбер (например, по типу связи или порогу частоты взаимодействий) должна быть предусмотрена.

4. Кластеризация: в системе должен быть реализован модуль кластеризации, использующий бикластеризацию. Алгоритм должен обеспечивать возможность конфигурации параметров, а также поддерживать повторные запуски с целью повышения точности.

5. Хранение данных: необходимо спроектировать и реализовать систему хранения исходных данных, промежуточных графовых структур и результатов

кластеризации. Предполагается использование реляционной или документной базы данных с возможностью восстановления и повторной обработки данных.

6. Визуализация результатов: предусмотрен пользовательский интерфейс, позволяющий просматривать структуру полученных сообществ, метаинформацию о кластерах, а также статистику по численности, плотности связей и другим параметрам. Визуализация должна быть реализована в форме интерактивного графа или таблицы с фильтрами.
7. Повторяемость анализа: важной функцией системы является возможность повторного запуска анализа с сохранёнными параметрами, а также ведение журнала предыдущих запусков для сравнения и анализа изменений.

Разработка должна вестись с учётом модульности архитектуры и возможности последующего расширения функционала. Помимо основных требований, к системе предъявляются негласные критерии качества, такие как отказоустойчивость при работе с API, производительность при обработке большого объёма данных и понятный пользовательский интерфейс для взаимодействия с результатами анализа.

Помимо функциональных возможностей, система должна соответствовать ряду нефункциональных требований, обеспечивающих её надёжность, эффективность и пригодность для практического применения. Нефункциональные требования формулируются таким образом, чтобы их можно было верифицировать количественно либо качественно, что позволяет объективно оценить качество реализации программного продукта.

Во-первых, система должна удовлетворять требованиям по производительности. Время сбора и предварительной обработки данных для одного запроса пользователя не должно превышать 5 минут при использовании стабильного соединения с API ВКонтакте. Это обеспечит приемлемое время отклика системы даже при анализе крупных фрагментов социальной сети.

Во-вторых, к числу критически важных требований относится отказоустойчивость и стабильность. При сбоях в работе API система должна автоматически повторять запрос с задержкой или корректно завершать операцию с

сохранением логов ошибки. Уровень отказоустойчивости оценивается как не менее 90% успешных операций при нестабильной сети или ограничениях API.

В-третьих, система должна быть масштабируемой. Архитектура решения должна позволять обрабатывать увеличивающийся объём данных без полной переработки кода или архитектуры.

Также значимым параметром является удобство взаимодействия с результатами анализа. Пользовательский интерфейс должен обеспечивать доступ ко всем этапам анализа (выбор параметров, запуск кластеризации, просмотр результатов) и быть интуитивно понятным без необходимости обращения к инструкции. Для оценки используется метрика времени на освоение интерфейса — не более 10 минут для нового пользователя с базовым уровнем цифровой грамотности.

Кроме того, система должна обеспечивать корректную работу с частично неполными данными. Если отсутствует информация о части связей или пользователей, общее снижение точности кластеризации по метрике модульности не должно превышать 10% по сравнению с анализом полной выборки.

Наконец, требования к безопасности данных предполагают, что система не сохраняет персональные данные пользователей ВКонтакте, не передаёт их третьим лицам и использует токены доступа только в рамках действующего API. Все данные должны быть обезличены, а система — соответствовать требованиям информационной этики и политики конфиденциальности.

Соблюдение указанных нефункциональных требований позволит обеспечить надёжность и применимость разработанного программного решения в исследовательской и прикладной деятельности, что делает их неотъемлемой частью проектирования системы.

2.2 Архитектура и компоненты системы

Разработка программной системы кластеризации сообществ в социальной сети требует построения продуманной архитектуры, обеспечивающей логическую целостность, расширяемость и надёжность всех компонентов. Архитектурные решения должны соответствовать как функциональным, так и

нефункциональным требованиям, описанным ранее, а также учитывать этапность обработки данных: от получения информации из внешнего источника до визуализации результатов пользователю.

Система проектируется по модульному принципу, где каждый компонент отвечает за конкретную задачу и может быть независимо модифицирован или расширен. Такая архитектура способствует удобству сопровождения, повторному использованию кода и обеспечивает гибкость при внедрении новых алгоритмов или подключении альтернативных источников данных. Структурно система состоит из нескольких логически взаимосвязанных уровней: уровня сбора данных, уровня обработки, аналитического уровня и уровня представления результатов.

На первом уровне реализуется модуль интеграции с API социальной сети ВКонтакте. Он обеспечивает автоматизированный сбор данных о пользователях, их связях, участии в сообществах, а также другой доступной активности, отражающей структуру взаимодействий. Данный модуль реализует механизмы авторизации, управления токенами, ограничения частоты запросов и кэширования повторяющейся информации. Это позволяет обеспечить стабильную и отказоустойчивую работу системы при взаимодействии с внешним API, в том числе при возникновении ограничений или временных сбоев.

Следующий уровень отвечает за предварительную обработку полученных данных. Здесь производится фильтрация, очистка и нормализация информации, что необходимо для устранения дублирующихся и нерелевантных записей, а также для приведения данных к единому формату. Особое внимание уделяется текстовой информации, включающей названия, описания и публикации сообществ: она обрабатывается с использованием методов векторизации, таких как TF-IDF, и масштабируется с целью дальнейшего использования в модели. На данном этапе формируется матрица признаков, объединяющая семантическую и числовую информацию, которая используется для проведения дальнейшего анализа.

На аналитическом уровне располагается модуль кластеризации, реализующий метод спектральной бикластеризации. В отличие от классических подходов, данный метод позволяет выявлять не просто однородные группы объектов, а подмножества строк и столбцов, демонстрирующие схожее поведение. В контексте социальной

сети ВКонтакте это означает, что одновременно анализируются как группы, так и признаки, характеризующие их (например, тематика или численность аудитории), что даёт более точную и интерпретируемую сегментацию. В процессе кластеризации учитываются семантические взаимосвязи между сообществами, а также численные параметры, влияющие на структуру данных. Результатом работы модуля является множество бикластеров — локальных кластеров сообществ, объединённых по смысловому и структурному сходству.

Для хранения и управления данными используется отдельный компонент, реализующий взаимодействие с системой управления базами данных. Он отвечает за сохранение как исходных данных, так и промежуточных результатов — графовых структур, параметров запусков и итогов кластеризации. Это позволяет не только восстанавливать состояние системы, но и проводить сравнительный анализ между различными итерациями алгоритма или исследуемыми выборками.

Финальный уровень архитектуры — модуль визуализации и пользовательского взаимодействия. Он предоставляет доступ к результатам анализа в удобной и наглядной форме. Пользователь может просматривать общую структуру графа, выделенные кластеры, метаинформацию по каждому сообществу, а также статистику по количеству участников, плотности связей и другим метрикам. Предусмотрен также интерфейс для запуска анализа, настройки параметров и управления результатами.

Взаимодействие между компонентами организовано с учётом принципов слабой связности и чёткой ответственности. Это обеспечивает устойчивость системы к ошибкам отдельных модулей и возможность их независимого тестирования и замены. Такая архитектура позволяет системе масштабироваться, адаптироваться к новым условиям и быть устойчивой к изменениям структуры данных или требований к алгоритмам.

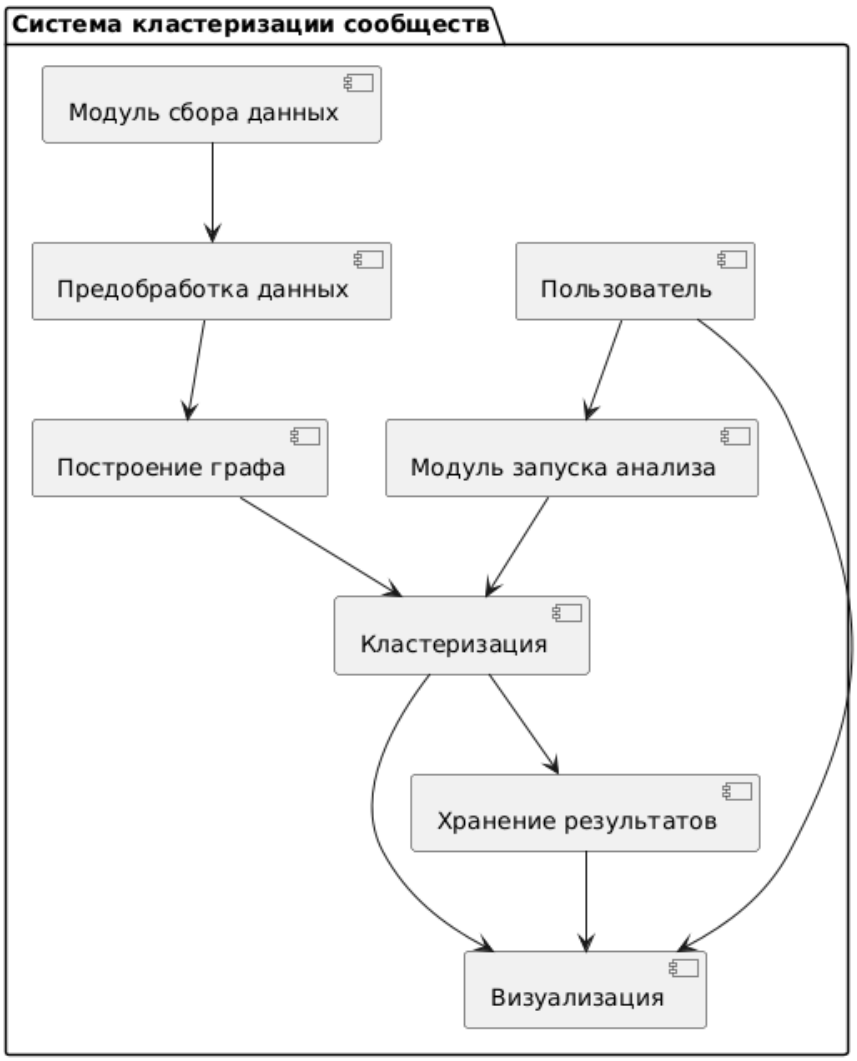

Рис. 1 Диаграмма компонентов системы

Диаграмма компонентов системы, представленная выше, визуализирует архитектуру решения для кластеризации сообществ в социальной сети ВКонтакте. Она демонстрирует основные модули, их взаимодействие и логическую последовательность обработки данных от начального этапа до выдачи результата пользователю.

Диаграмма последовательности (рис. 2) иллюстрирует временной порядок взаимодействий между пользователем и модулями системы при выполнении анализа. Пользователь инициирует процесс через интерфейс, после чего модуль управления направляет запрос на сбор данных, их предобработку, построение графа и последующую кластеризацию. Результаты сохраняются в базе данных и отображаются пользователю. Диаграмма подчёркивает этапность и логику обмена сообщениями между компонентами.

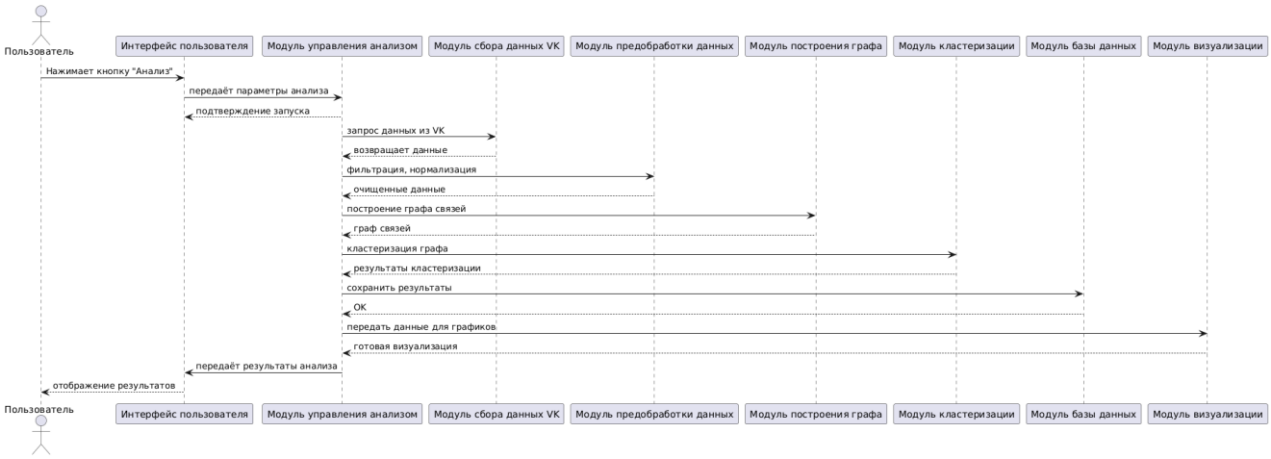

Рис. 2 Диаграмма последовательности

Диаграмма прецедентов (рис. 3) описывает основные действия, доступные пользователю. Система предоставляет функции запуска анализа, настройки параметров кластеризации, просмотра и сохранения результатов, а также повторного запуска процесса с другими параметрами. Эта диаграмма показывает, что пользователь является активным участником анализа и может гибко управлять процессом в интерактивном режиме.

Рис 3. Диаграмма прецендентов

Диаграмма архитектуры (рис. 4) отражает физическую структуру системы. Она состоит из пользовательского устройства, серверной части и базы данных. На клиентской стороне расположен пользовательский интерфейс, взаимодействующий с сервером, где размещены модули анализа. База данных служит для хранения как исходных, так и обработанных данных. Такое разделение обеспечивает масштабируемость, отказоустойчивость и безопасность при работе с данными социальной сети.

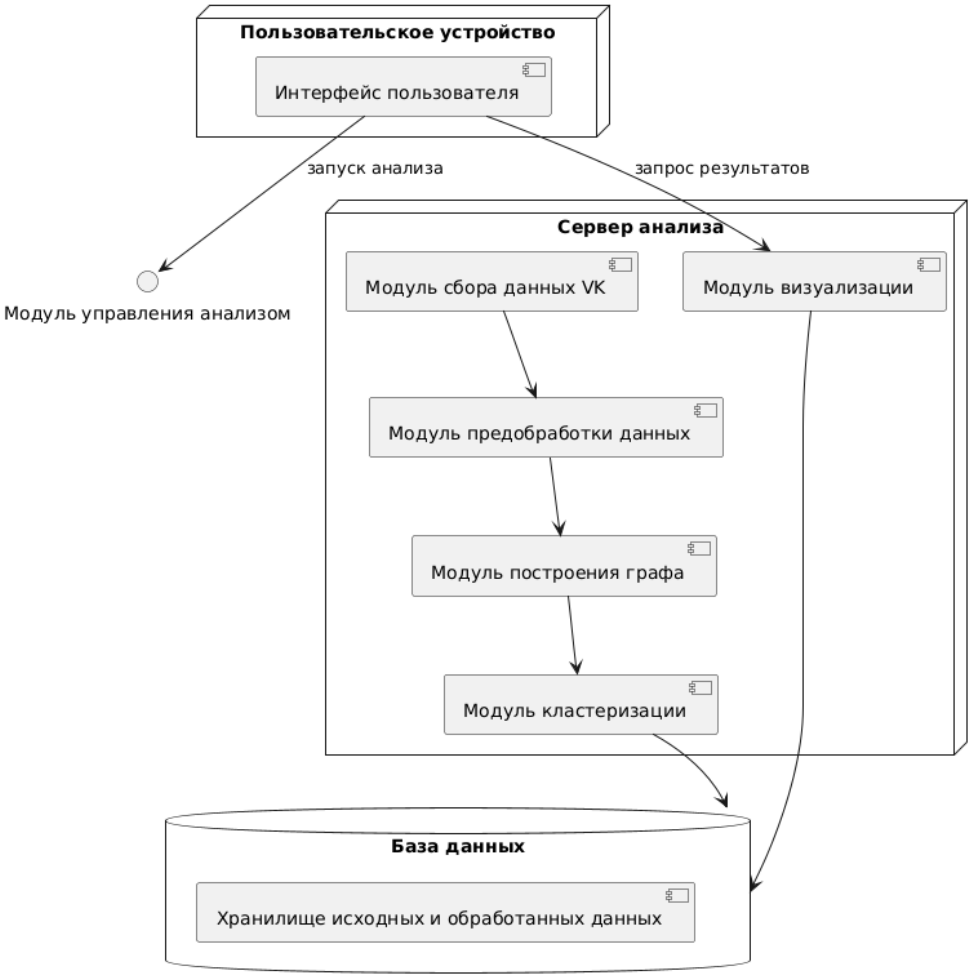

Рис. 4 Диаграмма архитектуры

В центре архитектуры находится аналитический модуль кластеризации, которому предшествуют этапы сбора и предварительной обработки данных, а также построение графовой модели. Визуализация результатов и управление анализом обеспечиваются отдельным пользовательским уровнем, взаимодействующим как с системой визуализации, так и напрямую с модулем запуска анализа. Результаты сохраняются в хранилище, откуда они могут быть как повторно загружены, так и использованы для последующего анализа.

Такое архитектурное решение позволяет достичь высокой модульности и изоляции логики каждого этапа. Благодаря чётко определённым границам между компонентами система может быть масштабирована и модифицирована: например, можно заменить алгоритм кластеризации без изменения модулей визуализации или сбора данных. Кроме того, наличие хранилища результатов обеспечивает возможность исторического анализа и сравнений между различными итерациями.

2.3 Выбор технологий и обоснование инструментов разработки

Разработка программной системы для кластеризации сообществ в социальной сети ВКонтакте требует подбора инструментов и технологий, которые обеспечат надёжную реализацию поставленных задач, соответствие функциональным и нефункциональным требованиям, а также возможность масштабирования и модификации проекта. При выборе средств разработки учитывались такие параметры, как доступность, поддержка необходимого функционала, наличие открытых библиотек, документации и активного сообщества разработчиков.

В качестве основного языка программирования выбран Python. Это решение обусловлено его широким применением в области анализа данных, машинного обучения и сетевой аналитики, а также наличием развитой экосистемы библиотек. Python позволяет эффективно обрабатывать структурированные и неструктурированные данные, легко интегрируется с API сторонних сервисов и поддерживает реализацию алгоритмов любой сложности, включая метды бикластеризации.

Для построения и обработки графовых структур используется библиотека NetworkX, являющаяся стандартом де-факто для работы с графами в Python. Она предоставляет все необходимые инструменты для создания, визуализации и анализа графов, а также для реализации собственных алгоритмов кластеризации. Благодаря NetworkX можно быстро прототипировать и отлаживать логику разбиения графа, отслеживать метрики (такие как модульность, плотность, центральность) и визуализировать полученные результаты.

Сбор данных из социальной сети ВКонтакте осуществляется с использованием официального API и библиотеки vk_api, которая предоставляет удобный интерфейс для авторизации, отправки запросов и обработки ответов. Библиотека поддерживает работу с методами получения информации о друзьях, подписках, участии в сообществах и другой активности пользователей. Это позволяет гибко формировать входной граф, отражающий реальные связи между участниками сети.

Для хранения данных на этапе реализации предусмотрено использование SQLite как лёгкой встраиваемой системы управления базами данных. SQLite позволяет быстро сохранять промежуточные результаты, настройки алгоритмов и метаинформацию, не требуя развёртывания отдельного сервера. При необходимости масштабирования система может быть адаптирована под использование полнофункциональной СУБД, такой как PostgreSQL.

Реализация биоинспирированных алгоритмов (например, генетических, муравьиных или алгоритма роя частиц) ведётся на основе библиотеки DEAP (Distributed Evolutionary Algorithms in Python), а также с использованием собственных реализаций, адаптированных под задачи кластеризации графов. DEAP предоставляет средства для определения структуры популяций, операторов мутации, селекции и кроссовера, что облегчает реализацию алгоритмов, имитирующих эволюционные процессы.

Для визуализации результатов кластеризации используются интерактивные графические библиотеки Plotly и PyVis. Они позволяют строить масштабируемые графы с возможностью навигации, отображения атрибутов узлов и рёбер, выделения сообществ цветом и фильтрации данных. Это особенно важно для пользовательского анализа, так как облегчает интерпретацию полученных кластеров и понимание структуры социальной сети.

Организация пользовательского взаимодействия с системой может быть выполнена с помощью простого графического интерфейса на базе Tkinter или через веб-интерфейс с использованием Flask, что обеспечит доступ к функционалу через браузер. Такой подход позволит в дальнейшем внедрить систему как самостоятельный аналитический инструмент.

Таким образом, выбранный стек технологий охватывает все ключевые аспекты: от сбора и хранения данных до анализа и визуального представления результатов. Он обеспечивает надёжность, гибкость, возможность масштабирования и совместимость с современными стандартами разработки программных решений в сфере анализа социальных сетей.

2.4 Разработка программного модуля

Реализация программного решения для анализа сообществ в социальной сети ВКонтакте представляет собой многоэтапный процесс, в котором каждый этап имеет важное значение для достижения общей цели исследования. Для успешного выявления скрытых структур внутри социальной сети необходимо обеспечить системный подход к подготовке данных, включающий их поиск, очистку, предварительный анализ и организацию в удобный для обработки формат. Каждый этап предварительной обработки критически важен для повышения качества и достоверности последующих аналитических процедур.

На первоначальной стадии работы осуществляется получение данных о сообществах, их содержании и активности пользователей. Сбор информации производится с использованием официального API социальной сети, что позволяет получать актуальные и структурированные данные. Однако особенности социальных сетей, такие как неполнота информации, различия в оформлении контента и высокая динамичность, требуют внимательной фильтрации и нормализации собранных данных, чтобы исключить искажения в итоговом анализе.

После этапа сбора и очистки подготовленные данные подвергаются первичному анализу и структурированию. Результатом этого процесса становится создание формализованного представления сетевой структуры, которое ляжет в основу последующего построения графа связей между пользователями. Именно на этой базе далее будет осуществляться кластеризация, направленная на выявление тематически или социально связанных групп пользователей, что является одной из ключевых задач исследования.

2.4.1 Поиск и сбор данных о группах ВКонтакте

Одним из первых этапов реализации программного решения является автоматизированный сбор данных о сообществах социальной сети ВКонтакте. Основная цель данного этапа — формирование исходного набора групп, релевантных тематике исследования, который в дальнейшем будет использоваться для анализа структуры, поведения участников и оценки параметров кластеризации.

Для поиска и сбора информации был разработан специализированный скрипт на языке программирования Python с использованием библиотеки Selenium. Выбор Selenium обусловлен тем, что данная библиотека позволяет имитировать действия пользователя в браузере, обходя ограничения стандартного API ВКонтакте, и получать доступ к данным, представленным в интерфейсе социальной сети. Это особенно важно в условиях, когда часть информации о группах недоступна через официальные API-интерфейсы.

Процесс поиска данных реализован следующим образом. Сначала с помощью Selenium осуществляется программное открытие страницы поиска сообществ в ВКонтакте по введённому пользователем запросу. Для стабильной и эффективной работы браузера используются настройки безголового режима (headless mode), что позволяет выполнять операции без открытия графического интерфейса и ускоряет процесс парсинга.

```
chrome_options = Options()
chrome_options.add_argument("--headless")
chrome_options.add_argument("--disable-gpu")
chrome_options.add_argument("--no-sandbox")

driver = webdriver.Chrome(service=Service(), options=chrome_options)
```

После загрузки страницы выполняется автоматическая прокрутка вниз для динамической подгрузки всех результатов поиска. Прокрутка осуществляется с помощью команды JavaScript window.scrollTo, выполняемой внутри браузерной сессии. Для избежания преждевременного завершения скроллинга реализована проверка изменения высоты страницы между итерациями.

```
def scroll_and_collect(url, max_scrolls=40):
    driver.get(url)
    sleep(3)

    last_height = driver.execute_script("return document.body.scrollHeight")
    for _ in range(max_scrolls):
        driver.execute_script("window.scrollTo(0, document.body.scrollHeight);")
        sleep(2)
        new_height = driver.execute_script("return document.body.scrollHeight")
        if new_height == last_height:
            break
        last_height = new_height

    items = driver.find_elements(By.CLASS_NAME, "groups_row")
    print(f"Найдено групп: {len(items)}")
    return items
```

Как только страница загружена полностью, осуществляется извлечение данных о каждой найденной группе. Из каждой карточки сообщества парсятся следующие атрибуты:

1. Название группы;
2. Ссылка на группу;
3. Описание сообщества;
4. Количество участников.

Для повышения устойчивости работы предусмотрена обработка исключений, позволяющая игнорировать неполные или повреждённые элементы и продолжать сбор данных без прерывания выполнения программы.

Все собранные данные последовательно сохраняются в промежуточную таблицу формата CSV, что обеспечивает удобство последующего анализа и возможность ручной проверки полученной информации. Для сохранения используется стандартный модуль csv с предварительным созданием заголовков таблицы.

Пользовательский сценарий работы с программой организован через интерфейс, где пользователь последовательно вводит поисковые запросы. Для каждого запроса происходит отдельный сбор данных. По завершении всех запросов результаты суммируются и сохраняются в единый CSV-файл.

Таким образом, реализованная система сбора данных обеспечивает следующие ключевые преимущества:

1. Возможность гибкого поиска групп по различным тематикам и регионам;
2. Автоматизацию процесса сбора информации без необходимости ручного копирования данных;
3. Формирование структурированной базы для последующей очистки, обработки и анализа;
4. Устойчивость к ошибкам и нестабильности загрузки страницы.

Данный подход позволил собрать первичный массив данных о группах ВКонтакте, необходимых для проведения дальнейшего исследования структуры

сообществ, их тематической близости, активности участников и эмоционального фона публикуемого контента.

2.4.2 Очистка и нормализация данных

После выполнения первичного поиска и сбора данных о сообществах социальной сети ВКонтакте возникает необходимость в проведении этапа очистки и нормализации информации. Данный процесс направлен на повышение качества исходной выборки, устранение ошибок и аномалий в данных, а также подготовку их к дальнейшему анализу.

Очистка данных начинается с устранения дубликатов. Поскольку в процессе поиска возможно повторное попадание одной и той же группы по разным поисковым запросам, в таблице могут содержаться несколько записей с одинаковыми идентификаторами. Для удаления таких дублей используется функция drop_duplicates библиотеки pandas, с указанием столбца "ID" в качестве критерия уникальности. На данном этапе также производится фиксация базовой статистики: общее количество записей до очистки и количество оставшихся записей после удаления дубликатов. Это позволяет количественно оценить степень загрязнения исходной выборки.

После удаления повторов проводится предварительный анализ содержимого столбца "Описание". Для каждой уникальной записи подсчитывается частота её появления в датасете. Такая проверка позволяет выявить группы с отсутствующими или слишком короткими описаниями, что в дальнейшем может повлиять на полноту текстового анализа при построении тепловой карты схожести между группами.

Следующим важным этапом является нормализация числовых данных о количестве участников в группах. Поскольку поле "Участники" при парсинге сохраняется в текстовом формате и может содержать пробелы, символы неразрывного пробела (\xa0) или другие артефакты форматирования, реализована функция для приведения значений к чистому числовому типу. Функция parse_members удаляет все нецифровые символы и преобразует строку в целочисленный формат (int). Это обеспечивает корректность сортировки и последующей аналитики.

```python
def parse_members(x):
    try:
        x = str(x).replace("\xa0", "").replace(" ", "")
        x = ''.join(filter(str.isdigit, x))
        return int(x) if x else 0
    except:
        return 0

df_cleaned["Участники"] = df_cleaned["Участники"].apply(parse_members)
```

Приведённые к единому формату данные о численности участников используются для формирования упорядоченного списка групп. Записи сортируются по убыванию количества участников, что позволяет выделить наиболее популярные сообщества. На текущем этапе выбраны 20 крупнейших групп, однако в дальнейшем реализация будет доработана таким образом, чтобы пользователь мог самостоятельно задавать желаемое количество групп для анализа.

После проведения очистки и нормализации очищенная таблица сохраняется в отдельный файл, что позволяет использовать полученные данные на следующих этапах обработки. Помимо этого проводится дополнительный анализ уникальности описаний в отобранных сообществах для оценки их разнообразия и подготовки к дальнейшему текстовому анализу.

Таким образом, этап очистки и нормализации данных включает:

1. удаление дубликатов по уникальному идентификатору группы;
2. приведение числовых полей к корректному формату;
3. фильтрацию и сортировку по количеству участников;
4. сохранение промежуточных результатов для удобства последующей обработки.

Корректная подготовка данных на этом этапе обеспечивает высокое качество входных данных, снижает вероятность возникновения ошибок на следующих шагах анализа и способствует более точному выявлению закономерностей внутри сетевой структуры.

2.4.3 Построение тепловой карты схожести групп

После этапа нормализации данных следующим шагом реализации программного решения является построение тепловой карты схожести между группами социальной сети ВКонтакте на основе анализа текстового контента. Данный этап направлен на выявление тематических кластеров сообществ путём комплексного изучения их названий, описаний и содержимого публикаций.

Исходной базой для работы служит очищенная таблица, содержащая информацию о выбранных группах. Для повышения качества анализа к ранее собранным данным добавляется текстовый массив публикаций, размещённых в каждой группе. Для этого реализована функция get_group_posts, использующая методы API ВКонтакте (utils.resolveScreenName и wall.get) для получения первых 30 постов каждой группы. Полученные тексты объединяются с названием и описанием группы, формируя единый текстовый корпус для дальнейшей обработки.

На этапе очистки текста производится нормализация данных: приведение текста к нижнему регистру, удаление символов, не относящихся к буквам русского или английского алфавита, а также фильтрация стоп-слов и удаление коротких слов длиной менее трёх символов. Для этого используется регулярная очистка и индивидуально подобранный список стоп-слов, наиболее типичных для русскоязычных текстов социальной сети.

```
def clean_text(text):
    text = text.lower()
    text = re.sub(r"[^а-яёa-z\s]", " ", text)
    tokens = text.split()
    tokens = [word for word in tokens if word not in custom_stop_words and len(word) > 2]
    return " ".join(tokens)
```

Преобразование текстов в векторное пространство осуществляется с помощью алгоритма TF-IDF (Term Frequency-Inverse Document Frequency), который позволяет учитывать как частоту появления слов в отдельных текстах, так и их распространённость по всему корпусу. Размерность пространства ограничена 1000 наиболее информативными признаками для сохранения качества модели при разумной вычислительной нагрузке.

Параллельно проводится нормализация численного признака — количества участников групп. Для этого используется метод масштабирования Min-Max, позволяющий привести данные к единому масштабу, пригодному для дальнейшего объединения с текстовыми признаками.

Финальное векторное пространство создаётся путём объединения текстовой матрицы признаков и нормализованных количественных данных в единый разреженный формат. Это позволяет одновременно учитывать тематическую направленность контента и популярность сообществ при построении модели.

На основе объединённого пространства признаков выполняется бикластеризация с использованием алгоритма Spectral Co-Clustering, который позволяет одновременно выявлять связанные подмножества строк и столбцов в матрице. Количество кластеров выбрано равным пяти, что обеспечивает баланс между детализацией и обобщением тематической структуры.

```
model = SpectralCoclustering(n_clusters=5, random_state=42)
model.fit(X)
df["Кластер"] = model.row_labels_
```

Результаты кластеризации закрепляются за каждой группой в виде метки кластера, добавленной в итоговую таблицу. Далее рассчитывается матрица косинусных сходств между текстовыми векторами групп, что позволяет наглядно отразить степень тематического пересечения между ними.

Визуализация результатов реализована в виде тепловой карты с использованием библиотеки seaborn. На графике пары групп с высокой степенью сходства выделяются насыщенными цветами, тогда как группы с минимальным совпадением остаются малозаметными. Подписи групп автоматически обрезаются при превышении 30 символов для повышения читаемости графика.

Итоговая тепловая карта сохраняется в файл vk_biclustering_labeled.png, а результирующая таблица с указанием кластера для каждой группы экспортируется в vk_clustered.csv.

2.4.4 Выявление наиболее влиятельных пользователей

Важным этапом исследования цифровых сообществ является выявление пользователей, оказывающих наибольшее влияние на коммуникационные процессы внутри социальной сети. В рамках данного проекта под влиянием понимается активность участников, выражающаяся в количестве оставленных комментариев и полученных реакций на них (лайков). Выявление таких пользователей позволяет анализировать структуру сетевого взаимодействия, а также строить гипотезы о распространении информации в сообществе.

Для реализации задачи была разработана программа на языке Python с использованием библиотеки requests для взаимодействия с API ВКонтакте. Исходной базой для анализа служит таблица cleaned_top10.csv, содержащая перечень исследуемых групп.

Первоначально для каждой группы производится получение её числового идентификатора (numeric ID) через метод utils.resolveScreenName. Это необходимо, поскольку многие группы в социальной сети идентифицируются по коротким символьным именам (screen_name), тогда как для работы с методами API требуется числовой формат ID.

```
def get_numeric_group_id(screen_name):
    try:
        r = requests.get("https://api.vk.com/method/utils.resolveScreenName", params={
            "access_token": VK_TOKEN,
            "v": VK_VERSION,
            "screen_name": screen_name
        }).json()
        return r.get("response", {}).get("object_id")
    except Exception:
        return None
```

Далее для каждой группы осуществляется сбор публикаций при помощи метода wall.get. Из выборки до 100 последних постов извлекаются идентификаторы публикаций, для которых запрашиваются комментарии через метод wall.getComments. В процессе обработки каждого комментария фиксируются следующие показатели:

1. Идентификатор пользователя, оставившего комментарий;
2. Количество комментариев, оставленных пользователем;

3. Количество лайков, полученных его комментариями.

Для накопления статистики используется структура данных defaultdict, позволяющая учитывать показатели активности для каждого пользователя без необходимости предварительной инициализации.

С целью обеспечения корректности работы и снижения нагрузки на API между запросами к серверу установлена искусственная задержка времени (sleep(0.4)).

После завершения сбора данных для всех групп производится расчёт итогового индекса влияния (influence_score) для каждого пользователя по формуле:

$$Influence\_Score = 2 \times Количество\ комментариев + 1 \times Количество\ лайков$$

Где вес комментариев выше, чем вес лайков, что отражает предположение о большей значимости непосредственного участия в дискуссии по сравнению с пассивной поддержкой.

Формула расчёта позволяет учитывать как активность пользователей, так и отклик аудитории на их высказывания, что даёт более полную картину уровня влияния. Пользователи ранжируются по убыванию итогового балла, формируя рейтинг наиболее активных участников в рамках исследуемых групп.

Итоговый список влиятельных пользователей сохраняется в файл vk_top_comment_influencers.csv. В таблице содержится информация об идентификаторе пользователя, числе оставленных комментариев, количестве лайков на них и рассчитанном индексе влияния.

Проведённый анализ позволяет не только выявить лидеров мнений внутри исследуемых сообществ, но и предоставляет базу для дальнейшего исследования их роли в процессе распространения информации, моделирования сетевого взаимодействия и построения прогностических моделей эволюции сообществ.

2.4.5 Анализ тональности публикаций

Анализ тональности текстов в социальных сетях является важным инструментом для понимания эмоционального фона, преобладающих настроений и степени вовлечённости аудитории. В рамках данной работы проведён анализ

тональности публикаций и комментариев в сообществах социальной сети ВКонтакте с использованием методов автоматизированной обработки текстов.

Для выполнения задачи была разработана программа на языке Python, использующая библиотеки requests, pandas, numpy, scikit-learn, seaborn и wordcloud. Сбор данных осуществлялся через официальное API ВКонтакте. Для каждой выбранной группы были загружены тексты постов (до 500 публикаций) и комментариев (до 20 комментариев под каждым постом), опубликованных в рамках сообщества.

На этапе предобработки текстов использовалась очистка от лишних символов: тексты приводились к нижнему регистру, удалялись все знаки препинания, цифры и прочие нерелевантные элементы, оставляя только кириллические слова. Далее применялась токенизация — разбиение текстов на отдельные слова.

Анализ тональности осуществлялся на основе словарного подхода. Были сформированы два словаря: положительных и отрицательных слов, содержащие наиболее часто встречающиеся в русском языке выражения положительных и негативных эмоций. Каждое слово в тексте проверялось на наличие в данных словарях, и по количеству совпадений определялась общая эмоциональная окраска текста. В зависимости от преобладания положительных или отрицательных слов публикация классифицировалась как позитивная, негативная или нейтральная.

Собранные результаты аггрегировались в виде матрицы, в которой для каждой группы подсчитывалось количество публикаций и комментариев с позитивной, негативной и нейтральной тональностью. Далее данная матрица была подвергнута кластеризации с использованием алгоритма Spectral Co-Clustering для выявления групп со схожим распределением эмоциональных настроений.

Для визуализации результатов анализа была построена комплексная схема, включающая:

1. Тепловую карту распределения тональности по группам;
2. Круговую диаграмму общего распределения эмоций;
3. Гистограмму накопленной тональности по группам;
4. Облака наиболее часто встречающихся позитивных и негативных слов;

5. График соотношения позитивных и негативных сообщений в каждом сообществе.

Визуализация позволяет быстро выявить сообщества с преимущественно позитивным, негативным или нейтральным настроением, а также провести сравнительный анализ между ними. На круговой диаграмме отчётливо видна доля каждой категории эмоциональной окраски среди всех проанализированных текстов, а гистограмма демонстрирует распределение эмоций по каждой конкретной группе.

Для каждой группы дополнительно рассчитано соотношение позитивных сообщений к негативным, что позволяет определить эмоциональное доминирование в сообществе. Значения коэффициента были нормированы и визуализированы на отдельной диаграмме для более наглядного представления различий между группами.

Итоговые результаты анализа сохранены в следующие файлы:

1. sentiment_matrix.csv — матрица распределения тональности по группам;
2. positive_words.csv и negative_words.csv — частотные списки слов с положительной и отрицательной окраской;
3. analysis_results.png — комплексная визуализация анализа.

Таким образом, проведённый этап анализа тональности позволил получить ценную информацию о характере коммуникации в сообществах ВКонтакте, определить эмоциональный фон обсуждений и выявить скрытые закономерности в настроениях аудитории, что в дальнейшем может использоваться для углублённого анализа структуры сетевых взаимодействий и прогнозирования развития сообществ.

2.4.6 Конструирование базы данных

Для обеспечения хранения, структурирования и последующего анализа информации, полученной в ходе работы программной системы, была спроектирована реляционная база данных, отражающая логику взаимодействия между пользователями, проектами и результатами анализа.

В основу модели положены принципы нормализации данных и обеспечения логической целостности. В результате проектирования была построена структура из

9 связанных между собой таблиц (см. рис. 5), каждая из которых соответствует одному из аспектов анализа.

Основные сущности:

1. users – содержит информацию об авторизированных пользователях в системе (имя, фамилия, email и пароль)
2. project – содержит сведения о созданных проектах (название проекта, дата создания и связанный пользователь)
3. search_keywords – хранит в себе ключевые слова, которые использовались для поиска
4. vk_groups – хранит данные о найденных группах ВКонтакте (ID, название, описание и количество участников)
5. heatmap_results – содержит путь к изображению тепловой карты для каждого проекта
6. influencer_analysis – таблица для хранения активности влиятельных пользователей (ID, количество комментариев, лайков, итоговый индекс влияния)
7. influencer_results – путь к CSV-файлу со списком влиятельных пользователей
8. sentiment_analysis – содержит ссылки на CSV-файлы с позитивными/негативными словами, матрицей тональности и визуализациями
9. sentiment_analysis_results – сохраняет изображения и матрицы по итогам анализа тональности

Все таблицы содержат поле created_at, фиксирующее дату и время создания записи, что упрощает аудит и отслеживание истории проекта.

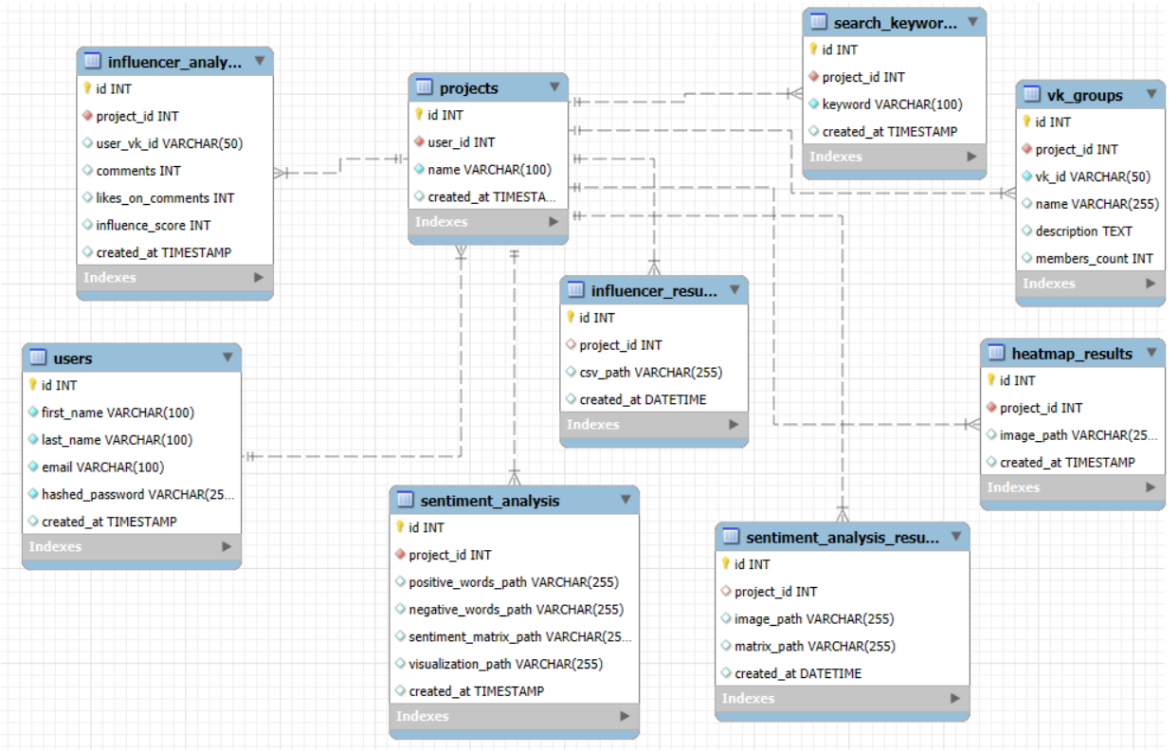

Рис. 5 – Диаграмма базы данных

База данных спроектирована таким образом, чтобы обеспечить эффективное и структурированное хранение всех ключевых типов информации, включая данные о пользователях, анализируемых группах и результатах обработки. Такая организация позволяет охватить весь цикл работы с проектом — от ввода ключевых слов до получения и визуализации итоговых результатов. Модель обладает гибкостью и легко поддаётся расширению, что создаёт возможности для последующего добавления новых сущностей. Благодаря чёткой структуре хранения значительно упрощается реализация поиска, фильтрации и восстановления ранее сохранённых данных, что положительно сказывается на удобстве использования системы через пользовательский интерфейс.

Глава 3 Тестирование и оценка результатов

На данном этапе работы проводится тестирование разработанной программной системы и оценка качества полученных результатов кластеризации сообществ в социальной сети ВКонтакте. Тестирование направлено на подтверждение работоспособности всех основных компонентов системы, а также на выявление соответствия реализованного решения функциональным и нефункциональным требованиям, определённым на этапе проектирования.

Для проверки эффективности системы использовались реальные данные о сообществах ВКонтакте, охватывающие различные тематики и размеры аудиторий. Такой подход позволил оценить универсальность разработанного решения и его устойчивость к изменениям структуры данных.

В процессе тестирования особое внимание уделялось следующим аспектам: корректности сбора и обработки данных, качеству построения графовой модели и выполнения кластеризации, точности выделения влиятельных пользователей, результатам анализа тональности публикаций, а также производительности работы системы при различных объёмах обрабатываемой информации.

Кроме функциональной проверки, проведён сравнительный анализ различных алгоритмов кластеризации для обоснования выбора наиболее подходящего метода. Завершающим этапом стало обсуждение интерпретации полученных результатов, выявление возможных ограничений текущего решения и формулирование направлений для его дальнейшего развития.

3.1 Методика тестирования программной среды

Тестирование программной системы направлено на проверку корректности реализации основных функций и оценку качества работы на практике. В рамках данного этапа разрабатывается методика, позволяющая всесторонне оценить работоспособность системы, выявить возможные ошибки и определить степень соответствия системы предъявляемым требованиям.

Тестирование проводилось на реальных данных социальной сети ВКонтакте, что позволяет обеспечить приближённость условий испытаний к реальным

сценариям использования. В качестве исходных данных использовались списки публичных групп, собранные с помощью разработанного модуля поиска и предварительной очистки информации. При выборе групп учитывалось разнообразие тематик и размеров аудиторий, что позволило оценить универсальность работы алгоритмов при различных характеристиках графов.

Методика тестирования включает проверку следующих аспектов:

1. корректность сбора данных из внешнего источника
2. полноту и точность предварительной обработки и нормализации данных
3. правильность построения графовой модели пользователей и связей
4. корректную работу алгоритмов кластеризации
5. точность выделения наиболее влиятельных пользователей по активности в сообществах
6. качество анализа тональности публикаций и комментариев
7. стабильность работы системы при изменении объёма данных;
8. производительность системы при увеличении размера обрабатываемой выборки.

Оценка корректности работы каждого модуля производилась с помощью ручной проверки выборочных данных на каждом этапе обработки, сопоставления результатов с ожидаемыми значениями и анализа метрик качества кластеризации. Для оценки качества кластеров использовались показатели модульности и однородности сообществ.

Особое внимание уделялось устойчивости системы к возможным ошибкам при взаимодействии с внешними API, таким как превышение лимитов запросов или недоступность сервера. В этих случаях проверялась способность системы корректно обрабатывать исключения и продолжать выполнение работы без критических сбоев.

Производительность системы оценивалась по времени выполнения основных операций: сбора данных, построения графа, проведения кластеризации и визуализации результатов. При этом проводилось тестирование на различных размерах выборок с целью анализа масштабируемости решения.

Таким образом, разработанная методика тестирования позволяет комплексно оценить качество программной реализации, выявить возможные узкие места и подтвердить соответствие разработанного решения предъявляемым требованиям.

3.2 Проведение кластеризации сообществ на данных ВКонтакте

Перед проведением кластеризации необходимо было сформировать тестовую выборку данных, максимально приближенную к реальным условиям функционирования социальной сети. В качестве источника данных была использована платформа ВКонтакте, предоставляющая доступ к информации о сообществах через открытый интерфейс поиска.

Для получения начального набора групп была разработана вспомогательная программа, основанная на использовании браузерной автоматизации через библиотеку Selenium. Пользователь на этапе подготовки данных имел возможность ввести ключевой запрос, на основе которого система автоматически осуществляла прокрутку страницы поиска и извлечение первых 240 результатов, соответствующих введённому запросу. Такой подход позволил быстро сформировать репрезентативную базу сообществ, тематически связанных с интересующей областью.

Алгоритм был построен таким образом, чтобы обеспечить возможность многократного выполнения поиска: после каждого завершённого запроса пользователю предоставлялась возможность ввести новый поисковый термин. Это позволило гибко наращивать объём выборки, охватывая различные тематики и направления сообществ. По завершении процесса сбора данные автоматически сохранялись в таблицу в формате CSV, содержащую основные характеристики групп: идентификатор сообщества, название, краткое описание и примерное количество участников.

Такой метод сбора данных обеспечил высокую вариативность и полноту информации, а также позволил оперативно сформировать объёмную и тематически разнообразную выборку для последующего этапа предварительной обработки и анализа.

После сбора данных следующим этапом стала предварительная обработка и нормализация информации, необходимая для корректной работы алгоритмов машинного обучения. На этом этапе производилась очистка данных от дублирующихся записей, которые могли возникнуть при многократных поисковых запросах или при наличии идентичных сообществ в результатах поиска. Первоначально для каждой группы проверялся уникальный идентификатор, и в случае обнаружения совпадений из выборки удалялись все дубли, оставляя только одну запись для каждого сообщества. Данный шаг позволил исключить искажения при построении графовой модели и анализе связей между группами.

Далее была произведена очистка числового признака, отражающего количество подписчиков в сообществе. Поскольку исходные данные могли содержать текстовые форматы отображения числа участников (например, с пробелами или символами), выполнялась приведение данных к единому числовому виду и нормализация для последующего масштабирования признаков.

Для удобства дальнейшего анализа и визуализации данных из общей выборки были отобраны двадцать сообществ с наибольшим количеством подписчиков. Такой подход позволил сосредоточиться на наиболее активных и представительных группах, обеспечив достаточную плотность взаимодействий и более наглядную структуру связей для последующего этапа кластеризации.

После подготовки данных следующим этапом стало построение тепловой карты схожести между выбранными сообществами. Для этого использовалась как текстовая информация о группах (названия, описания, публикации), так и числовые характеристики, например, количество подписчиков.

На основе собранного текста для каждой группы были сформированы векторные представления с использованием метода TF-IDF. Этот метод позволяет определить важность каждого слова для описания группы, при этом часто встречающиеся общеупотребительные слова были заранее удалены с помощью списка стоп-слов. Чтобы дополнительно учитывать размер аудитории, данные о количестве участников в группах были нормализованы и добавлены к текстовым признакам.

На объединённых признаках была проведена кластеризация с помощью алгоритма Spectral Co-Clustering. В результате работы алгоритма все сообщества были разбиты на пять кластеров, в которых группы оказались наиболее похожими друг на друга по содержанию своих публикаций и общей активности.

Для наглядного представления результатов была построена тепловая карта на основе матрицы косинусной схожести между текстами групп. На тепловой карте оси подписаны сокращёнными названиями сообществ, а цвет показывает степень их схожести: чем ярче цвет, тем более похожи между собой группы. Объекты на карте были отсортированы в соответствии с кластерами, что позволило легко увидеть компактные области, где сообщества оказались тематически близки.

На рисунке 6 представлена полученная тепловая карта кластеров сообществ:

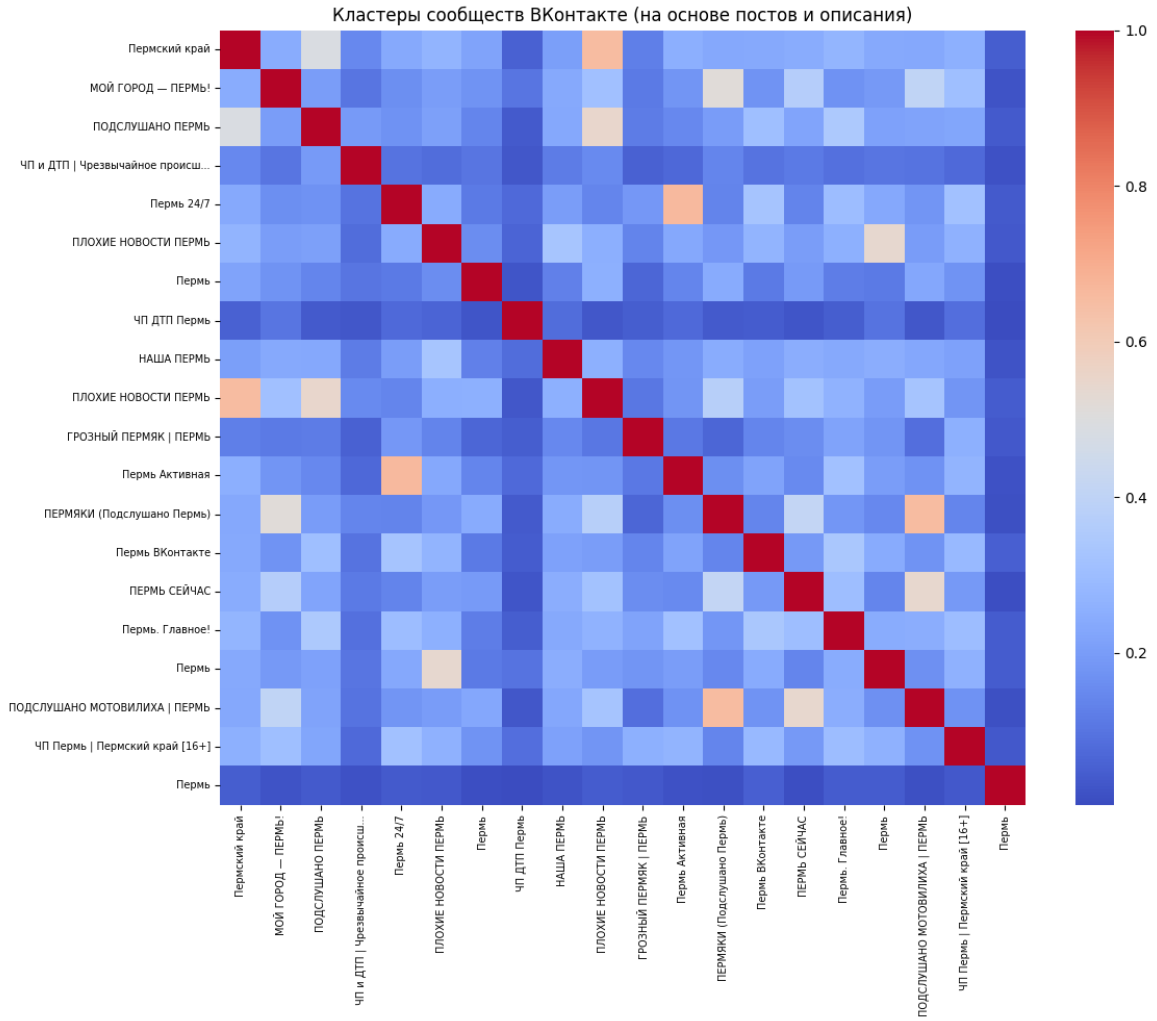

Рис. 6 — Тепловая карта кластеров сообществ ВКонтакте

Как видно из полученного результата, некоторые группы демонстрируют высокую степень текстовой и тематической близости, что проявляется в виде плотных цветных квадратов вдоль диагонали тепловой карты. Это подтверждает корректность работы алгоритма кластеризации и позволяет заключить, что разработанная методика успешно справляется с задачей выделения сообществ, объединённых общими интересами или тематикой.

Далее была проведена процедура выявления наиболее влиятельных пользователей на основе анализа активности в комментариях к публикациям сообществ. Основная идея заключалась в учёте как количества оставленных комментариев, так и отклика аудитории в виде лайков на эти комментарии.

На первом этапе для каждого сообщества был определён его числовой идентификатор для обращения к API ВКонтакте. Затем с использованием программного скрипта производился сбор публикаций, а также комментариев к ним. Для каждого комментария фиксировался автор и количество полученных лайков.

Для расчёта общей меры влияния пользователя применялась специальная метрика — influence score, которая определялась по формуле: сумма удвоенного количества оставленных комментариев и количества лайков на них. Такой подход позволял одновременно учитывать как саму активность пользователя, так и уровень отклика на его сообщения. При этом активность в комментариях и количество получаемых лайков не всегда оказывались прямо пропорциональными: в некоторых случаях пользователи оставляли небольшое количество комментариев, но получали высокую оценку аудитории, а иногда напротив — проявляли высокую активность, но собирали относительно мало лайков.

После обработки всех данных был сформирован рейтинг пользователей, упорядоченный по убыванию их influence score. Пользователи с наибольшими значениями этого показателя были определены как наиболее влиятельные в рамках выбранной выборки.

На рисунке 7 приведён пример полученной таблицы:

user_id	comments	likes_on_comments	influence_score
149024736	7	924	938
483993140	60	452	572
683675310	112	225	449
664115693	46	355	447
106039709	1	442	444
608340937	1	416	418
39656180	4	402	410
113580471	3	383	389
180527307	19	303	341
244102480	2	320	324

Рис. 7 — Список наиболее влиятельных пользователей по результатам анализа комментариев

Анализ таблицы показал, что высокое влияние могут иметь как пользователи с небольшой, но качественной активностью (получающие много лайков на отдельные комментарии), так и пользователи, активно вовлечённые в дискуссии, но набирающие умеренное количество откликов. Это разнообразие стратегий поведения отражает сложную природу вовлечённости участников в цифровом пространстве и подчёркивает важность комплексных метрик для оценки влияния в социальных сетях.

Также, одним из этапов анализа сообществ стало исследование тональности публикуемого контента. Целью данного этапа было определение преобладающего эмоционального окраса сообщений в разных группах социальной сети ВКонтакте, а также выявление общих закономерностей и различий между сообществами.

Для реализации анализа была разработана программа, которая автоматически собирала текстовые данные из выбранных групп: публикации на стене сообществ и комментарии пользователей. Для каждой группы обрабатывались до 500 постов и до 20 комментариев под каждым постом, что позволило получить достаточно репрезентативную выборку текстов.

Очищенные тексты подвергались обработке методом простого лексического анализа. Для определения эмоциональной окраски сообщений использовались специально составленные словари позитивной и негативной лексики. На основе подсчёта вхождений слов из этих списков определялась тональность каждого текста:

позитивная, негативная или нейтральная. В случае преобладания позитивной лексики сообщение относилось к категории «позитив», при доминировании негативной — к категории «негатив», а при их равенстве либо отсутствии ключевых слов — к категории «нейтраль».

На следующем этапе результаты по каждой группе агрегировались в виде матрицы распределения сообщений по категориям тональности. Для улучшения восприятия данных и выявления скрытых закономерностей была применена кластеризация с использованием метода спектральной бикластеризации. Это позволило сгруппировать схожие по эмоциональному профилю сообщества.

Визуализация результатов анализа была представлена в нескольких форматах. Построена тепловая карта, отражающая распределение позитивных, нейтральных и негативных сообщений по группам, что позволило наглядно оценить эмоциональный фон каждой из исследуемых групп. Также были сгенерированы круговые диаграммы, показывающие долю каждого типа сообщений, и гистограммы накопленного распределения тональностей.

Для более детального анализа текстов дополнительно были построены облака слов отдельно для позитивной и негативной лексики, на которых были выделены наиболее часто употребляемые слова. Это позволило получить представление о темах и настроениях, характерных для исследуемых сообществ.

Пример визуализации результатов анализа представлен на рисунке 8:

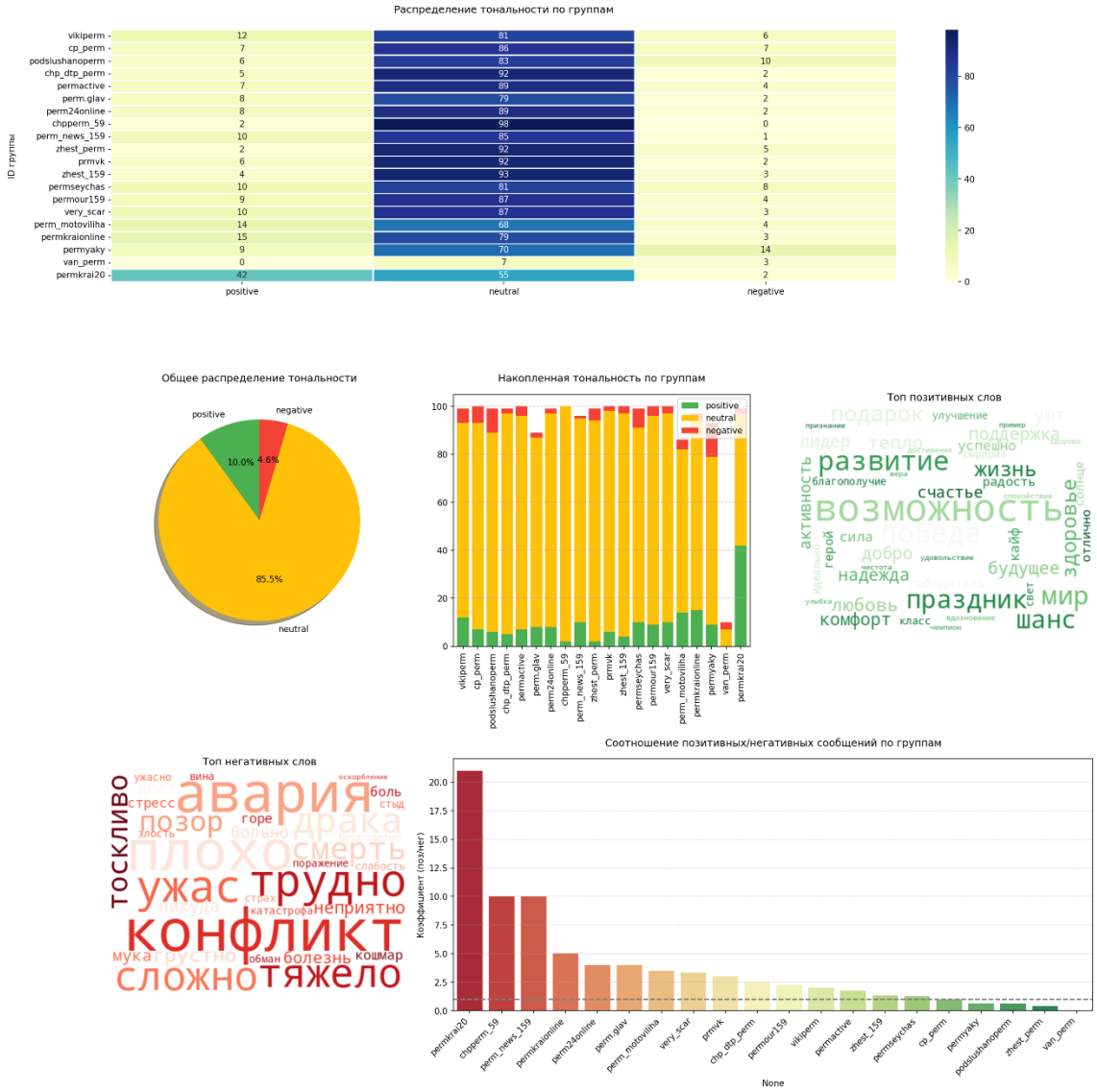

Рис. 8 — Комплексная визуализация результатов анализа тональности публикаций в сообществах

Анализ полученных данных показал, что эмоциональный фон в группах может существенно различаться. В некоторых сообществах преобладали позитивные настроения, связанные с продвижением культурных или образовательных инициатив, тогда как в других группах наблюдалось большое количество негативных сообщений, часто связанных с социальными проблемами или политическими дискуссиями.

3.3 Интерпретация результатов и анализ качества кластеров

Завершающим этапом практической части исследования стал анализ результатов, полученных в ходе работы программной системы. Были проведены три ключевых направления анализа: кластеризация сообществ, выявление влиятельных пользователей и исследование тональности публикаций. Каждый из этих этапов позволил сформировать более полную картину внутренней структуры и динамики онлайн-среды, представленной группами социальной сети «ВКонтакте».

В рамках кластеризации сообществ на основе текстовой информации — названий, описаний и постов — была построена тепловая карта, отображающая степень схожести между группами. Анализ данной карты показывает наличие локальных областей высокой плотности, где наблюдается сильная семантическая связь между сообществами. Это может свидетельствовать о тематической близости, общих интересах или совпадении лексических паттернов публикаций. Группы с высокой взаимной схожестью были объединены в кластеры с помощью бикластеризации. Благодаря этому подходу удалось структурировать выборку в логически осмысленные блоки, что особенно ценно при анализе большого массива неоднородных данных.

Следующим этапом стало выявление наиболее влиятельных пользователей на основе активности в комментариях. Влияние измерялось через комбинированный показатель, включающий число оставленных комментариев и суммарное количество лайков на них. Наибольший вклад во взаимодействие с сообществами внесли пользователи, которые не обязательно были наиболее активными по количеству сообщений. Например, один из пользователей с относительно небольшим числом комментариев (всего 7) получил на них 924 лайка, что позволило ему занять лидирующую позицию по совокупному «influence score». Это подтверждает, что не только частота участия, но и реакция аудитории играют важную роль в формировании пользовательского влияния. Подобная информация может быть применима в задачах таргетирования, выявления лидеров мнений и оценки потенциальных каналов распространения информации.

Финальной частью стал анализ тональности сообщений, охватывающий как посты, так и комментарии. Результаты представлены в виде набора визуализаций,

среди которых — круговая диаграмма общего распределения, гистограмма накопленной тональности по группам и облака слов для положительной и отрицательной лексики. Анализ показал, что преобладающее большинство сообщений носят нейтральный характер (85,5%), тогда как положительные составляют 10%, а негативные — лишь 4,6%. Такая картина характерна для большинства неангажированных сообществ, ориентированных на информирование или обсуждение повседневных тем. В отдельных группах наблюдаются заметные отклонения — например, повышенный уровень негативных сообщений, что может указывать на тематическую специфику (новости, происшествия) или высокий уровень эмоционального вовлечения аудитории.

Особый интерес представляет диаграмма соотношения позитивных и негативных сообщений по группам. Она выявляет, в каких сообществах наблюдается сильный перекос в ту или иную сторону. Это может служить индикатором общественных настроений, эмоционального фона обсуждаемых тем или качества модерации контента. Также в рамках анализа были выделены наиболее часто встречающиеся позитивные и негативные слова, что позволяет лучше понять характер дискуссий внутри сообществ.

Таким образом, проведённый анализ продемонстрировал практическую применимость разработанного программного решения для структурирования и оценки больших объёмов данных из социальных сетей. Полученные результаты могут быть использованы для дальнейших исследований, построения поведенческих моделей пользователей, а также в рамках прикладных задач, таких как мониторинг общественного мнения или выявление тематических кластеров.

3.4 Ограничения и возможные улучшения

Несмотря на успешную реализацию программной системы и получение интерпретируемых результатов, в ходе работы были выявлены определённые ограничения, которые стоит учитывать как при анализе данных, так и при дальнейшем развитии проекта.

Во-первых, сбор информации из социальной сети ВКонтакте осуществляется через публичный API, что накладывает жёсткие лимиты на количество запросов и

объём получаемых данных. Это особенно ощутимо при обработке больших групп с активным обсуждением, где число комментариев может исчисляться тысячами. В результате полнота данных может быть ограничена, что влияет на точность выявления наиболее влиятельных пользователей и качество тонального анализа. Для преодоления этого ограничения в будущем возможно использование авторизованных приложений с расширенными правами доступа или применение методов распределённого сбора данных с использованием нескольких токенов.

Во-вторых, алгоритм сбора информации из API ВКонтакте, включающий последовательную загрузку постов и комментариев, отличается высокой ресурсоёмкостью и продолжительностью выполнения. При анализе десятков сообществ общее время может достигать нескольких часов, особенно при учёте задержек между запросами во избежание блокировки. Это ограничивает оперативность анализа и требует дополнительной оптимизации, например, через параллельную обработку или предварительное кэширование полученных данных.

Третьим ограничением является то, что система анализа тональности основана на словарном подходе, что обеспечивает простоту и скорость обработки, но накладывает ограничения на гибкость и точность. Такой подход не учитывает контекст, ироничные конструкции, сарказм или двойные смыслы, что может привести к искажению оценки тональности. Решением данной проблемы может стать внедрение современных нейросетевых моделей, таких как BERT или RuRoBERTa, специально обученных для задач сентимент-анализа на русском языке.

Кроме того, при кластеризации сообществ используется алгоритм бикластеризации на основе текстовых признаков и численного показателя подписчиков. Хотя данный метод показал свою эффективность на ограниченной выборке, он может терять устойчивость при масштабировании на большие данные или при резком росте тематики сообществ. Возможным направлением развития является внедрение методов тематического моделирования (например, LDA) или комбинированных подходов, сочетающих топологические и семантические признаки для построения более устойчивых кластеров.

Также стоит отметить, что влияние пользователей в текущей реализации определяется только по активности в комментариях. Это упрощённая модель, не

учитывающая другие формы взаимодействия таких как репосты, создание собственных публикаций, влияние на других участников. Более точный подход к анализу вовлечённости может включать построение графа взаимодействий между пользователями и анализ центральности или PageRank-подобных метрик.

Таким образом, несмотря на достигнутые результаты, разработанная система обладает рядом точек роста, которые при их реализации могут значительно расширить область применения, повысить точность анализа и улучшить обоснованность выводов.

Список литературы

1. Statista. VKontakte users in Russia 2018–2028 [Электронный ресурс]. – Режим доступа: https://www.statista.com/forecasts/1144242/vkontakte-users-in-russia. – Дата обращения: 09.04.2025.
2. Datareportal. Digital 2024: Global Overview Report. Social Media Users [Электронный ресурс]. – Режим доступа: https://datareportal.com/social-media-users. – Дата обращения: 09.04.2025.
3. Забарная Э. Н., Куриленко И. В. Социальные сети: основные понятия, характеристики и современные исследования [Электронный ресурс] // CyberLeninka. – Режим доступа: https://cyberleninka.ru/article/n/sotsialnye-seti-osnovnye-ponyatiya-harakteristiki-i-sovremennye-issledovaniya/viewer. – Дата обращения: 09.04.2025.
4. Мясникова Е. И. Понятие «социальная сеть» в социологических теориях и интернет-практиках [Электронный ресурс] // CyberLeninka. – Режим доступа: https://cyberleninka.ru/article/n/ponyatie-sotsialnaya-set-v-sotsiologicheskih-teoriyah-i-internet-praktikah/viewer. – Дата обращения: 09.04.2025.
5. Ремизова И. М. Влияние социальных сетей на эмоциональные переживания пользователей: социологический аспект [Электронный ресурс] // CyberLeninka. – Режим доступа: https://cyberleninka.ru/article/n/vliyanie-sotsialnyh-setey-na-emotsionalnye-perezhivaniya-polzovateley-sotsiologicheskiy-aspekt/viewer. – Дата обращения: 09.04.2025.
6. Бронников М. А. Применение социальных сетей в целях управления маркетингом // Актуальные исследования. – 2024. – № 20 (202). – С. 44–48.
7. Элбон К. Машинное обучение с использованием Python. Сборник рецептов: практические решения от предобработки до глубокого обучения / пер. с англ. – М. : БХВ-Петербург, O'Reilly, 2019. – 369 с.
8. Бурков А. Машинное обучение без лишних слов. – СПб. : Питер, 2020. – 192 с.
9. Ullah F., Lee S. Community clustering based on trust modeling weighted by user interests in online social networks // Chaos, Solitons and Fractals. – 2017. – Vol. 103. – P. 184–204.

10. Elgazzar H., Spurlock K., Bogart T. Evolutionary clustering and community detection algorithms for social media health surveillance // Machine Learning with Applications. – 2021. – Vol. 6. – Article number: 100084.
11. Кацов И. Машинное обучение для бизнеса и маркетинга. – СПб. : Питер, 2019. – 224 с.
12. Нанцекин А. Н., Курчистый И. Ю. Комбинированный биоинспирированный алгоритм для решения задачи кластеризации данных // Современные информационные технологии и ИТ-образование. – 2018. – № 3(38). – С. 185–197.
13. Usanin A., Zimin I., Zamjatina E. Study of Strategies for Disseminating Information in Social Networks Using Simulation Tools // van der Aalst W.M.P. (ed.). AIST 2020. Lecture Notes in Computer Science, vol. 12602. – Cham : Springer, 2021.
14. Григорян А. С., Курейчик В. М., Курейчик В. В. Программный комплекс решения задач кластеризации // Проблемы управления и систем. – 2017. – № 2(89). – С. 261–269.
15. Марков В. В., Кравченко Ю. А., Кузьмина М. А. Развитие методов семантической фильтрации на основе решения задач кластеризации биоинспирированными алгоритмами // Раздел IV. Анализ данных и управление знаниями. – 2018. – С. 1–9.
16. Плаксин М. А. Тестирование и отладка программ для профессионалов будущих и настоящих. – 2-е изд. – М. :БИНОМ, Лаборатория знаний, 2013. – 167с.