МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение
высшего образования
**«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИМЕНИ Н.Г. ЧЕРНЫШЕВСКОГО»**

Кафедра социальной информатики

# ИССЛЕДОВАНИЕ МОДЕЛЕЙ КЛАСТЕРИЗАЦИИ СОЦИАЛЬНЫХ СЕТЕЙ

(автореферат бакалаврской работы)

студента 5 курса 531 группы
направления 09.03.03 - Прикладная информатика
профиль Прикладная информатика в социологии
Социологического факультета
Михайлова Данилы Андреевича

Научный руководитель
кандидат социологических наук, доцент _____ И.Г. Малинский
подпись, дата

Зав. кафедрой
кандидат социологических наук, доцент _____ И.Г. Малинский
подпись, дата

Саратов 2025

# ВВЕДЕНИЕ

**Актуальность проблемы.** Современное общество переживает стремительную цифровизацию, ключевым проявлением которой стало повсеместное распространение социальных сетей. Такие платформы, как «ВКонтакте», «Одноклассники», «Яндекс.Дзен», Telegram и другие, объединяют десятки миллионов пользователей, ежедневно генерируя огромные объемы разнородных данных. Общественные коммуникации, создание сообществ и обмен информацией все чаще происходят именно через социальные сети. Актуальность данной проблемы обусловлена необходимостью извлекать ценную информацию из этих больших массивов данных социальных медиа. Для этого требуются эффективные методы анализа, в частности методы кластеризации, позволяющие автоматически выявлять скрытые структуры и сообщества в соцсетевых данных. Кластерный анализ, будучи одной из базовых задач интеллектуальной обработки данных, приобретает все большее значение в сфере анализа социальных сетей, где традиционные подходы недостаточны для осмысления сложной структуры связей и контента.

**Степень научной разработанности.** Тематика кластеризации и анализа социальных сетей активно развивается на стыке социологии, информатики и прикладной математики. Широкий пласт научных исследований посвящен кластерному анализу данных в различных областях. В сетевой науке и социологии разрабатываются алгоритмы обнаружения сообществ в графах социальных связей (классические работы М. Гирвана и М. Ньюмана по разбиению сетей на сообщества, обзоры С. Фортунато и др.). В области машинного обучения и анализа данных сформирован обширный набор методов кластеризации: от классических алгоритмов (метод k-средних Маккуина, иерархические методы) до современных подходов к кластеризации графов (алгоритмы Louvain и Leiden, метод распространения меток Рагхавана и др.) и плотностных алгоритмов (DBSCAN Эстера с соавт., HDBSCAN Кампелло и др.). Различные исследователи улучшали эти методы, адаптируя

их к большим данным и специфике социальных сетей. Отечественные ученые также внесли вклад в изучение данной проблемы: например, Королёва К.В. с соавторами рассматривала кластеризацию пользователей социальных сетей по интересам, Шашкин М.А. изучал сегментацию новостного контента методом кластеризации и т.д. Таким образом, на сегодняшний день сформировалась солидная научная база, посвященная как развитию алгоритмов кластеризации, так и их применению для анализа социальных сетей.

**Цель исследования** заключается в комплексном анализе моделей кластерного анализа применительно к данным социальных сетей (с акцентом на российские платформы) и выявлении наиболее эффективных подходов для работы с соцсетевыми данными.

**Задачи исследования:**

1. Изучить теоретические основы кластеризации в социальных сетях: дать определения основным понятиям, рассмотреть свойства социальных сетей и характер генерируемых ими данных, обозначить ключевые задачи кластеризации и ее роль в аналитике социальных медиа.
2. Провести обзор наиболее распространенных моделей и алгоритмов кластеризации (k-средних, DBSCAN, HDBSCAN, спектральные методы, алгоритмы Louvain, Leiden, Label Propagation, ансамблевые подходы), проанализировать принципы их работы, преимущества и ограничения при применении к данным социальных сетей.
3. Выполнить экспериментальное исследование: реализовать выбранные алгоритмы на реальных данных, собранных из открытых источников российских социальных сетей, оценить качество полученной кластеризации, интерпретировать выявленные кластеры и разработать рекомендации по применению методов кластеризации в различных сферах (государственное управление, безопасность, социология, бизнес).

**Объект исследования** – социальные сети как источник больших массивов данных о пользователях, их связях и контенте.

**Предмет исследования** – модели и алгоритмы кластерного анализа, применяемые к данным социальных сетей, а также их эффективность и особенности в контексте соцсетевых данных.

**Методы исследования.** Для достижения поставленных целей использовались следующие методы: анализ научной литературы по тематике кластеризации и социального анализа сетей; сравнительный анализ и систематизация алгоритмов кластеризации; компьютерное моделирование (эксперимент) на реальных данных социальных сетей; методы визуализации и статистической оценки результатов кластеризации для интерпретации полученных групп.

**Эмпирическая база** выпускной квалификационной работы включает данные, собранные из открытых источников российских социальных платформ. В частности, для экспериментов использовались: выборка из ~10 000 пользователей сети «ВКонтакте» (со сведениями об их дружеских связях, базовых профилях и подписках на сообщества) и корпус из ~1 000 статей платформы «Яндекс.Дзен» (содержащих тексты публикаций, теги и статистику просмотров). Эти данные позволили протестировать различные алгоритмы кластеризации на практике и сопоставить их результаты.

**Теоретическая значимость** исследования заключается в обобщении и систематизации знаний о современных моделях кластеризации применительно к социальным сетям. Основные положения и выводы работы могут послужить основой для дальнейших научных исследований в области анализа социальных сетей и разработки новых методов кластеризации соцсетевых данных.

**Практическая значимость** работы определяется тем, что полученные результаты и рекомендации могут быть непосредственно использованы в прикладных сферах. Выявленные подходы к кластеризации социальных сетей могут применяться при мониторинге общественного мнения и поиске сообществ в государственном управлении, для обнаружения координированных групп или аномалий в обеспечении информационной

безопасности, для сегментации аудитории и персонализации контента в бизнесе и маркетинге, а также в социологических исследованиях онлайн-сообществ. Реализация алгоритмов, проведенная в рамках работы, и сделанные выводы могут быть внедрены в аналитические системы социальных медиаплатформ для улучшения рекомендаций и таргетирования.

Структура выпускной квалификационной работы представлена введением, тремя главами, заключением и списком использованных источников.

## ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

В первом разделе **«Теоретические основы кластеризации в социальных сетях»** рассматриваются базовые понятия и принципы кластерного анализа применительно к данным соцсетей. Дано определение кластеризации как метода автоматической группировки объектов по степени схожести без заранее заданных классов. Описаны ключевые свойства данных социальных сетей: графовая природа (пользователи и связи образуют социальный граф), огромный масштаб и разреженность таких графов, высокая кластеризация (склонность образовывать плотные сообщества), динамичность и гетерогенность информации (наличие разнообразных типов данных – связи, тексты, мультимедиа). Отмечено особое значение кластерного анализа для понимания сложной структуры социальных связей и поведения пользователей в онлайновых сообществах. Выделены основные задачи кластеризации социальных сетей – обнаружение сообществ пользователей по интересам или социальным связям, сегментация аудитории для маркетинговых или информационных целей, группировка контента по тематикам, выявление аномальных групп или аккаунтов. Показано, что кластеризация преобразует «сырые» большие данные соцсетей в осмысленные группы (кластеры), которые облегчают дальнейший анализ и принятие решений. Также подчеркнуто, что результаты кластеризации требуют содержательной интерпретации экспертами, поскольку каждая выявленная группа должна быть осмыслена с точки зрения предметной области.

Во втором разделе **«Обзор и сравнительный анализ моделей кластеризации»** представлено подробное рассмотрение современных алгоритмов кластерного анализа и оценка их применимости к социальным сетям. В работе проведен сравнительный анализ следующих моделей кластеризации: классический алгоритм k-средних, алгоритмы кластеризации на основе плотности DBSCAN и его иерархическая модификация HDBSCAN, метод спектральной кластеризации, алгоритмы обнаружения сообществ в графах Louvain и Leiden, метод распространения меток (Label Propagation), а также подходы ансамблевой кластеризации (объединяющей результаты нескольких алгоритмов). Для каждого алгоритма изложены принцип действия и ключевые параметры, рассмотрены достоинства и ограничения в контексте социальных данных. В частности, отмечено, что метод k-средних отличается простотой и скоростью, но требует предварительного задания числа кластеров и плохо обнаруживает кластеры сложной формы; алгоритм DBSCAN способен выделять кластеры произвольной формы и отсеивать «шум» данных, однако чувствителен к выбору параметров плотности; HDBSCAN избавляет от необходимости задавать число кластеров, строя иерархию плотностных областей; спектральные методы эффективны для обнаружения структур в графах через использование свойств матриц смежности; алгоритмы Louvain и Leiden оптимизируют модульность и хорошо справляются с разбиением крупных социальных графов на сообщества; метод распространения меток предоставляет быстрый, хотя и несколько грубый, способ найти сообщества; ансамблевые методы позволяют комбинировать преимущества разных алгоритмов, повышая надежность кластеризации. Проведен сравнительный разбор работы алгоритмов на типичных сценариях соцсетевых данных и сделан вывод об отсутствии универсального метода: выбор оптимального алгоритма должен учитывать природу данных и цели анализа.

В третьем разделе **«Экспериментальное исследование кластеризации социальных сетей»** описано практическое применение рассмотренных алгоритмов на реальных данных и анализ полученных результатов.

Представлены используемые датасеты: социальный граф пользователей «ВКонтакте» (~10 тыс. аккаунтов с ~100 тыс. связей дружбы) и коллекция текстовых публикаций «Яндекс.Дзен» (~1 тыс. статей с тегами). Изложена методика эксперимента: для графа дружбы во «ВКонтакте» применены алгоритмы кластеризации графов (Louvain, Leiden, Label Propagation) с целью выявления сообществ пользователей; для множества пользователей также проведена кластеризация по их интересам (на основе списков сообществ, в которых состоят пользователи) с использованием алгоритмов k-средних, DBSCAN, HDBSCAN; для набора статей Дзена выполнена контентная кластеризация – тексты преобразованы в векторы признаков (TF-IDF), после чего группированы методом k-средних, DBSCAN, а также иерархической кластеризацией для наглядности. Проведена оценка качества кластеризации: для графа «ВКонтакте» измерена модульность разбиения (алгоритм Louvain обнаружил порядка 7 крупных сообществ с значением модульности около 0.4, что свидетельствует о четкой кластерной структуре социальной сети); для кластеров пользователей по интересам сравнение осуществлялось с известными атрибутами (например, география – города проживания, что показало тенденцию образования кластеров по региональному признаку); для кластеризации статей проверена тематическая однородность кластеров на основе имеющихся у статей тегов и вычислен показатель согласованности (например, значение нормализованной взаимной информации между кластеризацией и разбиением по основным тегам). Полученные кластеры подвергнуты содержательной интерпретации: для сообществ пользователей «ВКонтакте» определены характерные профили (например, выделены кластеры, соответствующие географически локализованным группам студентов, сообществам по интересам — любители авто, музыки и т.п.); для кластеров статей Дзена построены облака наиболее частотных слов и выявлены доминирующие темы, что позволило охарактеризовать каждый кластер тематически (например, «кулинария», «путешествия», «технологии» и др.). Эксперимент подтвердил применимость рассмотренных алгоритмов к

данным российских социальных сетей и проиллюстрировал различия в их работе. В завершение раздела приведены рекомендации по практическому применению кластеризации: показано, что для комплексного анализа соцсетей целесообразно сочетать несколько методов (например, объединять результаты кластеризации графа дружбы и интересов пользователей через ансамблевые подходы) для получения более полноценной картины; также обозначены перспективы дальнейших исследований, включая изучение динамики кластеров во времени и расширение комбинированных методов на больший набор данных (учет переписки, содержимого мультимедиа и т.д.).

## ЗАКЛЮЧЕНИЕ

Выполненное исследование посвящено анализу моделей кластеризации и их применению к данным социальных сетей. В ходе работы был проведен теоретический обзор кластерного анализа, рассмотрены особенности социальных сетей как объекта кластеризации, изучены и сравнены современные алгоритмы, а также проведен эксперимент на реальных данных российских соцплатформ.

Основные результаты работы:

Теоретические основы. Даны определения кластеризации и социальных сетей, описаны характерные свойства соцсетевых данных (графовая структура, масштаб, гетерогенность, динамичность) и показано, почему кластеризация является ключевым методом для их анализа. Выделены основные задачи: обнаружение сообществ, сегментация пользователей, группировка контента, поиск аномалий - и продемонстрировано значение решения этих задач для прикладных сфер (маркетинга, социологии, безопасности). Кластеризация позволяет упростить сложную картину социальных данных, выявив укрупненные группы - что делает большие данные понятными и полезными.

Экспериментальное подтверждение. На реальных данных ~10 тыс. пользователей VK и 1 тыс. статей Дзена были применены указанные алгоритмы. Эксперимент показал:

• Социальный граф VK естественно распадается на сообщества, которые алгоритмы корректно обнаруживают (модулярность ~0.4). Сообщества соответствуют либо территориальным, либо социально-групповым объединениям. Алгоритм Leiden дал более детальную структуру (разделив некоторые крупные сообщества), а Label Propagation - очень быстрое, но менее стабильное разбиение.

• По интересам пользователей выделены несколько сегментов (музыкальные, новостные, спортивные и т.д.). K-средних обеспечил общую сегментацию, DBSCAN/HDBSCAN нашли узкие группы (фан-клубы, нишевые хобби) и отделили ~20% «непохожих» пользователей как шум. Это показывает, что в соцсетях значительная часть аудитории принадлежит к очевидным кластерам, но есть и много уникальных пользователей.

• Ансамблевая кластеризация (объединение графового и интересового анализа) выделила наиболее сплоченные сообщества: группы людей, связанных и дружбой, и общими интересами. Такие группы - ценные целевые аудитории, хотя их не так много.

• Кластеризация контента успешно сгруппировала статьи по темам, практически совпав с человеческой классификацией по тегам. Это подтверждает применимость кластер-методов для автоматической тематической разбивки большого потока пользовательских материалов.

• Качественные и количественные метрики (силуэт, модульность, NMI) в целом подтвердили друг друга: наилучшие по значениям разбиения были и наиболее интерпретируемыми. Например, кластеризация с NMI=0.7 с городами явно отражала города, а та, у которой NMI=0.3 - давала смешанные группы.

4. Рекомендации и применения. На основе анализа были предложены конкретные рекомендации для различных сфер:

• В государственном управлении - использовать знания о кластерах для более адресного общения с населением (например, учитывая локальные сообщества при распространении муниципальной информации).

- В безопасности - применять кластеризацию для мониторинга потенциально опасных групп (алгоритмы выделят их как плотные сообщества, что поможет надзору), а также для прогнозирования распространения влияний по сети.
- В социологии - использовать кластерный анализ соцсетей как инструмент выявления социальных групп и структур (цифровое отражение групп по интересам, возрасту, ценностям), что может дополнять традиционные методы опроса.
- В бизнесе и маркетинге - активно сегментировать онлайн-аудиторию на основе кластеризации, чтобы адаптировать рекламные кампании под каждое сообщество интересов; улучшать рекомендательные сервисы за счет кластеризации контента и пользователей; выявлять лидеров мнений внутри кластеров для целевого влияния (например, программы рефералов).

Кластеризация социальных сетей - чрезвычайно полезный метод, позволяющий преобразовать большие неструктурированные данные о пользователях и их связях в осмысленные группы. Различные алгоритмы взаимодополняют друг друга: нет универсального решения, но есть набор инструментов, из которого выбирают оптимальный под задачу. В частности, для социальных графов оптимальны Louvain/Leiden, для атрибутов - k-средних при наличии четких сегментов или DBSCAN при неявных формах кластеров, для максимальной полноты - комбинация методов. Кластеризация дает возможность выявить скрытые сообщества, которые ранее могли не учитываться - от локальных активистов до интернет-фандомов - и тем самым позволяет принимать более информированные решения (будь то маркетинговые стратегии или управленческие вмешательства).

В работе сознательно сделан акцент на российских соцсетях (VK, Одноклассники, Яндекс.Дзен, Telegram) вместо зарубежных (Facebook, Instagram и др.), что соответствует текущим реалиям информационной сферы. Показано, что все рассмотренные методы вполне применимы и к русскоязычным платформам, а примеры кластеров иллюстрированы на основе

отечественного контента. Это демонстрирует, что аналитика социальных сетей универсальна и инструментарий кластеризации успешно служит для исследования любых социальных графов и сообществ - независимо от страны и платформы.

Перспективы и дальнейшая работа. В будущем исследовании планируется углубиться в несколько направлений:

- Изучение динамики кластеров во времени: как сообщества в VK распадаются или объединяются, мигрируют ли пользователи между кластерами (для этого потребуются методы динамической кластеризации или отслеживание «жизни» кластеров).
- Применение алгоритмов кластеризации к мессенджерам (например, Telegram) - там связи неявные, но можно строить графы пересечений аудиторий каналов или графы перепостов. Интересно проверить, выделятся ли в Telegram сообщества по интересам аналогично VK.
- Расширение ансамблевых методов на больший набор видов данных: например, учесть помимо дружбы и интересов еще и переписку (топики сообщений) или геолокацию чек-инов - т.е. создавать мультимодальные кластеры. Это позволит еще точнее описать социальные группы (скажем, «друзья, которые часто бывают в одних местах и обсуждают одни темы» - очень сильная связка).
- Оптимизация алгоритмов под большие российские сети: возможно, разработка специальных реализаций Louvain/Leiden с учетом распределенных вычислений для обработки, к примеру, всей сети «ВКонтакте» (сотни миллионов узлов).
- Анализ качества кластеризации глазами самих пользователей: например, проводить опросы участников одного автоматически найденного кластера - считают ли они себя частью единого сообщества? Это поможет валидировать социальную значимость кластеров (не только метрики).

Подводя итог, можно отметить, что кластеризация - эффективный способ «приручить» большие данные социальных сетей. Применение

современных моделей кластеризации к российским социальным платформам позволяет извлечь новую ценную информацию о структуре онлайн-общения и интересов. Эти знания могут быть внедрены в различных областях - от улучшения пользовательского опыта до повышения эффективности управления и коммуникаций в цифровом обществе. Развитие методов кластеризации и их адаптация к специфике соцсетей продолжает оставаться актуальной научной и практической задачей в условиях роста роли социальных медиа в жизни общества.