УДК 004.032.26

Идентификация неявных сообществ и анализ социальных сетей с использованием искусственного интеллекта

Комардин Максим Алексеевич(*) komardin.macsim@gmail.com

МГТУ им. Н.Э. Баумана, Москва, Россия

Аннотация. *Статья предоставляет обзор современных методов и алгоритмов обнаружения скрытых сообществ в социальных сетях, что является важной задачей в условиях быстрого роста объемов данных и сложности сетевых структур. В статье рассматривается выявление неявных связей между пользователями, сформированных общими интересами, социальными окружениями или поведенческими паттернами. Предлагается всесторонняя оценка подходов к моделированию сетевых структур и интерпретации межличностных связей с использованием анализа текстового контента, социальных взаимодействий и метрик влияния. Экспериментальные исследования подтверждают разработанные алгоритмы на различных структурах социальных графов, включая реальные и синтетические данные, подчеркивая влияние топологических и поведенческих факторов на качество обнаружения сообществ.*

***Ключевые слова*:** *социальные сети, кластеризация, графовые алгоритмы, межличностные связи, машинное обучение, векторные представления, семантический анализ*

Implicit community identification and social network analysis using artificial intelligence

Kamardin Maxim Alekseevich komardin.macsim@gmail.com

BMSTU, Moscow, Russia

Abstract. The article provides an overview of modern methods and algorithms for detecting hidden communities in social networks, a crucial task given the rapid growth of data volumes and the complexity of network structures. The paper explores the identification of implicit connections among users formed by shared interests, social environments, or behavioral patterns. A comprehensive evaluation of approaches to modeling network structures and interpreting interpersonal connections is provided, leveraging text content analysis, social interaction, and influence metrics. Experimental studies validate the developed algorithms on various social graph structures, including real-world and synthetic data, highlighting the influence of topological and behavioral factors on community detection quality.

Keywords: social networks, clustering, graph algorithms, interpersonal communication, machine learning, vector representations, semantic analysis

Введение. Социальные сети стали неотъемлемой частью современного общения, генерируя динамичные и сложные структуры данных взаимодействия пользователей. Эти данные содержат информацию о поведении и предпочтениях людей, их связях и тенденциях в различных сообществах. Сообщества в социальных медиа можно определить как группы пользователей, которые разделяют общие интересы, социальные роли, схожее поведение или другие факторы, влияющие на их взаимодействия. Выявление таких сообществ является сложной задачей, поскольку не все из них формируются явно. Неявные сообщества включают пользователей, связанных через косвенные отношения, такие как пересекающиеся интересы, схожие предпочтения или частые взаимные действия (лайки, комментарии, упоминания).

Выявление неявных сообществ в социальных сетях представляет собой исследовательский интерес из-за масштаба и сложности данных. Такие сети характеризуются высокой плотностью и гетерогенностью связей, что затрудняет использование традиционных методов анализа. Кроме того, динамическая природа сетей, где взаимодействия между пользователями и их активность постоянно меняются, требует адаптивных подходов и методов, которые могут быстро обрабатывать поступающие данные и корректировать модели в реальном времени.

Помимо академических исследований, задача выявления скрытых сообществ в социальных сетях имеет большое прикладное значение. Например, результаты такого анализа могут быть использованы в бизнес-приложениях для создания рекомендательных систем, предлагающих персонализированный контент с учетом скрытых интересов и предпочтений пользователей. В социальной аналитике знания о неявных сообществах важны для понимания структуры социальных связей, что может быть полезно для выявления ключевых лидеров мнений, анализа распространения информации, прогнозирования электорального поведения, а также для предсказания социальных и маркетинговых трендов.

Однако сложность задачи выявления неявных сообществ увеличивается с ростом объема данных и эволюцией самих социальных сетей. Влияние и поведение пользователей могут изменяться, что требует гибких и масштабируемых подходов для анализа и выявления таких сообществ. Эти подходы должны учитывать, как топологические характеристики сетевых структур, так и контент, который пользователи создают и обмениваются друг с другом.

Материалы и методы. Существует множество моделей и методов для выявления таких неявных сообществ, каждая из которых имеет свои особенности, ограничения и область применения. Основные подходы можно разделить на несколько категорий: алгоритмы на основе кластеризации, методы распространения меток, графовые нейронные сети и методы, учитывающие модульность сети.

Один из наиболее известных методов обнаружения сообществ — это метод кластеризации, который включает, например, алгоритмы k-средних, DBSCAN и Louvain. Эти методы используются для разделения сети на груп-

пы на основе близости между узлами, которая чаще всего измеряется количеством общих связей.

Недостатками многих методов кластеризации являются ограниченность в обнаружении только плотных сообществ, плохая устойчивость к шуму и высокая чувствительность к некоторым параметрам, таким как выбор начальных точек для k-средних или влияние различных параметров плотности в DBSCAN. Это может существенно повлиять на итоговый результат и точность обнаружения сообществ [1].

Методы распространения меток, такие как алгоритм распространения меток (Label Propagation Algorithm, LPA), основаны на передаче уникальных идентификаторов или «меток» от узла к узлу в графе. Этот подход основывается на том, что в социальных сетях информация или влияние передаются по цепочке между пользователями, постепенно охватывая все больший сегмент сети. Этот алгоритм эффективен и масштабируем, но его основной недостаток — нестабильность. LPA может давать разные результаты при каждом запуске, что затрудняет его использование для воспроизведения результатов. Кроме того, он не учитывает степень связности между узлами, что может привести к неправильному определению сообществ в случаях с неявными связями между пользователями [2, 3].

Методы, основанные на графовых нейронных сетях (GNN), стали все более популярными в последние годы. Эти модели используют глубокое обучение и могут учитывать как топологические свойства графа, так и атрибуты узлов, что особенно полезно для социальных сетей, где контекст взаимодействий имеет большое значение. Основными недостатками графовых нейронных сетей являются их трудности с обучением, требующие значительных вычислительных ресурсов, а также высокая зависимость от объемов данных.

Модульность — это мера, которая оценивает, насколько сильно связаны узлы внутри одного сообщества по сравнению с внешними связями. Алгоритмы, основанные на модульности, стремятся максимизировать это значение, что позволяет выявлять плотные группы узлов с большим количеством внутренних связей (Newman-Girvan, Clauset-Newman-Moore (CNM)). Алгоритмы на основе модульности хорошо работают на относительно малых графах, но для больших социальных сетей, особенно с перекрывающимися сообществами, они требуют значительных доработок и адаптации [4].

Анализ сообществ в социальных сетях направлен на выявление кластеров пользователей в обширных социальных графах. Основной задачей является выбор метода, способного эффективно работать со сложными и крупными графовыми структурами с комплексной топологией. Особое внимание следует уделить скорости обработки и использованию вычислительных ресурсов, особенно учитывая отсутствие обширных размеченных наборов данных для обучения.

Алгоритмы, основанные на графовых нейронных сетях (GNN), требуют значительных вычислительных ресурсов и объемов данных для обучения, что затрудняет их применение в реальном времени для анализа крупных социальных сетей [5, 6]. Многие классические алгоритмы кластеризации, такие

как k-средних, имеют ограничения, включая высокую вычислительную сложность и низкую масштабируемость, что делает их трудными для применения к данным в виде графов без предварительного преобразования.

В связи с вышеупомянутыми трудностями, алгоритм DBSCAN (Density-Based Spatial Clustering of Applications with Noise) был выбран для анализа социальных сетей. DBSCAN эффективно идентифицирует кластеры на основе плотности точек, что позволяет выделять сообщества пользователей с общими интересами или поведением [7-9].

Принцип работы DBSCAN:

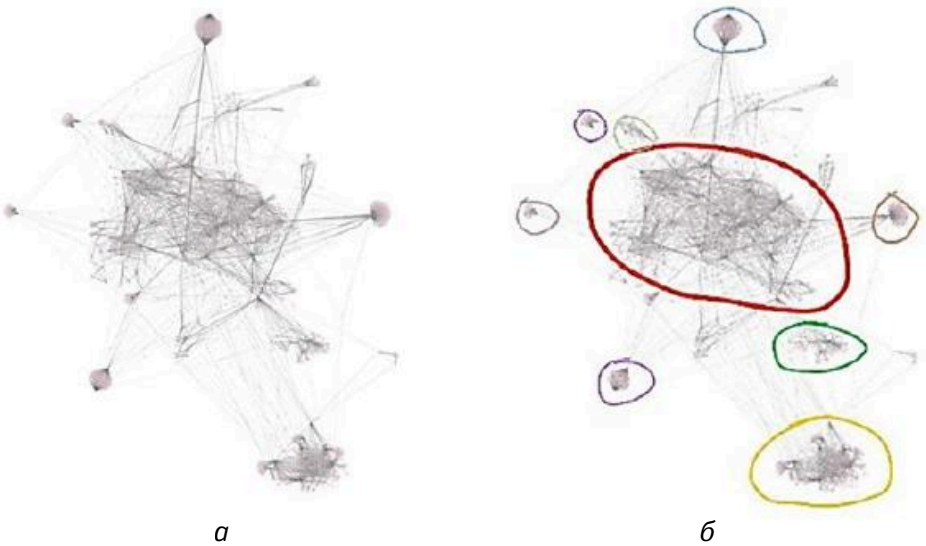
– выбор точки: Алгоритм начинает с произвольной непосещенной точки и определяет ее ε-окружность (окружность радиусом ε);

– проверка плотности: Если в этой окружности содержится не менее minPts точек, то точка считается ядром кластера, и начинается формирование нового кластера;

– расширение кластера: Все точки в ε-окружности ядра добавляются в кластер. Затем для каждой из этих точек проверяется, являются ли они ядрами. Если да, их окружности также добавляются к кластеру. Этот процесс продолжается до тех пор, пока кластер не будет полностью сформирован;

– обработка шума: Точки, не входящие в ни один кластер и не являющиеся ядрами, помечаются как шум.

Следующий вопрос заключается в том, как идентифицировать эти кластеры. Классические реализации являются примерами того, как реализовать алгоритм. Визуализация выше была создана с использованием библиотеки d3.js. Это, по сути, визуализация проекции графа в двумерное пространство с моделированием взаимодействий (сил притяжения и отталкивания) между вершинами (рис. 1). Этого достаточно для общего понимания и схематичного представления сложной социальной структуры, но такая визуализация не может служить для анализа.



а *б*

Рис. 1. Пример социального графа связей 5000 пользователей (*а*) и потенциальные кластеры (на основе локальной плотности графа) (*б*)

Для правильного отражения взаимосвязей n пользователей необходимо создать матрицу n×n, где в i-й строке для i-го пользователя id все значения для id «не друзья» равны нулю. Для визуализации этого без нарушения эвклидовой геометрии потребуется n-мерное пространство. Поэтому, поскольку проекция не точно отражает произвольно искривленное пространство, анализ близости вершин по длине ребер между ними на визуализации будет некорректным.

Подход к хранению данных в виде матрицы n×n является чрезвычайно затратным и неэффективным как с точки зрения объема необходимой памяти, так и с точки зрения вычислительных ресурсов, требуемых для анализа. В большинстве случаев подавляющее большинство «ячейки» будут равны 0, и каждый n-й пользователь увеличивает размер таблицы на $2n - 1$ ячеек (рис. 2). Использование n-мерного пространства позволяет корректно отображать взаимосвязи между всеми пользователями, но анализ такого пространства алгоритмически сложен, плохо масштабируем и, что самое важное, не востребован.

Для оптимизации хранения и обработки данных был реализован алгоритм расчета близости вершин, основанный на их структурной близости в графе; в качестве единицы хранения данных рассматриваются только существующие отношения, а параметр близости вершин определяется через пропорцию совпадений меток, назначенных в ходе нескольких итераций метода обхода в ширину. BFS (поиск в ширину) — это алгоритм обхода графа, который начинается с исходной вершины и посещает соседние вершины слоями, переходя к следующему слою, пока все вершины в графе не будут покрыты. Метод расширен использованием «цветного BFS» и оптимизирован введением параметра максимального радиуса распространения BFS.

Таким образом, вершины, которые находятся близко друг к другу и имеют много общих связей, получают одинаковые метки с большей вероятностью, чем те, которые находятся далеко друг от друга или не имеют общих связей (см. рис. 2).

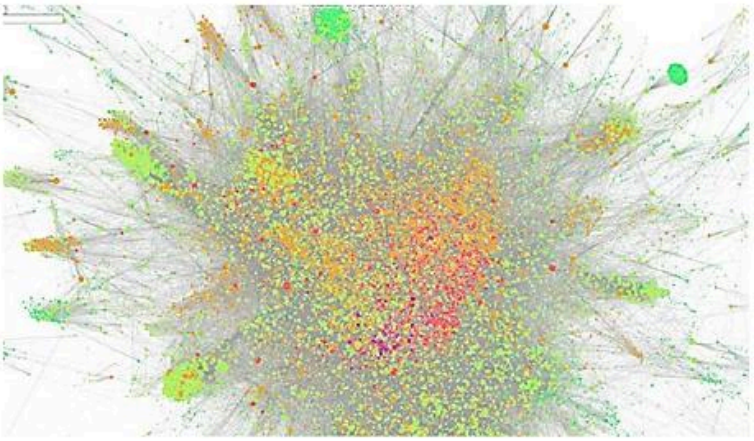


Рис. 2. Пример визуализации одной из итераций цветового bfs для 30000 уникальных идентификаторов пользователей

На каждой итерации цветовые метки узлов (пользователей) обновляются на основе сходства их связей с соседними пользователями (рис. 3).

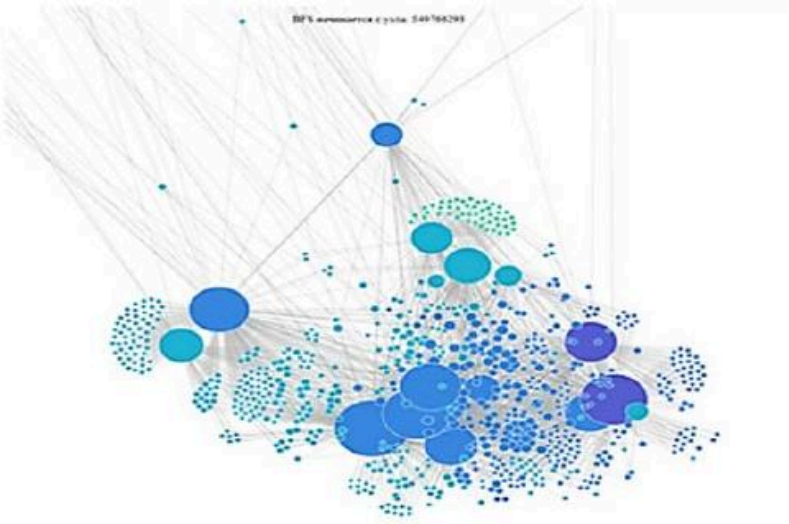


Рис. 3. Пример кластера

Постепенно это приводит к тому, что пользователи с близкими интересами получают схожий набор цветов, что позволяет оценить их близость и центральность в сообществе (рис. 4).

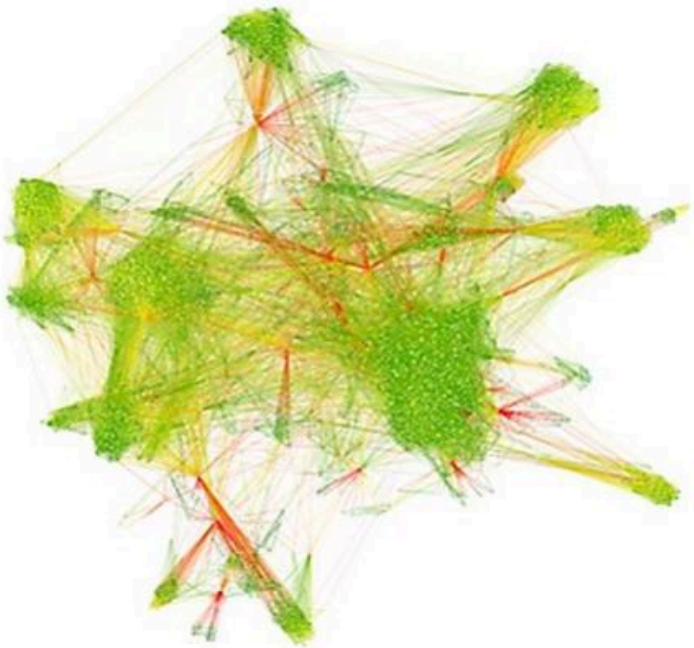


Рис. 4. Цветовая визуализация близости между вершинами

После этого, если обозначить процент совпадающих меток как степень близости вершин, то, задав eps-радиус в качестве минимального показателя

сходства для определения соседства, а minPts — как минимальное количество соседей вершины, необходимое для ее рассмотрения в качестве основной, становится возможным выполнить кластеризацию с помощью алгоритма DBSCAN (рис. 5).

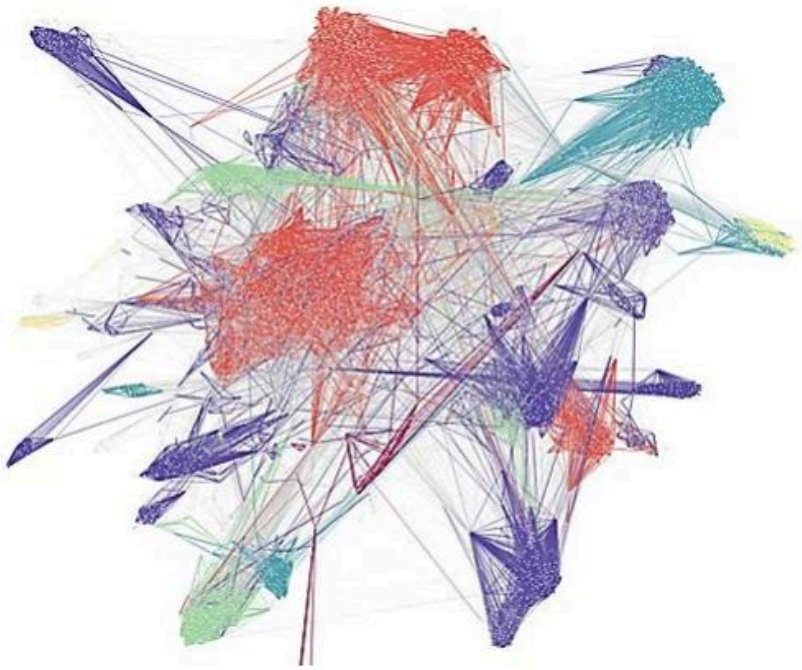


Рис. 5. Пример кластеризации с помощью DBSCAN

Результаты. После кластеризации социального графа становится возможным исследование взаимосвязей между структурными паттернами, моделями развития неявных сообществ и параметрами пользователей, входящих в их состав [10].

С помощью пользовательских меток можно формировать параметрические маски по месту жительства, полу, возрасту, образованию, семейному положению и другим персональным характеристикам. Однако наиболее интересным является анализ кластеров по критерию ценностно-ориентационного единства (ЦОЕ), а также кластеризация на основе семантического содержания публикаций пользователей и сообществ, на которые они подписаны.

В качестве базы данных для исследования используются наборы подписок и публикаций каждого пользователя, а также публикации сообществ, на которые они подписаны. Анализ осуществляется с использованием LLM-модели BERT. Тематическое моделирование с помощью векторных представлений BERT позволяет выявлять скрытые темы в публикациях сообществ, анализируя семантическое содержание текстов.

Метод основан на создании эмбеддингов — векторных представлений слов и предложений, отражающих их смысл с учетом контекста, в котором они чаще всего встречаются. Идея эмбеддингов заключается в преобразовании дискретных текстовых данных (слов, токенов) в числовую форму, при-

годную для обработки алгоритмами машинного обучения и нейронными сетями. Эмбеддинги обеспечивают, что слова, фразы или другие элементы данных, имеющие схожий смысл, получают близкие векторные представления. Например, слова «кот» и «кошка» будут иметь схожие эмбеддинги, поскольку они обозначают одно и то же и употребляются в схожих контекстах.

BERT (Bidirectional Encoder Representations from Transformers) — это архитектура модели для обработки естественного языка, основанная на механизме Transformer и предложенная Google в 2018 году. Главное преимущество BERT заключается в его двунаправленном подходе к обучению: модель анализирует текст одновременно слева направо и справа налево, что позволяет учитывать контекст слова со всех сторон.

Кластеризация пользователей осуществляется по среднему вектору их подписок, что позволяет группировать пользователей с похожими ценностями и интересами (рис. 6). Такой подход эффективен для глубокого анализа предпочтений и сегментации аудитории, поскольку учитывает не просто частоту слов или их простые ассоциации, а именно их контекстуальное значение [10].

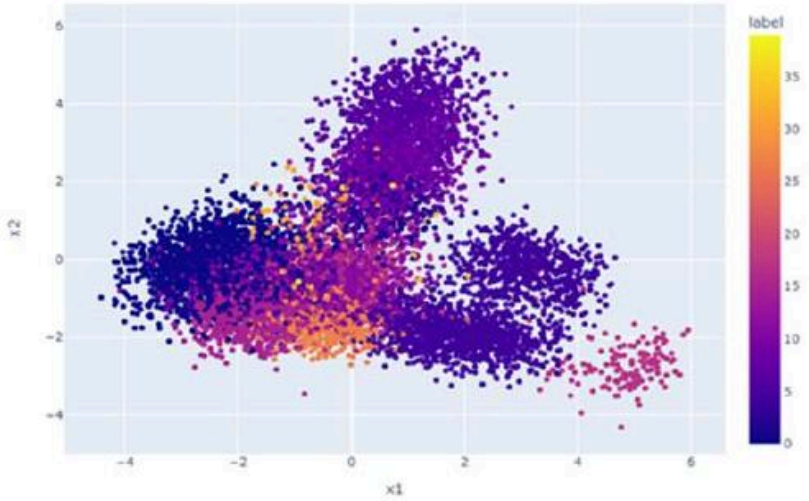


Рис. 6. Визуализация кластеризации векторных представлений новостей

Несмотря на все преимущества, у подхода есть и недостатки — модель BERT не учитывает форматы, отличные от текста. В XXI веке, с развитием интернет-технологий, невозможно представить крупную социальную сеть, не поддерживающую публикацию медийного контента. Игнорирование этого факта может привести к значительному снижению качества анализа, поэтому для повышения эффективности необходимо использовать мультимодальные ИИ-модели, способные анализировать публикации со сложной структурой в целом.

Обсуждение. Хотя глубокое обучение совершило революцию в области компьютерного зрения, текущие подходы сталкиваются с рядом серьезных

проблем. Классические наборы данных для компьютерного зрения требуют больших затрат на создание, при этом они охватывают лишь узкий набор визуальных концепций. Стандартные модели компьютерного зрения хорошо справляются только с одной задачей и требуют значительных доработок для адаптации к новым задачам. Однако даже специализированные модели, показывающие высокие результаты в тестах, обладают ограниченной гибкостью.

CLIP (Contrastive Language-Image Pretraining) — мультимодальная модель, разработанная OpenAI, которая решает эти проблемы, связывая текст и изображение в едином векторном пространстве. CLIP обучается с использованием контрастного метода: модель стремится сблизить эмбеддинги изображений и соответствующих им текстов, одновременно разделяя несоответствующие пары. Она также реализует совместное представление текстов и изображений: CLIP создает эмбеддинги таким образом, чтобы можно было напрямую сравнивать изображения и текстовые описания. Это позволяет определять степень соответствия изображения заданному описанию или искать похожие изображения по текстовому запросу.

Эмбеддинги позволяют модели понимать семантику слов в зависимости от контекста. Векторные представления данных являются плотными, что делает их более эффективными по сравнению с одноразрядными представлениями (one-hot encoding), где каждая категория кодируется отдельным вектором, независимым от других классов. Кроме того, эмбеддинги могут применяться для широкого круга задач, что ускоряет процесс предобучения и повышает точность модели в данной области.

Базовая модель была предварительно обучена на крупных наборах данных: CLIP тренируется на огромных объемах информации из интернета, где пары «изображение – текст» берутся из естественных источников (например, описания фотографий или публикации в социальных сетях). Это позволяет модели обладать широкими знаниями и решать множество задач. По своей конструкции сеть может получать инструкции на естественном языке для выполнения разнообразных задач.

CLIP состоит из двух отдельных моделей:

– текстовый энкодер — трансформер, аналогичный моделям BERT или GPT, который преобразует текстовое описание в эмбеддинг;

– изображенческий энкодер — сверточная нейросеть (ResNet) или Vision Transformer (ViT), который преобразует изображение в эмбеддинг.

Процесс обучения CLIP:

1. Подготовка данных: пары «изображение – текстовое описание» собираются из Интернета.

2. Извлечение эмбеддингов: текстовый энкодер преобразует текст в эмбеддинг, а энкодер изображений делает то же самое с изображением.

3. Контрастное обучение: создается матрица схожести, в которой вычисляется косинусное сходство между эмбеддингами всех изображений и текстов. Модель стремится максимизировать сходство между правильными парами (изображение и его описание) и минимизировать его для неверных пар.

4. Функция потерь: используется функция потерь на основе InfoNCE Loss (сопоставление по косинусному сходству), которая поощряет правильные соответствия и штрафует за ошибки. Потери считаются как для изображений, так и для текстов, что делает процесс симметричным.

5. Zero-shot обучение: после тренировки CLIP может выполнять широкий спектр задач без дообучения, что является одним из его ключевых преимуществ.

Таким образом, мультимодальные модели позволяют анализировать записи пользователей и сообществ, учитывая не только текст, но и визуальный контент.

Заключение. Кластеризация социальных графов позволяет проводить углубленный анализ структурных закономерностей, динамики формирования неявных сообществ и характеристик пользователей внутри этих сообществ. Использование пользовательских меток дает возможность генерировать параметрические маски на основе демографических данных, таких как место жительства, пол, возраст, образование и семейное положение. Особенно важным является анализ кластеров по признаку ценностно-ориентационного единства (VOE), а также семантическая кластеризация пользовательских публикаций и контента подписанных сообществ.

Семантический анализ, выполняемый с помощью моделей, таких как BERT и CLIP, играет ключевую роль в понимании поведения пользователей и сообществ. В то время как BERT эффективно анализирует текст благодаря контекстно-ориентированным эмбеддингам, его возможности ограничены при обработке мультимодального контента, широко распространенного в современных социальных сетях. Эти ограничения преодолеваются мультимодальными моделями ИИ, такими как CLIP, которые объединяют текстовые и визуальные данные в единое векторное пространство, значительно повышая точность и полноту профилирования пользователей и сообществ.

Применение этих методов кластеризации и семантического анализа позволяет глубже понять структуру социальных сетей и поведение пользователей, что способствует развитию систем сегментации аудитории, рекомендательных алгоритмов и аналитики социальных взаимодействий.

Список источников

[1] Omran M.G.H., Engelbrecht A.P., Salman A. An overview of clustering methods. Intelligent Data Analysis, 2007, vol. 11, no. 6, pp. 583–605.

[2] Schaeffer S.E. Graph Clustering. Computer Science Review, 2007, vol. 1 (1), pp. 27–64.

[3] Slota G.M., Rajamanickam S., Madduri K. BFS and coloring-based parallel algorithms for strongly connected components and related problems. IPDPS 2014, IEEE Computer Society, 2014, pp. 550–559. https://doi.org/10.1109/IPDPS.2014.64

[4] Wang K., Ding Y., Han S.C. Graph neural networks for text classification: a survey. Artif Intell Rev, 2024, vol. 57, art. no. 190. https://doi.org/10.1007/s10462-024-10808-0

[5] Webb S.D. SOAR—the solution to the high prevalence of heart disease in Appalachian Kentucky. J Public Health (Berl.) 29, 1331–1338 (2021). https://doi.org/10.1007/s10389-020-01248-5

[6] Toschev A., Talanov M. Artificial Cognitive Architectures Review. BioNanoSci. 2020. Vol. 10, pp. 811–823. https://doi.org/10.1007/s12668-020-00768-4

[7] Панилов П.А. Использование методов нейроэволюции для автоматизации оптимизации алгоритмов киберзащиты в когнитивных информационных центрах. *Известия Тульского государственного университета*. *Технические науки*, 2024, № 11, с. 16–25.

[8] Khan Z., Fu Y. Exploiting BERT for Multimodal Target Sentiment Classification through Input Space Translation. Proceedings of the 29th ACM International Conference on Multimedia, New York, NY, USA, 2021, pp. 3034–3042. https://doi.org/10.1145/3474085.3475692

[9] Суханов В.А., Цибизова Т.Ю. Архитектура системы развития и поддержки профессиональной интеллектуальной деятельности. *Динамика сложных систем* — XXI *век*, 2020, т. 14, № 1, с. 23–31.

[10] Коротеев М.В. Обзор некоторых современных тенденций в технологии машинного обучения. E-Management, 2018, т. 1, № 1, с. 26–35.