Федеральное государственное образовательное бюджетное учреждение
высшего образования
«ФИНАНСОВЫЙ УНИВЕРСИТЕТ ПРИ ПРАВИТЕЛЬСТВЕ РОССИЙСКОЙ ФЕДЕРАЦИИ»

Факультет информационных технологий и анализа больших данных
Кафедра анализа данных и машинного обучения

Выпускная квалификационная работа
на тему: «Построение модели определения отраслевой принадлежности компании на основе ее медийной активности»

Направление подготовки: 01.03.02 Прикладная математика и информатика
Профиль: Анализ данных и принятие решений в экономике и финансах

Выполнил студент учебной группы
ПМ20-1
Кудряшов Никита Александрович

Научный руководитель работы
ассистент
Блохин Никита Владимирович

ВКР соответствует предъявляемым требованиям:
Заведующий кафедрой анализа данных и машинного обучения, к.т.н., доцент

_____ Д. А. Петросов
«____» _____ 2024 г.

Москва 2024

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ 3
ГЛАВА 1. ТЕОРЕТИЧЕСКАЯ ОСНОВА ПРЕДСТАВЛЕНИЯ ДАННЫХ В ВИДЕ ГРАФА 6
1.1 Граф, как структура данных 6
1.2 Графовые нейронные сети (GNN): особенности и области применения 7
1.3 Предсказания связей в графах (Link Prediction) 10
1.3.2 Классические методы предсказаний связей в графах 11
1.3.3 GNN методы предсказаний связей в графах 17
1.3. Алгоритм распространения доверия (message passing) ..... **Ошибка! Закладка не определена.**
1.3. Генерация отрицательных примеров ..... **Ошибка! Закладка не определена.**
1.3. Метрики ..... **Ошибка! Закладка не определена.**
1.3. Постановка задачи к разработке ..... **Ошибка! Закладка не определена.**
1.3 Анализ существующих моделей предсказания связей в графах 17
1.4 Вывод 17
ГЛАВА 2. ПРОЕКТИРОВАНИЕ И РАЗРАБОТКА МОДЕЛИ ГРАФОВОЙ НЕЙРОННОЙ СЕТИ 18
ГЛАВА 3. АНАЛИЗ ПОЛУЧЕННЫХ РЕЗУЛЬТАТОВ 19
ЗАКЛЮЧЕНИЕ 20
СПИСОК ИСПОЛЬЗУЕМЫХ ИСТОЧНИКОВ 21
ПРИЛОЖЕНИЕ 25

ВВЕДЕНИЕ

Стремительное развитие информационных технологий и доступности информации сильно изменила модель ведения бизнеса. Компании активно используют различные медийные платформы для распространения информации о своей деятельности, достижениях, продуктах и услугах. Социальные сети, новостные сайты, блоги, форумы и другие онлайн-ресурсы стали неотъемлемой частью корпоративной коммуникации. Отображаемая там информация представляет компанию для конечного потребителя и возможных партнеров. Медийные платформы и хранящаяся там информация о компаниях выступают важным источником данных для аналитики и принятия стратегических решений внутри компании.

В условиях большого количества информации и конкурирующего рынка становиться необходимым разрабатывать эффективные методы анализа, позволяющие определить отраслевую принадлежность компании. Понимание, к какой отрасли можно отнести компанию, дает возможность не только определить ее конкурентное преимущество и позиционирование на рынке, но и принимать обоснованные стратегические решения.

Анализ медийной активности компании становится ключевым фактором для выявления тенденций в поведении потребителей, оценки влияния маркетинговых кампаний, а также для прогнозирования ее развития в контексте отраслевой динамики. Разработка модели определения отраслевой принадлежности компаний на основе медийной активности является важным шагом к повышению качества аналитики и принятия обоснованных управленческих решений в условиях развития цифровой инфраструктуры.

Актуальность работы заключается в том, что определение отраслевой принадлежности компании является нетривиальной задачей, которая требует привлечения современных методов и технологий: от сбора и формирования представления новостных данных до обучения глубокой нейронной сети.

Целью данной выпускной квалификационной работы является построение графой нейронной модели способной определить отраслевую

принадлежность компании на основе ее медийной активности.

Основные задачи для достижения поставленной цели:

- Исследование основных источников медийной активности различных компаний.
- Составление собственного датасета с информацией о медийной активности компаний.
- Анализ существующих инструментов для обучения моделей на основе графов.
- Разработка алгоритма, учитывающего взаимосвязи между компаниями в одной отрасли.
- Техническая реализация и совершенствование сформулированных идей.
- Анализ полученных результатов.

Объектами исследования являются источники медийной активности компании, а также способы обучения графовой нейронной модели.

Предмет исследования – построение и дальнейшее применение модели классификации отраслевой принадлежности компании.

Методология работы опирается на классические и современные подходы в задачах обработки и анализа связанных данных, представимых в виде графов. Для реализации задач на практике был использован язык python 3.10.

Выпускная квалификационная работа состоит из трёх основных глав, в каждой из которых содержатся параграфы.

В первой главе описывается решаемая задача. Вводятся основные теоретические понятия, такие как граф…

Во второй главе описываются основные идеи, математический и алгоритмический аппарат для реализации поставленной задачи.

Третья глава содержит описание данных, на которых будет тестироваться разработанная модель. В ней также будут описаны основные этапы разработки. Глава заканчивается описанием результатов экспериментов

и оценкой качества полученной модели.

ГЛАВА 1. ТЕОРЕТИЧЕСКАЯ ОСНОВА ПРЕДСТАВЛЕНИЯ ДАННЫХ В ВИДЕ ГРАФА

1.1 Граф, как структура данных

Граф – это математическая структура, состоящая из двух множеств: множества объектов или узлов ($V = \{v_1, \dots, v_2\}$) и множества связей или ребер между этими объектами ($E = \left(e_{ij}\right)^n, j = 1$) (1.1).

$$G = (V, E), V \in \{v_i | i \in N\}, E \subseteq \left\{\left(v_i, v_j\right) | \left(v_j, v_i\right) \in V^2\right\} \quad (1.1)$$

Графы являются одной из важнейших структур в математических и компьютерных науках, так как они лежат в основе многих процессов, моделей и алгоритмов [1]. При помощи графов возможно моделировать множество различных процессов: карты прокладки труб или дорог, социальные или компьютерные сети, такие как Интернет, сети компаний и финансовых операций, цепочки химических реакций и молекулярные взаимодействия, эпидемическое распространение болезней и многое другое.

Графы можно разделить на различные типы:

- Направленный – связи между вершинами в таком графе ориентированы и имеют определенное направление (1.2).

$$E \subseteq \left\{\left(v_i, v_j\right) | \left(v_i, v_j\right) \in V^2, \left(v_i, v_j\right) \neq \left(v_i, v_j\right), i \neq j\right\} \quad (1.2)$$

- Ненаправленный – связи между вершинами в таком графе не ориентированы и характеризуют отношение между узлами в обе стороны (1.3).

$$E \subseteq \left\{\left(v_i, v_j\right) | \left(v_i, v_j\right) \in V^2, \left(v_i, v_j\right) = \left(v_i, v_j\right), i \neq j\right\} \quad (1.3)$$

- Взвешенный – каждая из связей в таком графе имеет некоторый вес, который может быть выражен в различных характеристиках (1.4).

$$E \subseteq \left\{\left(v_i, v_j\right) | \left(v_i, v_j\right) \in V^2, W\left(v_i, v_j\right) \in R\right\} \quad (1.4)$$

Помимо этого, графы можно представить в виде матрицы смежности $A \in \{0,1\}^{N \times N}$, в которых $A_{ij} = 1$, если вершины i и j соединены и $A_{ij} = 0$ в ином случае.

1.2 Графовые нейронные сети (GNN): особенности и области применения

Использование графов в качестве структуры данных получило широкое распространение в области машинного и глубокого обучения благодаря возможности графов эффективно моделировать отношения и зависимости между наборами объектов. Подобная структура позволяет более детально представить сложные нелинейные взаимосвязи по сравнению с традиционными табличными данными, которые предполагают линейную взаимосвязь между данными.

Впервые использование графовых нейронных сетей было предложено в работе [2]. Графовые нейронные сети или Graph Neural Network (GNN) представляет собой класс моделей глубокого обучения, взаимодействующих с данными, представленными в виде графов.

Представление вершин, связей или в целом всего графа в виде эмбеддингов [3] позволяет облегчить работу графовых нейронных сетей. Эмбеддинг – есть математическая структура, которая находится внутри другой структуры. Когда говорится, что некоторый объект X является эмбеддингом объекта Y, то эмбеддинг выступает в виде инъективного и сохраняющего структуру отображения. Иными словами, эмбеддинг - непрерывное отображение некоторого объекта в вектор в пространство другой размерности. Векторы, в которые преобразуется структура графа, позволяют сохранить топологию и семантику исходной сети. Здесь к топологии относиться изначальный вид графа: типы или шаблоны связей, наличие кластеров и общее взаимное расположение узлов и ребер.

Базовая структура GNN включает в себя энкодер и декодер (рис. 1) [4]. Энкодер принимает на вход изначальный граф, генерирует и представляет эмбеддинги. Декодер в свою очередь обрабатывает и использует созданные эмбеддинги для составления прогнозов. Принцип кодирования информации с одинаковым вектором, основанным на сходстве окрестностей в графе, является основополагающим для сбора структурных данных о графе.

Благодаря такому подходу GNN применяет итеративные механизмы агрегации информации и распространения информации по сети, позволяя обучать узлы на основе их локального окружения. Конечные обученные эмбеддинги инкапсулируют информацию о структуре и связности графа, которую в дальнейшем можно использовать для решение привычных задач машинного обучения.

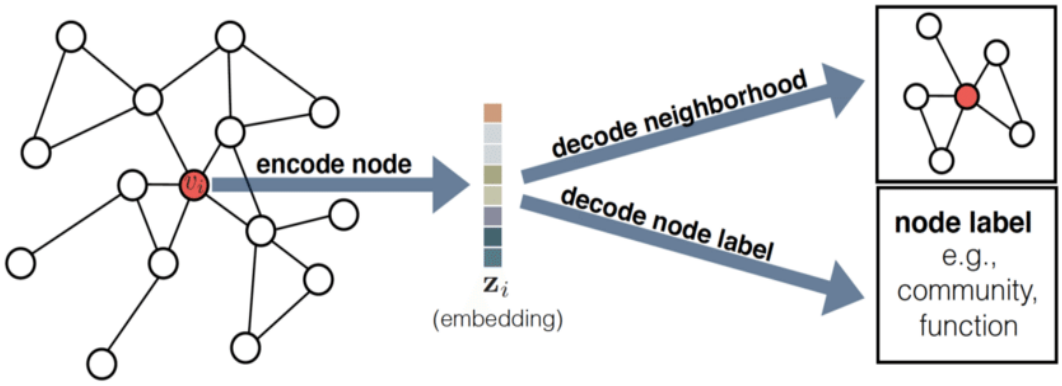

Рис. 1. Пример подхода к обработке графа.

Графовые нейронные сети – универсальные модели, которые могут выполнять различные задачи с использованием данных, структурированных в виде графа. Среди самых распространенных задач можно выделить следующие:

- Классификация узлов (Node Classification) - определение меток классов для каждого узла в графе на основе характеристик узла и топологии графа. Подобные модели могут быть использованы для определения категорий узлов в социальной сети [5], сети цитирований или фармакологических сетей [4].
- Предсказания связей в графах (Link Prediction) – задача заключается в прогнозировании установления связей между двумя узлами графа, которые в данный момент не имеют прямой связи (матрица смежности для них не завершена). Таким образом, подобные модели могут быть использованы в качестве рекомендательных систем, например, при подборе фильмов в онлайн кинотеатрах [15, 16], прогнозировании

потенциальных взаимосвязей в социальных сетях на основе общих интересов [6], прогнозирование соавторства в сетях цитирования [17] и многом другом.

- Классификация графов (Graph Classification) — это классификация целых сетей по предопределенным классам. К примерам решаемых задач можно отнести классификацию химических соединений на основе молекулярной структуры или лекарственных препаратов в целом [8], определение типов графов в социальных сетях, оценка безопасности графов на основе их общего статуса безопасности (на примере прогнозирования безопасности авиаперелетов на основе графов предыдущих полетов в определенных областях) [9], а также использование средств обработки естественного языка и графовых нейронных сетей для классификации текстов [10].
- Обнаружение сообществ (Community Detection) - идентификация групп узлов, плотно связанных внутри графа. Подобный подход может быть использован в обнаружении групп пользователей со схожими интересами в социальных сетях [11] или определения наиболее загруженных транспортных районов [12].
- Построение векторных представлений для графовых наборов данных (Graph Embedding) - создание и анализ низкоразмерных представлений для узлов, ребер или целых графов для сохранения информации о их структуре и топологии.
- Генерация графов (Graph Generation). Создание новых графов, которые имеют сходные структурные свойства с уже имеющимся набором данных, позволяет расширить изначальную выборку и тем самым улучшить качество создаваемых моделей. Подобный подход может использоваться в множестве задач, упомянутых ранее. В качестве примера реализации подобного подхода стоит упомянуть графовую генеративную модель, целью которой является изучение эффективных взаимосвязей узлов графа в end-to-end режиме для решения сложных

задач генерации архитектурных макетов с ограниченными графическими возможностями (Graph Transformer Generative Adversarial Network (GTGAN)) [13].

Из представленных подходов применения GNN подходящей для решения поставленных задач в рамках темы выпускной квалификационной работы является использование модели предсказаний связей в графах (Link Prediction).

1.3 Предсказания связей в графах (Link Prediction)

Как уже говорилось ранее, предсказание связей в графах – это задача прогнозирования существования связи между двумя узлами сети. Впервые эта задача была сформулирована в работе [14].

1.3.1 Гетерогенность и гомофильность графов

Link Prediction имеет множество названий в зависимости от области применения. Термин «предсказание связей» часто относится к предсказанию связей в однородных графах (homogeneous graphs), где узлы и связи между ними имеют только один тип. Примерами таких сетей можно назвать строение атома или граф остатков (рис. 2)

Сложнее бывает, когда сеть состоит из узлов, представляющих отдельные объекты с различными типами связей, соответствующих различным отношениям между объектами. Такие графы называются графами знаний. Иначе их можно назвать гетерогенными графами (heterogeneous graphs). В качестве примера можно привести сети регуляции генов или химических элементов (рис. 2) Различные вершины и связи в таких графах обладают собственными пространствами идентификаторов и набором характеристик (рис. 3), в отличие от однородных графов.

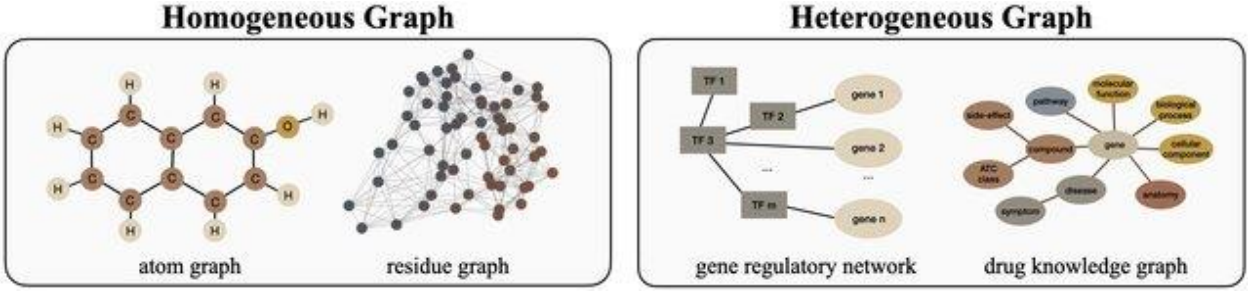

Рис 2. Примеры однородных и гетерогенных графов.

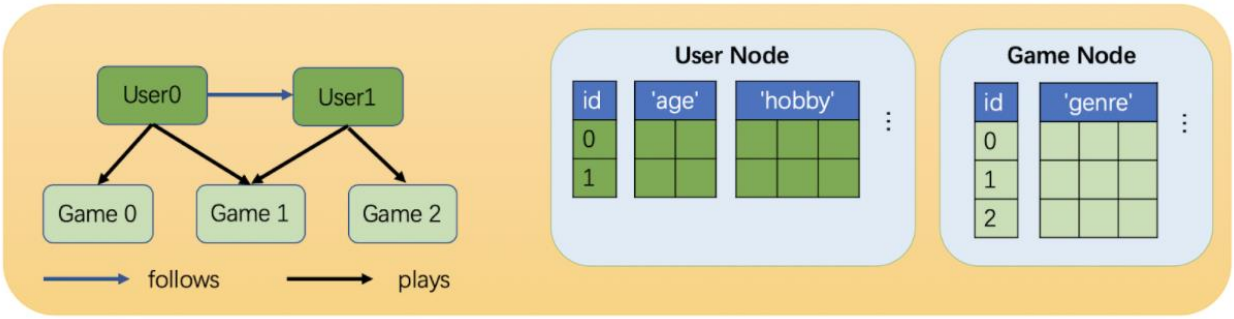

Рис. 3. Пример гетерогенного графа с двумя типами узлов («пользователь» и «игра») и двумя типами связей («подписан» и «играет»).

1.3.2 Классические методы предсказаний связей в графах

При рассмотрении различных методов решения задачи Link Prediction можно выделить два основных типа: классические методы и методы с использованием графовых нейронных сетей. Первые, в свою очередь, подразделяются на:

- Эвристические методы (Heuristic Methods) – методы используют простые, но эффективные оценки сходства узлов в качестве вероятности наличия связей [19].
- Методы изучения скрытых признаков (Latent-Feature Methods) в некоторой литературе данные методы также называются моделями скрытых факторов (latent-factor models) или эмбеддинговыми методами (embedding methods). Методы вычисляют скрытые свойства или представления узлов, часто получаемые путем разложения на множители определенной матрицы, полученной из сети, таких как матрица смежности или матрица Лапласа. Эти скрытые характеристики узлов не являются явно наблюдаемыми — они вычисляться путем оптимизации.
- Методы, основанные на содержании (Content-Based Methods), используют имеющиеся характеристики данных, связанные с узлами.

Рассмотрим некоторые примеры эвристических методов предсказывания связей. Обозначим x и y в качестве исходного и целевого узла,

между которыми можно спрогнозировать связь. Для обозначения множества соседей узла x будем использовать $\Gamma(x)$.

Первым и самым простым способом является метод общих соседей (common neighbors (CN)). Данный подход рассчитывает количество вершин, одновременно являющихся соседями двух узлов в качестве вероятности наличия между ними связи (1.5) (рис. 4).

$$f_{CN}(x, y) = |\Gamma(x) \cap \Gamma(y)| \quad (1.5)$$

CN широко используется в рекомендациях друзей в социальных сетях. Предполагается, что чем больше у двух людей общих друзей, тем больше вероятность, что они сами также являются друзьями.

Коэффициент Жаккара (Jaccard score) измеряет долю общих соседей (1.6).

$$f_{Jaccard}(x, y) = \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x) \cup \Gamma(y)|} \quad (1.6)$$

Метод предпочтительной привязанности (preferential attachment (PA)) использует произведение степеней узла[1] для измерения вероятности связи (1.7).

$$f_{PA}(x, y) = |\Gamma(x)| \cdot |\Gamma(y)| \quad (1.7)$$

PA подразумевает, что x с большей вероятностью будет связан с y, если y имеет высокую степень (рис. 4). Подобный подход применим в сетях цитирования - новая статья с большей вероятностью будет цитировать те статьи, которые уже имеют много упоминаний. Сети, сформированные с помощью механизма PA, называются свободными от масштабирования (scalefree networks) [20].

Существующие эвристические методы могут быть классифицированы на основе максимального количества переходов между соседними узлами, необходимого для вычисления оценки. CN, Jaccard и PA считаются методами первого порядка, так как они используют только самых ближайших соседей (с одним переходом для двух целевых узлов). Далее рассмотрим два метода

[1] Степень узла – количество ребер графа, инцидентных в конкретной вершине.

второго порядка.

Индекс Адамик-Адар (Adamic-Adar (AA)) [21] – метод, учитывающий веса общих соседей (1.8).

$$f_{AA}(x,y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{log|\Gamma(z)|} \tag{1.8}$$

Он определяется как сумма обратной логарифмической степени центральности соседей, разделяемых двумя узлами. Предполагается, что вершина с высокой степенью соединяется как с x, так и с y менее информативна, чем узел более низкой степенью (рис. 4).

Метод распределения ресурсов (Resource allocation (RA)) [22] использует более жесткий фактор снижения веса (1.9).

$$f_{RA}(x,y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{|\Gamma(z)|} \tag{1.9}$$

Он определяется как сумма обратной степени центральности соседей, разделяемых двумя узлами. Предполагается, что в отличие от подхода Адамик и Адара, метод RA еще больше будет отдавать предпочтения узлам с низкими степенями.

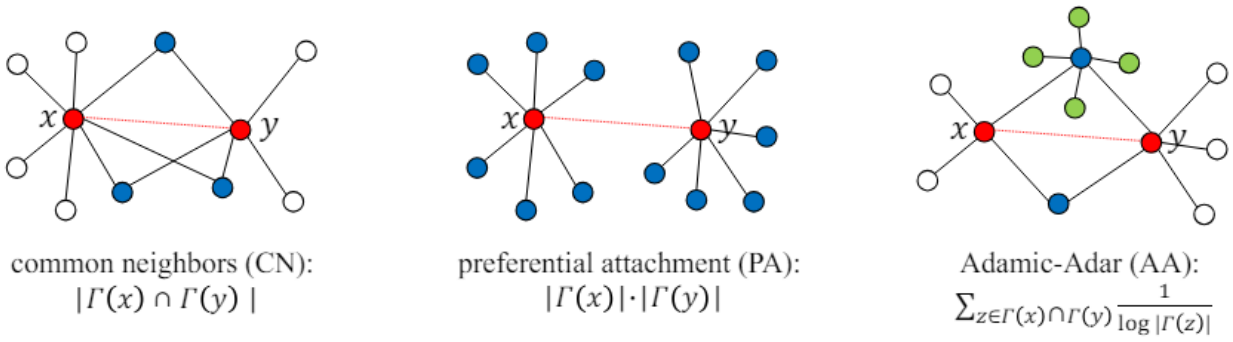

Рис. 4. Иллюстрация трех эвристических подходов в решении задачи предсказания связей: CN, PA и AA.

Способы первого и второго порядка являются локальными методами, поскольку все они высчитываться из локального подграфа ближайших соседей исходного и целевого узла, не учитывая при этом структуру всей сети. Существуют также методы, которые рассматривают всю сеть целиком. Такие подходы называться эвристическими методами высокого порядка (high-order

heuristics) и, как правило, являются более производительными, в отличие он низкоуровневых методов. К ним можно отнести индекс Каца (Katz index) [23], Rooted PageRank (RPR) [24] и SimRank (SR) [25].

Индекс Каца использует взвешенную сумму всех переходов между узлами x и y, где более продолжительный переход оценивается ниже (1.10).

$$f_{Katz}(x,y) = \sum_{l=1}^{\infty} \beta^l \left| walks^{\langle l \rangle}(x,y) \right| \tag{1.10}$$

Здесь β является убывающим коэффициентом между 0 и 1, а $\left| walks^{\langle l \rangle}(x,y) \right|$ подсчитывает l-длину переходов между x и y. Если же рассматривать переходы только длинной 2 ($l = 2$), то индекс Каца сводиться к методу CN.

Rooted PageRank является обобщением осинового алгоритма PageRank[2]. Сначала рассчитывается стационарное распределение π_x случайного блуждания[3], начиная с узла x, который двигается по текущим соседям с вероятностью α или возвращается в вершину x с вероятностью $1 - \alpha$. Далее π_x рассчитывается от узла y (обозначается как $[\pi_x]_y$), чтобы предсказать связь (x,y). Если же граф неориентированный, алгоритм симметрично повторяется относительно узла y (1.11).

$$f_{RPR}(x,y) = [\pi_x]_y + \left[\pi_y\right]_x \tag{1.11}$$

Оценка SimRank предполагает, что два узла схожи, если их соседи также схожи. Она рассчитывается рекурсивно (1.12):

$$\begin{aligned} &if\ x = y: && f_{SR}(x,y) = 1 \\ &otherwise: && f_{SR}(x,y) = \gamma \frac{\sum_{a \in \Gamma(x)} \sum_{b \in \Gamma(x)} f_{SR}(a,b)}{|\Gamma(x)| \cdot |\Gamma(y)|} \end{aligned} \tag{1.12}$$

где γ – константное значение между 0 и 1.

Помимо описанных выше эвристических методов существуют

[2] PageRank - один из алгоритмов ссылочного ранжирования. Алгоритм применяется к коллекции документов, связанных гиперссылками (таких, как веб-страницы из всемирной паутины), и назначает каждому из них некоторое численное значение, измеряющее его «важность» или «авторитетность» среди остальных документов.

[3] Случайное блуждение (random walk) - случайный процесс перехода между состояниями (в нашем случае - узлами сети / вершинами графа) определяемый матрицей перехода (row stochastic matrix), в которой фиксируется вероятность случайного перехода из узла в узел.

множество других подходов, описанных в работах [14, 26]. Эвристические методы можно рассматривать как вычисление предопределенных особенностей графовой структуры относительно расположения узлов и связей между ними. Несмотря на эффективность во многих областях, эти характеристики графовой структуры охватывают лишь небольшое подмножество всех возможных структурных шаблонов и не могут в полном объеме отобразить характеристики сети. Большинство эвристических методов работают только с однородными графами. Кроме того, эвристические методы хорошо работают только тогда, когда механизм формирования сети согласуется с эвристикой. Могут существовать сети со сложными механизмами формирования, на которых не получиться использовать ни один из методов.

Для того, чтобы в полной мере учесть все структурные особенности сети следует рассмотреть методы, учитывающие скрытые признаки графов (Latent-Feature Methods). Среди такого подхода к решению задачи предсказания связей можно выделить два основных метода: матричная факторизация (Matrix Factorization) и вычисление эмбеддингов графов (Network Embedding).

Матричная факторизация (MF) была рассмотрена в работах по рекомендательным системам [27, 28]. MF преобразует разряженную матрицу смежности A графа в произведение двух плотных матриц эмбеддиннгов Z. Исходные признаки матрицы (в нашем случае связи в графах) выражаются через латентные признаки линейным образом (1.13).

$$\hat{A}_{i,j} = z_i^T \cdot z_j \tag{1.13}$$

Затем происходит минимизация среднеквадратичной ошибки между создаваемой матрицей $\hat{A}_{i,j}$ и обычной матрицей смежности A по рассматриваемым связям для того, чтобы определить эмбеддинги (1.14).

$$\mathcal{L} = \frac{1}{\mathcal{E}} \sum_{(i,j)\in\mathcal{E}} \left(A_{i,j} - \hat{A}_{i,j}\right)^2 \tag{1.14}$$

Таким образом мы можем предсказать новые связи по матричному

произведению двух эмбеддингов интересующих узлов. Вариации матричной факторизации включают использование степеней в матрице A [29] или использовании матриц сходства узлов (Node similarity matrices) [30] для замены обычной матрицы смежности. Если мы заменим A на матрицу Лапласа L и определим потери следующим образом (1.15):

$$\mathcal{L} = \sum_{(i,j)\in\mathcal{E}} \left\| z_i - z_j \right\|_2^2, \tag{1.15}$$

Затем строится нетривиальное решение вышеприведенной задачи с использованием собственных векторов, соответствующих k наименьшим ненулевым собственным значениям L, которые восстанавливают метод лапласовой карты собственных значений (Laplacian eigenmap technique) [31] и решение для спектральной кластеризации (spectral clustering) [32].

Применения сетевых эмбеддингов (Network embedding) получили большую популярность в последнее время благодаря работе DeepWalk [33]. Подобные методы изучают низкоразмерные представления (эмбеддинги) для узлов, как правило основанные на обученной модели скип-грам[4] (skipgram model) [34] на основе последовательностей узлов, генерируемых случайным блужданием, так что узлы, которые часто появляются рядом друг с другом при случайном блуждании (т. е. узлы, расположенные близко в сети), будут иметь схожие эмбеддинги. Затем эти представления попарно объединяются для прогнозирования связей. В работе [35] также упоминается, что многие методы Network Embeddings (такие как LINE, DeepWalk и node2vec).

Таким образом, они также могут быть отнесены к методам со скрытыми признаками. Например, в DeepWalk приблизительно факторизует (1.16):

$$log\left(vol(G)\left(\frac{1}{w}\sum_{r=1}^{w}(D^{-1}A)^r\right)D^{-1}\right) - \log(b), \tag{1.16}$$

где:

- $vol(G)$ – сумма степеней узлов;

[4] Изначально скип-грам - один из методов обучения без учителя, который используется для поиска близких по тематике слов для заданного слова.

- D – диагональная степенная матрица;
- w – размерность окна скип-грам;
- b – некоторая константа.

Как мы можем видеть, DeepWalk, по сути, факторизует логарифм суммы некоторых нормализованных матриц смежности высокого порядка (вплоть до w). Это можно представить как случайное блуждание с расширением окрестности на w шагов, так что мы не только требуем, чтобы прямые соседи имели похожие эмбеддинги, но и требуем, чтобы узлы, находящиеся друг от друга через w шагов случайного блуждания, имели похожие вложения.

Аналогично, алгоритм LINE а своих формах второго порядка неявно факторизует (1.17):

$$log\left(vol(G)(D^{-1}AD^{-1})\right) - log(b). \qquad (1.17)$$

Другой популярный метод node2vec, который является улучшенной версией DeepWalk с добавлением негативного семплирования[5] и смещенным случайным блужданием[6], также неявно факторизует матрицу.

1.3.3 GNN методы предсказаний связей в графах

1.3 Анализ существующих моделей предсказания связей в графах

1.4 Вывод

[5] Негативное сэмплирование — это способ создать для обучения векторной модели отрицательные примеры, то есть показать ей пары узлов, которые не являются соседями по контексту.

[6] Смещенное случайное блуждание (biased random walk) – отличается от обычного случайного блуждания тем, что переменная с некоторой непостоянной вероятностью в некоторый момент может сменить свое текущее состояние на любое другое потенциальное состояние.

ГЛАВА 2. ПРОЕКТИРОВАНИЕ И РАЗРАБОТКА МОДЕЛИ ГРАФОВОЙ НЕЙРОННОЙ СЕТИ

ГЛАВА 3. АНАЛИЗ ПОЛУЧЕННЫХ РЕЗУЛЬТАТОВ

ЗАКЛЮЧЕНИЕ

СПИСОК ИСПОЛЬЗУЕМЫХ ИСТОЧНИКОВ

1. Fabian Beck, Michael Burch, Stephan Diehl, Daniel Weiskopf The State of the Art in Visualizing Dynamic Graphs // 2014
2. Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, Gabriele Monfardini The Graph Neural Network Model // 2009 https://ieeexplore.ieee.org/document/4700287
3. Hongyun Cai, Vincent W. Zheng, and Kevin Chen-Chuan Chang A Comprehensive Survey of Graph Embedding: Problems, Techniques and Applications, 2018. https://arxiv.org/pdf/1709.07604
4. Ritwik Raj Saxena, Ritcha Saxena Applying Graph Neural Networks in Pharmacology // 2024 https://www.researchgate.net/publication/378540568_Applying_Graph_Neural_Networks_in_Pharmacology
5. Aikta Aryaб, Pradumn Pandey, Akrati Saxena Node Classification Using Deep Learning in Social Networks // 2022 https://www.researchgate.net/publication/363652768_Node_Classification_Using_Deep_Learning_in_Social_Networks
6. Mohamed Badiy, Fatima Amounas, Ahmad El allaoui, Younes Bayane Neural Network for Link Prediction in Social Network // 2024 https://www.researchgate.net/publication/377803711_Neural_Network_for_Link_Prediction_in_Social_Network
7. Samarth Khanna, Sree Bhattacharyya, Sudipto Ghosh, Kushagra Agarwal, Asit Kumar Das Link Prediction for Social Networks using Representation Learning and Heuristic-based Features // 2024 https://arxiv.org/abs/2403.08613
8. Hoyeon Jeong, Young-Rae Cho, Jungsoo Gim, Seung-Kuy Cha, Maengsup Kim, Dae Ryong Kang GraphMHC: Neoantigen prediction model applying the graph neural network to molecular structure // 2024 https://www.researchgate.net/publication/379336536_GraphMHC_Neoantigen_prediction_model_applying_the_graph_neural_network_to_molecular_structu

re?_tp=eyJjb250ZXh0Ijp7ImZpcnN0UGFnZSI6InB1YmxpY2F0aW9uIiwicGFnZSI6InNlYXJjaCIsInBvc2l0aW9uIjoicGFnZUhlYWRlciJ9fQ

9. Wenfang Zhang, Hongwei Shi, Yang Yang, Yonghui Luo Research on the Classification of Aviation Safety Reports Based on Text and Knowledge Graph // 2020 https://iopscience.iop.org/article/10.1088/1742-6596/1646/1/012028
10. Ximing Li, Bing Wangm, Yang Wang, Meng Wang Graph-based Text Classification by Contrastive Learning with Text-level Graph Augmentation // 2023 https://www.researchgate.net/publication/376773608_Graph-based_Text_Classification_by_Contrastive_Learning_with_Text-level_Graph_Augmentation
11. Partha Basuchowdhuri, Satyaki Sikdar, Varsha Nagarajan, Khusbu Mishra, Surabhi Gupta, Subhashis Majumder Fast detection of community structures using graph traversal in social networks // 2019 https://www.researchgate.net/publication/318460944_Fast_Detection_of_Community_Structures_using_Graph_Traversal_in_Social_Networks?_tp=eyJjb250ZXh0Ijp7ImZpcnN0UGFnZSI6InB1YmxpY2F0aW9uIiwicGFnZSI6InNlYXJjaCIsInBvc2l0aW9uIjoicGFnZUhlYWRlciJ9fQ
12. Roman Sarkar, S M Sojib Ahamed, Md. Arman Hossain Possible Causes and Solutions of the Traffic Jam in Dhaka // 2024 https://www.researchgate.net/publication/373825265_Possible_Causes_and_Solutions_of_the_Traffic_Jam_in_Dhaka
13. Hao Tang, Ling Shao, Nicu Sebe, Luc Van Gool Graph Transformer GANs with Graph Masked Modeling for Architectural Layout Generation // 2024 https://www.researchgate.net/publication/377470350_Graph_Transformer_GANs_with_Graph_Masked_Modeling_for_Architectural_Layout_Generation
14. David Liben-Nowell, Jon Kleinberg The link-prediction problem for social networks // 2007 https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.20591
15. James Bennett, Stan Lanning The Netflix Prize // 2007 https://www.cs.uic.edu/~liub/KDD-cup-2007/NetflixPrize-description.pdf

16. Koppadi Bhavani, Kottu Aslesha, Lakshmi Sai Netflix Movies Recommendation System // 2024 https://www.researchgate.net/publication/379496999_Netflix_Movies_Recommendation_System
17. Mohammad Rezwanul Huq, Sanjeda Sara Jennifer, Shafiul Mahmud Partho, Fariha Fairuz A Comparative Study between Graph Database and Traditional Approach to forecast Coauthor Link Prediction based on Machine Learning Models // 2022 https://www.researchgate.net/publication/365676997_A_Comparative_Study_between_Graph_Database_and_Traditional_Approach_to_forecast_Coauthor_Link_Prediction_based_on_Machine_Learning_Models
18. https://docs.dgl.ai/guide/graph-heterogeneous.html#guide-graph-heterogeneous
19. David Liben-nowell, Jon Kleinberg The Link Prediction Problem for Social Networks // 2003 https://www.researchgate.net/publication/2930322_The_Link_Prediction_Problem_for_Social_Networks
20. Albert-Laszlo Barabasi, Reka Albert Emergence of Scaling in Random Networks // 1999 https://barabasi.com/f/67.pdf
21. Lada A Adamic, Eytan Adar Friends and neighbors on the Web // 2003 https://www.sciencedirect.com/science/article/abs/pii/S0378873303000091
22. Jing Zhou, Shung Jae Shin, Daniel Brass, Jaepil Choi, Zhi-Xue Zhang Social Networks, Personal Values, and Creativity: Evidence for Curvilinear and Interaction Effects // 2009 https://www.researchgate.net/publication/38092561_Social_Networks_Personal_Values_and_Creativity_Evidence_for_Curvilinear_and_Interaction_Effects
23. Leo Katz A new status index derived from sociometric analysis // 1953 https://link.springer.com/article/10.1007/BF02289026
24. Sergey Brin, Lawrence Page The anatomy of a large-scale hypertextual Web search engine // 1998 https://snap.stanford.edu/class/cs224w-readings/Brin98Anatomy.pdf

25. Glen Jeh Jennifer Widom Scaling Personalized Web Search // 2002 http://infolab.stanford.edu/~glenj/spws.pdf
26. Linyuan Lu, Tao Zhou Link Prediction in Complex Networks: A Survey // 2011 https://arxiv.org/pdf/1010.0725
27. Yehuda Koren, Robert Bell, Chris Volinsky Matrix Factorization techniques for recommender systems // 2009 https://datajobs.com/data-science-repo/Recommender-Systems-%5BNetflix%5D.pdf
28. Amr Ahmed, Nino Shervashidze, Shravan Narayanamurthy, Vanja Josifovski, Alexander J. Smola Distributed Large-scale Natural Graph Factorization // 2013
29. Catalina Cangea, Petar Velickovic, Nikola Jovanovic, Thomas Kipf, Pietro Liò Towards Sparse Hierarchical Graph Classifiers // 2018 https://arxiv.org/pdf/1811.01287
30. Mingdong Ou, Peng Cui, Jian Pei, Ziwei Zhang, Wenwu Zhu Asymmetric Transitivity Preserving Graph Embedding // 2016 https://www.kdd.org/kdd2016/papers/files/rfp0184-ouA.pdf
31. Mikhail Belkin, Partha Niyogi Laplacian Eigenmaps for Dimensionality Reduction and Data Representation // 2002 https://www2.imm.dtu.dk/projects/manifold/Papers/Laplacian.pdf
32. Ulrike von Luxburg A Tutorial on Spectral Clustering // 2007 https://arxiv.org/pdf/0711.0189
33. Bryan Perozzi, Rami Al-Rfou, Steven Skiena DeepWalk: Online Learning of Social Representations // 2014 https://arxiv.org/pdf/1403.6652
34. Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean Efficient Estimation of Word Representations in Vector Space // 2013 https://arxiv.org/pdf/1301.3781
35. Jiezhong Qiu, Yuxiao Dong, Hao Ma, Jian Li, Kuansan Wang, Jie Tang Network Embedding as Matrix Factorization: Unifying DeepWalk, LINE, PTE, and node2vec // 2018 https://arxiv.org/pdf/1710.02971
36.

ПРИЛОЖЕНИЕ