

УДК 004.89

doi: 10.15622/rcai.2025.085

ТЕНЗОРНОЕ ПРЕДСТАВЛЕНИЕ НЕОДНОРОДНЫХ СЕМАНТИЧЕСКИХ СЕТЕЙ ДЛЯ РЕШЕНИЯ ЗАДАЧИ СЛИЯНИЯ БАЗ ЗНАНИЙ

А.И. Пальчевский (*apalchev@gmail.com*)

А.И. Молодченков (*aim@tesyan.ru*)

Федеральный исследовательский центр
«Информатика и управление» РАН, Москва

В работе рассмотрен способ преобразования баз знаний, построенный на основе неоднородных семантических сетей, в семейство тензорных структур. Введена метрика оценки близости узлов базы знаний. Преобразование базы знаний в семейство тензоров является одним из этапов разработки алгоритма автоматизации слияния баз знаний. Предложенный подход может быть применен и к онтологиям.

Ключевые слова: тензоры, неоднородные семантические сети, машинное обучение, КИИ-2025.

Введение

Задача слияния графовых структур, к которым относятся семантические и неоднородные семантические сети, является не новой, но актуальной в настоящее время. Так, например, когда несколько экспертов работают над одной базой знаний или необходимо объединить несколько онтологий в одну.

Существует множество подходов к решению данной задачи, варьирующихся от простых эвристических правил до сложных методов машинного обучения. Одним из перспективных направлений является использование графовых нейронных сетей (GNN) [Li, 2023], позволяющих обучать модели, непосредственно работающие с графовой структурой данных. Для эффективного обучения GNN необходимо разработать способы представления графов в виде векторных или тензорных эмбедингов. Среди существующих методов генерации таких эмбедингов можно выделить несколько ключевых направлений:

1. CP-разложение (CANDECOMP/PARAFAC) [Faber, 2003], [Yang, 2021], [Zhoubao, 2003], [Геяа Hahn, 1997] представляет собой тензорное разложение высокого порядка, позволяющее декомпозировать граф на латентные компоненты. CP-разложение обладает относительно простой реализацией и интерпретируемостью, что делает его привлекательным для анализа графов с простой структурой. Однако, его эффективность может снижаться при работе со сложными графовыми структурами, содержащими большое количество связей и иерархических отношений.

2. Knowledge-Enriched Tensor Factorization [Padiа, 2019] представляет собой расширение методов тензорного разложения, позволяющее интегрировать внешние знания, такие как метаданные и семантическая информация, в процесс разложения. Интеграция внешних знаний позволяет улучшить качество получаемых эмбедингов и повысить точность анализа графов. Однако, эффективность этого подхода напрямую зависит от качества и доступности внешних знаний, а также от сложности их интеграции в тензорную модель.

3. Message Passing (MP) [Bordes, 2013] является основой многих современных GNN и позволяет узлам графа обмениваться информацией с соседними узлами, учитывая как локальные, так и глобальные зависимости в графе. Алгоритмы, основанные на этом подходе, часто превосходят стандартные методы матричного разложения и алгоритмы семейства TransX по качеству, особенно при наличии сложных зависимостей между узлами и небольшом объеме данных для обучения. Однако, MP требует значительных вычислительных ресурсов и чувствителен к качеству интеграции знаний, требуя преобразования текстовой или другой неструктурированной информации в матричный вид.

Несмотря на существенный прогресс в разработке методов слияния графовых структур, большинство существующих подходов ориентированы на работу с однородными графами, где все узлы и ребра имеют одинаковый тип и свойства. Однако, есть задачи, где часто встречаются неоднородные графы, содержащие узлы и ребра различных типов, характеризующиеся разными свойствами и семантикой. В данной работе представлен новый подход к преобразованию графовых структур в семейство тензоров, который подходит для работы с разнородной структурой узлов и свойств. Кроме этого, в работе представлена метрика оценки близости таких узлов.

1. Преобразование семантической сети в семейство тензоров

Неоднородная семантическая сеть – семейство графов, имеющих общее множество вершин; вершинам сопоставлены объекты моделируемой действительности, ребрам элементы некоторых бинарных отношений на множестве вершин; им же сопоставлены процедуры, предназначенные для

проверки корректности сети и порождения различного рода гипотез, повышающих эффективность процесса построения сети [Осипов, 2015]. В данной главе представлен метод преобразования неоднородных семантических сетей (НСС) в структуру, пригодную для обработки методами тензорной алгебры, обучения графовых, и не только, нейронных сетей и разработки алгоритмов слияния графовых структур. Ключевой особенностью подхода является представление НСС в виде семейства разделенных тензоров, позволяющих эффективно кодировать разнородную информацию, содержащуюся в структуре графа. В отличие от традиционных методов, использующих единый тензор для представления графа, наш подход разделяет информацию о связях, атрибутах узлов, текстовых дескрипторах и типах узлов на отдельные тензорные компоненты, что обеспечивает большую гибкость и выразительность представления.

1.1. Структура тензоров

Предложенная структура разделенных тензоров включает в себя три основных компонента:

1. Тензор связей – $A \in \mathbb{R}^{N \times N}$ – кодирует направленные связи между узлами в графе. N представляет собой общее количество узлов в сети. Элемент A_{ij} представляет собой значение, ассоциированное с направленной связью типа k от узла i к узлу j .

Например, $A_{3,5} = 0$ означает наличие связи типа "всегда наблюдается" от узла с идентификатором 3 к узлу 5. Соответственно связь $A_{5,3} = 1$ обозначает, обратную связь от узла с идентификатором 5 к узлу 3 с типом связи «Может наблюдаться».

2. Тензор узлов – $X \in \mathbb{R}^{N \times F}$. Этот тензор кодирует атрибуты, характеризующие каждый узел в графе. N представляет количество узлов, а F – количество постоянных свойств, описывающих каждый узел. Для нашего примера $F = 3$ (название, тип узла, описание). Тензор узлов может включать в себя как числовые, так и категориальные свойства. Числовые свойства нормализуются для приведения к единому масштабу.

Категориальные свойства кодируются с использованием методов one-hot или multi-hot кодирования. Например, $X_0 = [0.3, 1, 0, 1]$ может обозначать узел с численным атрибутом, имеющим значение 0.3, и дополнительными категориальными свойствами, представленными вектором $[1, 0, 1]$. Дополнительно, в нашем подходе мы используем расширенный набор признаков, включая статистические показатели, полученные из окрестности узла в графе, такие как центральность по степени, посредничеству и близости, что позволяет более полно охарактеризовать роль и положение узла в сети.

3. Тензор названий узлов – $E \in \mathbb{R}^{N \times D}$. Этот тензор кодирует семантическое значение названий узлов. N представляет количество узлов, а D – размерность векторного представления (эмбединга). Для генерации

эмбедингов используются предварительно обученные NLP-модели, такие как BERT, Word2Vec или GloVe. Использование NLP-моделей позволяет учитывать контекст и семантические связи между названиями узлов, что повышает качество представления графа. Мы также рассматриваем возможность использования трансформерных архитектур для генерации более контекстуально-зависимых эмбедингов.

1.2. Метод преобразования НСС в семейство тензоров

Рассмотрим пример структуры подсети НСС (рис. 1).

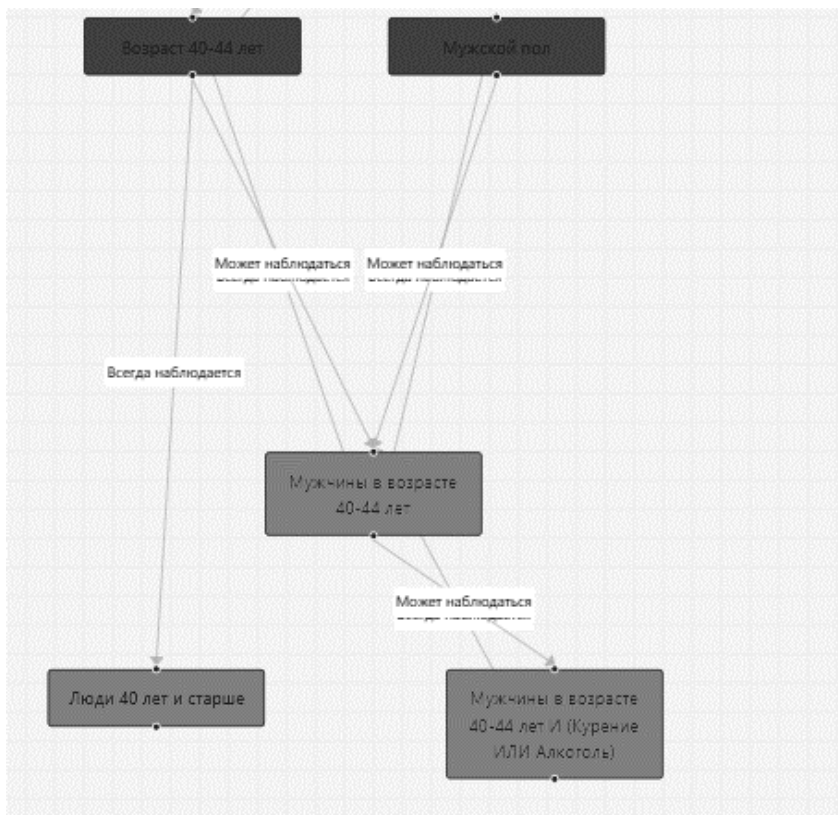


Рис. 1. Пример подсети. С тремя типами узлов

Метод состоит из следующих этапов.

На первом этапе создается матрица узлов. Название узла векторизуется, а затем связывается с соответствующим полем в матрице. Записываются значения типа узла и векторизованное описание.

Например, (табл. 1).

Таблица 1

Индекс	Название	Тип узла	Векторное представление названия	Векторное представление описания
0	Мужчины в возрасте 40-44 лет	0	[векторизованный текст]	[векторизованный текст]
1	Возраст 40-44 лет	1	[векторизованный текст]	[векторизованный текст]

На втором этапе преобразование динамических свойств узла в тензор свойств. Например,

$$e \quad e \quad d \quad e \quad e \quad (1.1)$$

На третьем этапе строится матрица связей узлов с их динамическими свойствами. Эта матрица связывает индексы узлов и динамических свойств. В качестве значений выступают области значений этих свойств для узла. В примере выше свойства Пол, Возраст, Артериальное давление имеют индексы 0, 1, 2 соответственно. У нас имеются узлы «Мужской пол», «Возраст 40-44» которые имеют индексы в своей матрице 0, 1. Тогда для связи узла «Мужской пол» со свойством будет в матрице будет записана следующая строка

$$d \quad e \quad e \quad (1.2)$$

На четвертом этапе строится матрица связей, где указаны типы отношений между узлами. Например, создаем матрицу размерностью $[N \times M]$, где N – это номер узла.

$$(1.3)$$

где

а) Узлы:

- «Возраст 40-44 лет» – индекс 0;
- «Мужской пол» – индекс 1;
- «Мужчины в возрасте 40-44» – индекс 2;

б) Связи и их типы кодируются значениями в промежутке от 0 до 4 включительно, где 1 – Всегда наблюдается, 2 – Может наблюдаться, 3 – Может отсутствовать, 4 – всегда отсутствует. Соответственно матрица читается как: От узла 0 связь «Всегда наблюдается» к узлу 2 и от узла 3 связь «Может наблюдаться» к узлу 0.

2. Метрики близости для оценки близости узлов с гетерогенными свойствами

В задачах слияния графовых структур, особенно в контексте неоднородных семантических сетей (НСС), критически важным является точная оценка близости между узлами, характеризующимися гетерогенными свойствами. Традиционные метрики, такие как евклидово расстояние или косинусное сходство, часто оказываются неэффективными при сравнении узлов с разнородными атрибутами. В данной главе мы представляем взвешенную комбинированную метрику, специально предназначенную для оценки близости узлов в НСС, учитывающую гетерогенность свойств и структурные особенности графа.

Предлагаемый подход состоит в объединении нескольких специализированных метрик, каждая из которых предназначена для оценки сходства в определенном векторном пространстве, с последующим взвешенным агрегированием результатов. Это позволяет эффективно обрабатывать разнородные атрибуты узлов, такие как числовые значения, категориальные признаки, текстовые описания и интервальные данные, а также учитывать структурную информацию о графе.

Общая формула для оценки близости между узлами N и K выглядит следующим образом:

(2.1)

где

1. – косинусное сходство векторных представлений имен свойств.
2. – метрика близости гетерогенных скалярных значений.
3. – категориальная метрика близости для свойств типа интервальная, качественная и т.д.
4. – комбинированная метрика (перекрытие + расстояние).
5. – адаптивные веса для корректировки важности признака.

Структурные свойства узлов:

1. – Косинусное сходство векторных представлений имен узлов.
2. – функция, которая используется как бинарный множитель, зависящий от типа узла где 1 это совпадение типа узла и 0 для разных типов узлов.

Рассмотрим более подробно каждый компонент данной формулы.

Для оценки близости названий можно использовать косинусное сходство:

$$\text{sim}(u, v) = \frac{\langle \mathbf{u}, \mathbf{v} \rangle}{\|\mathbf{u}\| \|\mathbf{v}\|} \quad (2.2)$$

Для интервальных значений можно применить комбинированную метрику, которая будет учитывать, как перекрытие, так и расстояние интервалов:

$$\text{sim}(u, v) = \frac{\text{overlap}(u, v) + \text{dist}(u, v)}{\text{overlap}(u, v) + \text{dist}(u, v) + 1} \quad (2.3)$$

где w – вес перекрытия.

Для работы со скалярными значениями нужно ввести универсальную метрику, учитывающую, что скалярное значение может быть нескольких видов, например, непрерывная величина, категориальное значение или порядковая величина.

Формально метрику можно представить в виде:

$$(2.4)$$

Обоснование создания новой метрики для разных типов узлов заключается в том, что традиционные метрики не учитывают динамическую размерность и гетерогенность признаков. Поэтому предложено решение: каждое векторное пространство признаков \mathbf{X}_i – обрабатывается отдельно специализированной метрикой, уменьшая размерность итогового пространства к 2 размерностям. Веса w_i отражают относительную важность одной конкретной метрики признаков в домене. Например для признаков из разных векторных пространств \mathbf{X}_i .

Оптимизация w_i проводится на этапе экспериментов. Веса изначально инициализируются как $w_i = 1$. При подборе значений весов можно использовать L2 регуляризацию для предотвращения перекоса в пользу одного признака. Также использование относительных весов дает устойчивость к неполноте данных. Например: если некоторый \mathbf{X}_i отсутствует, то соответствующий ему w_i будет равен 0.

Структурные модуляторы мы можем использовать для корректировки общей относительной близости узла. $\text{sim}(u, v)$ – использует контекстные эмбединги (Bert или Sentence-BERT), учитывающие близость имён. Это позволяет нам корректировать итоговое значение близости двух узлов при совпадении их типов и близости динамических свойств.

Функция $\text{type_sim}(t_u, t_v)$ – вводит штраф за несовпадение типов. Мы берем упрощённую модель, где при разных типах мы можем сказать, что даже при совпадении свойств узлы разные.

Все , нормированы и , где 1 – это полное совпадение узлов. Базовая настройка порогового значения . При достижении порогового значения мы можем сказать, что узлы N и K являются одним узлом и мы можем провести операцию слияния.

Заключение

В настоящей работе был представлен подход к решению задачи преобразования графовых структур в семейство тензоров, ориентированный на работу с неоднородной структурой графов и семантических сетей.

Преобразование НСС в семейство тензоров позволило эффективно кодировать разнородную информацию, содержащуюся в базах знаний, путем разделения ее на отдельные тензорные компоненты, соответствующие связям, атрибутам узлов, текстовым дескрипторам и типам узлов.

Взвешенная комбинированная метрика, разработанная для оценки близости узлов, учитывает гетерогенность свойств и структурные особенности графа.

В дальнейших исследованиях планируется разработка алгоритма результаты для создания итеративного алгоритма слияния, который будет автоматически определять наиболее близкие узлы на основе разработанной метрики и выполнять их слияние с учетом структуры графа, а также планируется проведение экспериментальных исследований.

Список литературы

- [Осипов, 2015] Осипов Г.С. Методы искусственного интеллекта. – 2-е изд. – М.: ФИЗМАТЛИТ, 2015. – 296 с.
- [Abadal, 2021] Abadal, Sergi, et al. Computing graph neural networks: A survey from algorithms to accelerators // ACM Computing Surveys (CSUR). – 2021. – 54.9. – P. 1-38. – doi:10.1145/3477141.
- [Bordes, 2013] Bordes A., Usunier N., Garcia-Durán A., Weston J., & Yakhnenko O. Translating Embeddings for Modeling Multi-relational Data // Advances in Neural Information Processing Systems (NeurIPS). – Dec. 2013. – Vol. 26. – P. 2787-95. – <https://proceedings.neurips.cc/paper/2013/file/1cecc7a77928ca8133fa24680a88d2f9-Paper.pdf>.
- [Li, 2023] Li C., Wu S., Chen T., Wang R. and Cao J. Knowledge Fusion Algorithm Based on Entity Relation Mining and Graph Neural Network // 5th International Conference on Frontiers Technology of Information and Computer (ICFTIC), Qiangdao, China, 2023. – P. 413-417. – doi: 10.1109/ICFTIC59930.2023.10456136.
- [Faber, 2003] Faber, Nicolaas Klaas M., Rasmus Bro, and Philip K. Hopke. Recent developments in CANDECOMP/PARAFAC algorithms: a critical review // Chemometrics and Intelligent Laboratory Systems. – 2003. – 65.1. – P. 119-137. – doi: 10.1016/S0169-7439(02)00089-8.

- [**Geña Hahn, 1997**] Geña Hahn, Gert Sabidussi. Graph symmetry: algebraic methods and applications. – Springer, 1997. – Vol. 497. – P. 116. – (NATO Advanced Science Institutes Series). – ISBN 978-0-7923-4668-5.
- [**Padia, 2019**] Padia, Ankur, et al. Knowledge graph fact prediction via knowledge-enriched tensor factorization // Journal of Web Semantics. – 2019. – 59. – P. 100497. – doi: <https://doi.org/10.1016/j.websem.2019.01.004>.
- [**Yang, 2021**] Yang Han, and Junfei Liu. Knowledge graph representation learning as groupoid: unifying TransE, RotatE, QuatE, ComplEx // Proceedings of the 30th ACM international conference on information & knowledge management. – P. 2311-2320. – doi: 10.1145/3459637.3482442.
- [**Zhoubao, 2003**] Zhoubao Sun, Xiaodong Zhang, Haoyuan Li, Yan Xiao, Haifeng Guo. Recommender Systems Based on Tensor Decomposition // Tech Science Press Computers, Materials & Continua January 2020. – 66(1). – P. 621-630. – DOI: 10.32604/cmc.2020.012593.