УДК 004.655.3, 004.652.3

Использование больших языковых моделей для формирования запросов к графовым базам данных

Р.А. Мамадалиев
Научный руководитель: доцент, к.ф-м.н. М.Е. Семенов
Национальный исследовательский Томский политехнический университет,
Россия, г. Томск, пр. Ленина, 30, 634050
E-mail: ram17@tpu.ru

Using large language models for a graph database interaction

R.A. Mamadaliev
Scientific Supervisor: Ass. Prof., PhD M.E. Semenov
Tomsk Polytechnic University, Russia, Tomsk, Lenin str., 30, 634050
E-mail: ram17@tpu.ru

Abstract. *This study investigates the application of large language models for formulating queries to graph databases. A knowledge graph was constructed based on data pertaining to users within a social network community. An analysis and comparison were conducted between queries generated using a large language model and those composed manually.*
Key words: *graph database, large language models, Neo4j.*

Введение

На сегодняшний день графовые системы управления базами данных вышли на первый план в области управления данными, предлагая уникальный подход к визуализации и запросам взаимосвязанных данных. В отличие от традиционных реляционных баз данных, графовые способны обрабатывать сложные отношения между данными, что в свою очередь позволяет представлять данные в виде графа знаний – структуры данных, где объекты или понятия представлены как узлы, а связи между ними – как ребра.

Одним из инновационных подходов к извлечению данных из графа знаний является использование больших языковых моделей (large language model, LLM) – моделей обработки естественного языка используемых для выполнения различных задач, таких как генерация и понимание текста. Преимуществом данного подхода является возможность написания запроса без знания языка обращения к конкретной базе данных.

Целью данной работы является исследование эффективности использования больших языковых моделей в формировании запросов к графу знаний.

Экспериментальная часть

В работе в качестве исходных данных для построения графа знаний мы использовали информацию о пользователях сообщества «Типичный ТПУ» (7500 подписчиков) [1]. Для реализации алгоритмов получения данных был использован язык программирования Python. В качестве графоориентированной системы управления базы данных (СУБД) мы выбрали Neo4j [2, 3]. Для написания запросов и визуализации данных применялась встроенная утилита Neodash. В качестве большой языковой модели использовалась модель – OpenAI [4].

Принцип работы формирования запросов с помощью LLM заключается в преобразовании инструкций для модели, изложенных на естественном языке (промпта) в запрос, написанный на языке запросов конкретной СУБД [5]. В нашем случае был использован язык Cypher. Иллюстративный пример формирования запроса с использованием LLM приведена на рис. 1.

Рис 1. Иллюстративный пример формирования запроса с использованием LLM

Принцип работы самой модели можно описать последовательностью этапов. Модель LLM загружает предварительно обученные параметры и конфигурации для обработки естественного языка и генерации запросов к графовой базе данных. Инициализация модели подготавливает ее к обработке входного текста и настройке внутренних состояний для работы с графовой структурой. Входной запрос на естественном языке токенизируется и преобразуется в числовое представление с использованием векторного пространства слов. Каждый токен кодируется в числовое представление, которое модель LLM может понимать и обрабатывать. Закодированный запрос подается на вход модели LLM для обработки. Модель анализирует контекст и синтаксические зависимости в запросе, используя свою архитектуру, в частности, модель Transformer, состоящую из двух компонентов – кодировщика и декодировщика.

Кодировщик принимает на вход последовательность токенов и преобразует их в скрытые представления. Важным составляющим является механизм самовнимания который позволяет модели фокусироваться на значимых элементах последовательности. Декодировщик принимает на вход скрытое представление и генерирует выходную последовательность. Декодировщик также содержит механизм самовнимания, который позволяет модели учесть ранее сгенерированные токены.

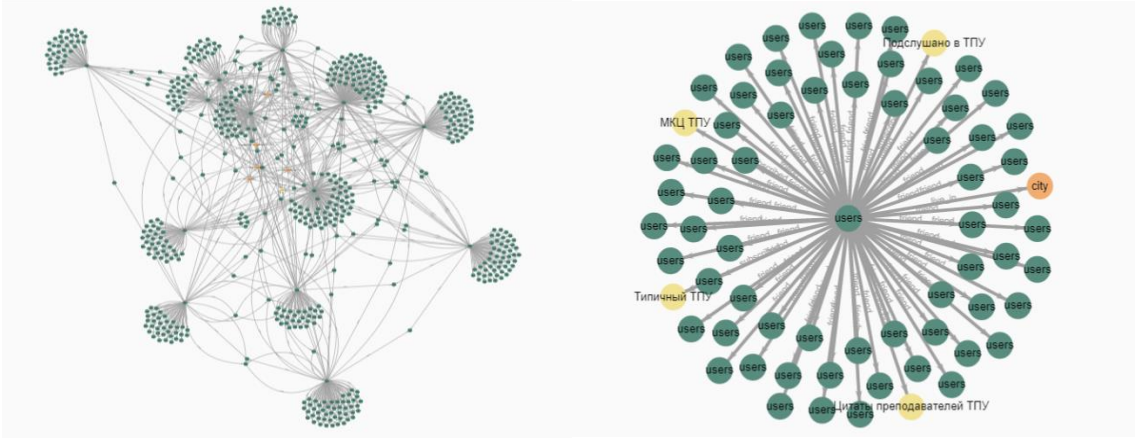
В процессе обработки модель извлекает ключевые элементы запроса и определяет его цель, например, поиск конкретных узлов или связей. Генерация происходит путем предсказания последовательности токенов, которые образуют запрос. Сгенерированные токены объединяются в итоговый запрос на языке Cypher.

Результаты

В результате исследования сформирован граф знаний сообщества социальной сети [1], проведен анализ и обработка данных на основе графового представления знаний сообщества (рис. 2). Также были исследованы основные характеристики графа знаний, вычислен эффективный диаметр графа.

Рис. 2. Граф знаний сообщества: пользователи сообщества (синие маркеры), пять самых популярных городов среди пользователей (зеленые маркеры), пять самых популярных сообществ среди пользователей (желтые маркеры)

В рамках работы были проанализированы возможности LLM OpenAI для формирования запросов к графовой базе данных Neo4j. Проведены сравнение и анализ результатов написания запросов с помощью LLM OpenAI и традиционным подходом. Использованная языковая модель справляется с простыми запросами, более сложные запросы (в частности, с использованием методов топологического анализа данных) требуют верификации сгенерированного кода человеком. На рис. 2 приведен результат выполнения запроса: «show all users who live in Tomsk and have more than 50 friends and how they are connected to each other», на рис. 3 приведено визуализация результата запроса: «show the user with the highest betweenness rate and all his connections using graph data science function».



а *б*

Рис. 3. Результаты выполнения запросов, сформированных с использованием LLM модели: а – запрос без вычислений, б – запрос с вычислением характеристик графа

Таким образом, большие языковые модели способны формулировать как простые, так и сложные многоуровневые запросы с использованием методов графового анализа данных, что позволяет существенно сократить время написания запроса и его тестирования.

Заключение

Проведённое исследование показывает, что использование больших языковых моделей в формировании запросов к базе данных значительно проще и эффективнее по сравнению с традиционным (ручным) формированием запросов. Одним из ключевых преимуществ исследуемого подхода является возможность формирования запросов без знаний языка написания запросов. Это открывает новые возможности в сфере обработки и структуризации данных и делает процесс более интуитивным.

Список литературы

1. Сообщество «Типичный ТПУ» // URL: https://vk.com/tipicaltpu / (дата обращения: 08.02.2024).
2. Ломов П.А. Применение графовых СУБД в задачах анализа данных // Труды Кольского научного центра РАН. – 2019. – № 9 (9). – С. 137–145.
3. Matasova E.A., Sabinin O.Yu. Research of efficiency of Oracle and Neo4j DBMS // Theoretical & Applied Science. – 2022. – № 5(109). – P. 742–752.
4. Михайлова С.С., Халмакшинов Е.А. Алгоритм анализа данных на графовых структурах /С.С. Михайлова, // Наука и бизнес: пути развития. – 2022. – № 4(130). – С. 25-28.
5. Объединение LLM и графов знаний для GenAI: примеры использования и лучшие практики // neo4j.com URL: https://neo4j.com/blog/unifying-llm-knowledge-graph/ (дата обращения: 12.02.2024).