

Основы GraphRAG

Улучшенный RAG на базе графов знаний

ТОМАЖ БРАТАНИЧ, ОСКАР ХАНЕ
ПРЕДИСЛОВИЕ ПАКО НАТАНА



Санкт-Петербург • Москва • Минск

2026

Краткое содержание

Предисловие	10
Введение.....	12
От издательства	20
Глава 1. Повышение точности больших языковых моделей.....	21
Глава 2. Поиск по векторному сходству и гибридный поиск.....	39
Глава 3. Улучшенные стратегии поиска по векторному сходству	54
Глава 4. Генерирование запросов на Cypher из вопросов на естественном языке.....	70
Глава 5. Агентная система RAG.....	82
Глава 6. Построение графов знаний с помощью LLM.....	97
Глава 7. Реализация GraphRAG от Microsoft	117
Глава 8. Оценка RAG-приложений.....	148
Приложение. Среда Neo4j	162
Дополнительные ресурсы	187

Оглавление

Предисловие	10
Введение.....	12
О книге	13
О коде	15
Благодарности.....	16
Об авторах	18
Иллюстрация на обложке	19
От издательства	20
О научном редакторе русскоязычного издания	20
Глава 1. Повышение точности больших языковых моделей.....	21
1.1. Введение в LLM.....	22
1.2. Ограничения LLM	26
1.2.1. Проблема отсечки знаний	26
1.2.2. Устаревшая информация.....	26
1.2.3. Галлюцинации	27
1.2.4. Отсутствие приватной информации	28
1.3. Преодоление ограничений LLM	30
1.3.1. Дообучение с учителем	30
1.3.2. Механизм RAG.....	31
1.4. Графы знаний как хранилище данных для RAG-приложений	35
Резюме	37

Глава 2. Поиск по векторному сходству и гибридный поиск	39
2.1. Компоненты архитектуры RAG	40
2.1.1. Ретривер	40
2.1.2. Генератор	42
2.2. RAG с использованием поиска по векторному сходству	43
2.2.1. Подготовка данных приложения	44
2.2.2. Корпус текстов	44
2.2.3. Чанкинг текста	44
2.2.4. Эмбединговая модель	46
2.2.5. База данных с возможностью поиска по векторному сходству	46
2.2.6. Выполнение векторного поиска	48
2.2.7. Генерирование ответа с помощью LLM	49
2.3. Добавление полнотекстового поиска в RAG-приложение для включения гибридного поиска	50
2.3.1. Индекс для полнотекстового поиска	50
2.3.2. Выполнение гибридного поиска	51
2.4. Заключительные мысли	52
Резюме	53
Глава 3. Улучшенные стратегии поиска по векторному сходству	54
3.1. Step-back-промптинг	59
3.2. Ретривер родительских документов	61
3.2.1. Извлечение данных через родительские документы	66
3.3. Полный пайплайн RAG	67
Резюме	69
Глава 4. Генерирование запросов на Cypher из вопросов на естественном языке	70
4.1. Основы генерирования запросов	71
4.2. Место генерирования запросов на Cypher в RAG-пайплайне	72
4.3. Полезные практики для генерирования запросов на Cypher	73
4.3.1. Использование few-shot-примеров для обучения в контексте	73
4.3.2. Использование схемы базы данных в промпте, чтобы показать LLM структуру графа знаний	74
4.3.3. Добавление терминологических соответствий для семантического отображения пользовательского запроса на схему	76
4.3.4. Инструкции по форматированию	77
4.4. Реализация генератора text2cypher на базовой модели	77

8 Оглавление

4.5. Специализированные (дообученные) LLM для text2cypher	80
4.6. Возможности, которые открывает text2cypher	80
Резюме	81
Глава 5. Агентная система RAG	82
5.1. Что такое агентная RAG	82
5.1.1. Агенты-ретриверы	84
5.1.2. Маршрутизатор ретриверов	84
5.1.3. Критик ответов	85
5.2. Зачем нужна агентная RAG	85
5.3. Как реализовать агентную RAG	86
5.3.1. Реализация ретриверов	86
5.3.2. Реализация маршрутизатора ретриверов	89
5.3.3. Реализация критика ответов	93
5.3.4. Объединение всех компонентов	95
Резюме	96
Глава 6. Построение графов знаний с помощью LLM	97
6.1. Извлечение структурированных данных из текста	98
6.1.1. Определение модели структурированных выходных данных	101
6.1.2. Запрос на извлечение структурированных выходных данных	106
6.1.3. Набор данных CUAD	107
6.2. Построение графа	109
6.2.1. Импорт данных	110
6.2.2. Разрешение сущностей	113
6.2.3. Добавление неструктурированных данных в граф	114
Резюме	116
Глава 7. Реализация GraphRAG от Microsoft	117
7.1. Выбор набора данных	118
7.2. Индексирование графа	119
7.2.1. Чанкинг	120
7.2.2. Извлечение сущностей и связей	121
7.2.3. Суммаризация сущностей и связей	126
7.2.4. Обнаружение сообществ и суммаризация	130
7.3. Графовые ретриверы	134
7.3.1. Глобальный поиск	134
7.3.2. Локальный поиск	140
Резюме	146

Глава 8. Оценка RAG-приложений	148
8.1. Проектирование набора данных для бенчмарка	150
8.1.1. Создание тестовых примеров.....	151
8.2. Метрики оценки	154
8.2.1. Полнота контекста	154
8.2.2. Достоверность ответа.....	155
8.2.3. Правильность ответа.....	155
8.3. Выполнение оценки.....	156
8.3.1. Загрузка набора данных.....	156
8.3.2. Запуск оценки.....	157
8.3.3. Наблюдения	158
8.4. Дальнейшие действия.....	160
Резюме	160
Приложение. Среда Neo4j	162
П.1. Язык запросов Cypher	163
П.2. Установка Neo4j.....	164
П.2.1. Установка Neo4j Desktop.....	164
П.2.2. Установка Neo4j через Docker	168
П.2.3. Установка Neo4j Aura	169
П.3. Настройка Neo4j Browser	169
П.4. Набор данных Movies	169
П.4.1. Загрузка через Neo4j Query Guide	170
П.4.2. Использование онлайн-версии	170
П.4.3. Загрузка с помощью Cypher	170
Дополнительные ресурсы	187