УДК 004.82
DOI 10.18523/2617-3808.2025.8.84-87

Андрощук М. В.

ВПЛИВ МЕТОДІВ ДОБУВАННЯ ЗНАНЬ НА ЕФЕКТИВНІСТЬ RAG-СИСТЕМ НА ОСНОВІ ГРАФІВ

Стаття досліджує, як методи добування знань впливають на ефективність RAG-систем, що використовують графи знань. Вона показує, що якість графа знань, сформованого різними методами добування знань, є ключовою для подолання обмежень великих мовних моделей (LLM), таких як «галюцинації». Робота аналізує архітектури LightRAG і GraphRAG та підкреслює, що вибір оптимальної KE-стратегії залежить від конкретних завдань і предметної області.

Ключові слова: RAG, графи знань, добування знань, ВММ.

Вступ

Останні роки демонструють прорив у галузі штучного інтелекту, зокрема завдяки стрімкому розвитку великих мовних моделей (LLMs). Моделі, такі як GPT-4.1, Claude Opus 4, LLaMA 4 та інші, генерують текст, який важко відрізнити від написаного людиною, перекладають тексти різними мовами, відповідають на запитання та виконують широкий спектр завдань, пов'язаних з обробленням природної мови. Їхня архітектура, що базується переважно на трансформерах, і навчання на величезних масивах текстових даних дозволили досягти безпрецедентної продуктивності. Однак, попри значні успіхи, ВММ мають низку фундаментальних обмежень, які стримують їхнє надійне застосування у критично важливих сферах, де важливі точність і пояснення результатів.

Одним із найсуттєвіших недоліків є схильність ВММ до так званих галюцинацій — генерування правдоподібної, але фактично неправильної або безглуздої інформації. Це явище пов'язане з тим, що ВММ, по суті, є статистичними моделями, які навчаються передбачати наступне слово в послідовності, не маючи справжнього розуміння світу чи доступу до актуальної фактичної бази даних. Другим важливим обмеженням є проблема «застарілих знань». Знання ВММ фіксуються на момент завершення їхнього тренування, і вони не можуть динамічно оновлюватися новою інформацією, що з'являється у світі. Це робить їх менш надійними для завдань, що потребують актуальних даних. Крім того, часто бракує прозорості у процесі прийняття рішень ВММ, що ускладнює верифікацію та довіру до їхніх відповідей [4].

Для розв'язку проблеми актуальності й точності створено системи генерації з доповненням пошуком, особливо на основі графів знань (KG-RAG). Та попри це залишається недостатньо дослідженою фундаментальна залежність ефективності таких систем від конкретних методів і стратегій, застосованих для добування знань та побудови самого графа. Необхідне глибоке розуміння того, як різні підходи до побудови та збагачення графів знань впливають на якість відповідей і загальну продуктивність таких систем.

Генерація з доповненням пошуком

Для подолання зазначених обмежень ВММ було запропоновано концепцію генерації з доповненням пошуком (Retrieval-Augmented Generation, RAG) [2]. RAG — це архітектурний підхід, який поєднує потужності генеративних моделей ВММ із зовнішньою базою знань. Замість того, щоб покладатися виключно на параметричні знання, засвоєні під час тренування, RAG-система спочатку здійснює пошук релевантної інформації із зовнішнього джерела у відповідь на запит користувача. Потім ця добута інформація передається ВММ разом із початковим запитом як контекст, на основі якого модель генерує остаточну відповідь.

Такий підхід має кілька ключових переваг: зниження ймовірності галюцинацій, забезпечення фактичної точності, вирішення проблеми застарілих знань шляхом легкого оновлення зовнішньої бази, а також підвищення прозорості через можливість надання посилань на джерела.

Графи знань

Графи знань є потужним інструментом для представлення інформації у структурованому вигляді, що становить великий інтерес для RAG-систем. Графи складаються з сутностей (вузлів) і відношень (ребер), що їх пов'язують, утворюючи семантичну мережу. Сутності можуть представляти реальні об'єкти, поняття, події, а відношення описують зв'язки між ними (наприклад, «Альберт Ейнштейн — є автором — теорії відносності»). Така структура дає змогу не лише зберігати факти, а і явно моделювати складні взаємозв'язки.

Потенціал графів знань для покращення RAG-систем є значним. На відміну від простого пошуку за ключовими словами у текстових корпусах, графи знань дають можливість здійснювати більш точний і контекстуально обґрунтований ретривінг. Завдяки чітко визначеним відношенням, можна виконувати складні запити, здійснювати багатоетапні логічні висновки та виявляти неявні зв'язки. Інтеграція графів знань у RAG-системи може забезпечити ВММ більш багатим і структурованим контекстом.

Методи добування знань для графів знань

Побудова високоякісного графу знань визначається застосованими методами добування знань з різнорідних джерел. Основні завдання добування знань такі: розпізнавання іменованих сутностей (NER), виявлення відношень (RE), зв'язування сутностей (EL) і виявлення подій (Event Extraction).

Основні методи добування знань:

– Методи, основані на правилах і лінгвістичних патернах (Rule-based KE): використання вручну створених або напівавтоматично індукованих шаблонів. Переваги: висока точність у вузьких доменах, інтерпретованість. Недоліки: низька повнота, погана масштабованість, трудомісткість. Для RAG: створюють високоточні, але розріджені графи знань.
– Методи на основі класичного машинного навчання (Classical ML-based KE): використання алгоритмів типу CRF, SVM. Переваги: краща узагальнювальна здатність, навчання на анотованих даних. Недоліки: залежність від інженерії ознак, потреба в анотованих даних. Для RAG: якість графу залежить від даних та ознак, можливий шум.
– Методи на основі глибокого навчання (Deep Learning-based KE): Ззастосування нейронних мереж (RNN, Transformers, BERT). Переваги: state-of-the-art результати, автоматичне вивчення ознак. Недоліки: потреба у великих даних, обчислювальні витрати, менша інтерпретованість. Для RAG: потенціал для повних і багатих графів знань, але помилки моделей можуть проникати в граф.
– Методи зв'язування сутностей (EL) і розширення графів (KG Completion): EL зіставляє сутності з канонічними ідентифікаторами. KG Completion передбачає відсутні зв'язки. Для RAG: EL критичний для усунення неоднозначності; KG Completion може збагатити, але й внести неточні зв'язки [5].

Вибір і комбінація цих методів визначають щільність, точність, повноту, актуальність та узгодженість графу, що безпосередньо впливає на здатність RAG-системи ефективно знаходити та використовувати релевантну інформацію до запиту користувача.

Архітектури наявних RAG-систем на основі графів знань

Загальний принцип роботи KG-RAG систем передбачає інтерпретацію запиту, пошук фактів з графу знань, формування доповненого контексту та генерацію відповіді ВММ.

У відкритому доступі є дві основні RAG-системи, що використовують графи знань, — GraphRAG [1] і LightRAG [3]. Відмінності у підходах до інтеграції графу знань полягають у тому, що LightRAG використовує його як швидкий довідник, GraphRAG — як модель предметної області. Тож для LightRAG критичною є точність видобутку знань, оскільки шум сильно деградує якість пошуку. Для GraphRAG важлива повнота видобутих знань для формування якісних результатів; певний рівень шуму може бути менш критичним, але систематичні помилки, що спотворюють структуру, також шкідливі.

При отриманні користувацького запиту LightRAG розбиває його на ключові слова, абстрактні і конкретні, для кращої контекстної обізнаності. Після цього у векторному сховищі знаходить найбільш подібні результати. Далі для них іде пошук в рамках графу знань і текстових фрагментів, що разом складуть контекст для відповіді ВММ. У GraphRAG аналіз покладається на роботу з узагальненнями, що були створені на етапі індексації, і ВММ отримує узагальнення та запит користувача і дає об'єднану відповідь.

Теоретичний аналіз впливу методів добування знань на ефективність RAG-систем

Граф знань для систем, що використовують його, є ключовим для відповідей через використання його в момент оброблення запиту користувача. Це призводить до необхідності аналізу цих процесів. Різні методи видобутку знань генерують граф знань із різною точністю, повнотою, щільністю, наявністю помилок та рівнем гранулярності. Наприклад, rule-based можуть дати високу точність, але низьку повноту, тоді як DL-методи можуть забезпечити вищу повноту, але з ризиком внесення шуму.

Кожна властивість графу знань впливатиме на ефективність роботи RAG-систем. Високоточні графи знань зменшують ризик передання ВММ помилкових фактів, знижуючи галюцинації. Більш повні графи знань надають багатший контекст, дозволяючи відповідати на складні запити (особливо multi-hop). Якість зв'язування сутностей забезпечує коректну інтеграцію інформації та навігацію всередині графу, уникаючи розпорошення інформації. Гранулярність відношень між сутностями дозволяє ВММ краще розуміти семантику при відповіді. За структурної цілісності ми отримуємо зв'язаний граф, який полегшує етап пошуку релевантних запиту сутностей і відношень.

Відповідно до цих властивостей графу і з'являються потенційні проблеми: шум у вилучених даних (помилкові факти, неправильні відношення), пропущені сутності / відношення, проблеми масштабування та підтримки правил. Для запобігання таким проблемам можна регулювати щільність графа, якість вихідного тексту для видобутку знань, складність і тип користувацьких запитів.

Аналіз видобутку знань у наявних системах

Основою RAG-систем є опрацювання документів, що надає користувач для подальшого пошуку в них. Подальше оброблення документів і перетворення їх на граф знань є частиною індексації вхідних даних.

У GraphRAG і LightRAG основна робота з видобутку знань лягає на ВММ, яка працює з частиною документа. GraphRAG орієнтований на одноразову побудову графу знань, і операції індексації нових документів вимагають перебудови графу знань для включення нового документа. ВММ у Graph є набором запитів до ВММ, де вказується структура очікуваної відповіді та налаштування контексту ВММ як асистента. При цьому GraphRAG допускає зміну цих запитів для покращення результатів при налаштуванні користувацьких потреб [1].

LightRAG теж покладається на ВММ модель, але додатково оперує «контекстом», що позначає суміжні сутності і відношення у графі знань. Додатково індексування нового документа вимагає менше кроків через пошук відповідних сутностей і відношень у графі. Це дає змогу органічно доповнювати граф знань без потреби повної перебудови графу знань.

Обидві системи покладаються на ВММ, що призводить до залежності від якості ВММ. Автори систем використовували моделі gpt-4-turbo для GraphRAG і GPT-4o-mini для LightRAG. ВММ чутливі до даних, на яких вони навчалися, що призводить до різних результатів. Іншим параметром, який впливатиме на результат, буде розмір фрагментів, на які розбивається документ: він має бути досить великим для збереження сенсу тексту у фрагменті і досить малим для оброблення ВММ без галюцинування [1; 3].

Висновки

Проведений аналіз підтверджує критичну залежність ефективності RAG-систем від методів видобування знань. Якість, повнота і точність графу знань безпосередньо впливають на здатність RAG-системи надавати релевантні, точні та правдиві відповіді. Різні архітектури RAG-систем (наприклад, LightRAG і GraphRAG) висувають різні вимоги до характеристик графів знань і, відповідно, до пріоритетних аспектів видобутку знань (точність або повнота).

Список літератури

1. Edge D. From Local to Global: A Graph RAG Approach to Query-Focused Summarization [Electronic resource] / D. Edge, H. Trinh, N. Cheng, J. Bradley, A. Chao, A. Mody, S. Truitt, D. Metropolitansky, R. O. Ness, J. Larson // arXiv.org. — Mode of access: https://doi.org/10.48550/arXiv.2404.16130 (date of access: 10.06.2025).
2. Gao Y. Retrieval-Augmented Generation for Large Language Models: A Survey. [Electronic resource] / Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, M. Wang, H. Wang // arXiv.org. — Mode of access: https://doi.org/10.48550/arXiv.2312.10997 (date of access: 01.06.2025).
3. Guo Z. LightRAG: Simple and Fast Retrieval-Augmented Generation. [Electronic resource] /Z. Guo, L. Xia, Y. Yu, T. Ao, C. Huang // arXiv.org. — Mode of access: https://doi.org/10.48550/arXiv.2410.05779 (date of access: 10.06.2025).

4. Huang L. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions [Electronic resource] / L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, T. Liu // arXiv.org. — Mode of access: https://doi.org/10.48550/arXiv.2311.05232 (date of access: 07.06.2025).
5. Liu P. A Survey on Open Information Extraction from Rule-based Model to Large Language Model [Electronic resource] / P. Liu, W. Gao, W. Dong, L. Ai, Z. Gong, S. Huang, Z. Li, E. Hoque, J. Hirschberg, Y. Zhang // arXiv.org. — Mode of access: https://doi.org/10.48550/arXiv.2208.08690 (date of access: 25.05.2025).

References

Edge, D., Trinh, H., Cheng, N., Bradley, J., Chao, A., Mody, A., Truitt, S., Metropolitansky, D., Ness, R. O., & Larson, J. (2024). From Local to Global: A Graph RAG Approach to Query-Focused Summarization. *arXiv preprint arXiv:2404.16130* (date of access: 10.06.2025). https://doi.org/10.48550/arXiv.2404.16130.

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., & Wang, H. (2023). Retrieval-Augmented Generation for Large Language Models: A Survey. *arXiv preprint arXiv:2312.10997* (date of access: 01.06.2025). https://doi.org/10.48550/arXiv.2312.10997.

Guo, Z., Xia, L., Yu, Y., Ao, T., & Huang, C. (2024). LightRAG: Simple and Fast Retrieval-Augmented Generation. *arXiv preprint arXiv:2410.05779* (date of access: 10.06.2025). https://doi.org/10.48550/arXiv.2410.05779.

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., & Liu, T. (2023). A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *arXiv preprint arXiv:2311.05232* (date of access: 07.06.2025). https://doi.org/10.48550/arXiv.2311.05232.

Liu, P., Gao, W., Dong, W., Ai, L., Gong, Z., Huang, S., Li, Z., Hoque, E., Hirschberg, J., & Zhang, Y. (2022). A Survey on Open Information Extraction from Rule-based Model to Large Language Model. *arXiv preprint arXiv:2208.08690* (date of access: 25.05.2025). https://doi.org/10.48550/arXiv.2208.08690.

M. Androshchuk

INFLUENCE OF KNOWLEDGE EXTRACTION METHODS ON THE EFFECTIVENESS OF GRAPH-BASED RAG SYSTEMS

This paper investigates the impact of knowledge extraction methods on the effectiveness of RAG (Retrieval-Augmented Generation) systems that utilize knowledge graphs. It highlights that the quality of the knowledge graph, which is formed using various knowledge extraction methods, is crucial for overcoming limitations of large language models (LLMs), such as "hallucinations". The paper analyzes the architectures of LightRAG and GraphRAG, emphasizing that the selection of an optimal knowledge extraction strategy depends on specific tasks and the subject area. LLMs have advanced significantly, but they have limitations, including generating factually incorrect information ("hallucinations") and possessing "outdated knowledge". RAG systems were proposed to address these issues by combining LLMs with external knowledge bases. This approach reduces hallucinations, ensures factual accuracy, solves the problem of outdated knowledge, and increases transparency. Knowledge graphs are powerful tools for structuring information, consisting of entities (nodes) and relations (edges). They enhance RAG systems by enabling more precise and contextually grounded retrieval compared to keyword-based searches. The quality of a knowledge graph depends on the knowledge extraction methods used, which include named entity recognition (NER), relation extraction (RE), entity linking (EL), and event extraction. Different methods, such as rule-based, classical machine learning, and deep learning approaches, have varying trade-offs in terms of accuracy, completeness, and scalability. Entity linking and knowledge graph completion are also crucial for accuracy and richness. LightRAG and GraphRAG are two main graph-based RAG systems. LightRAG uses the knowledge graph as a quick reference, requiring high precision in knowledge extraction to avoid noise degradation. GraphRAG uses the knowledge graph as a domain model, where completeness of extracted knowledge is more critical, though systematic errors are still harmful. Both systems rely on LLMs for knowledge extraction, which makes them dependent on the LLM's quality and the size of document fragments processed. The theoretical analysis confirms that the effectiveness of RAG systems is critically dependent on knowledge extraction methods. The quality, completeness, and accuracy of the knowledge graph directly influence the RAG system's ability to provide relevant, accurate, and truthful answers. Different RAG system architectures like LightRAG and GraphRAG have distinct requirements for knowledge graph characteristics, prioritizing either accuracy or completeness in knowledge extraction.

Keywords: RAG, knowledge graphs, knowledge extraction, LLM.

Матеріал надійшов 23.06.2025