

ПРОГРАМНИЙ ЗАСІБ ІНТЕЛЕКТУАЛЬНОГО ПОШУКУ НА ОСНОВІ RAG-МОДЕЛІ

Вінницькій національний технічний університет

Анотація

У роботі досліджено особливості розробки програмного засобу інтелектуального пошуку на основі підходу Retrieval-Augmented Generation (RAG). Проаналізовано принципи поєднання пошукових алгоритмів та генеративних моделей штучного інтелекту для підвищення точності та релевантності результатів. Обґрунтовано вибір архітектури системи та проведено оцінку ефективності її роботи.

Ключові слова: RAG, інтелектуальний пошук, штучний інтелект, NLP, генеративні моделі, векторні бази даних.

Abstract: *The paper investigates the development of an intelligent search system based on the Retrieval-Augmented Generation (RAG) approach. The principles of combining retrieval mechanisms and generative AI models to improve accuracy and relevance are analyzed. The system architecture is justified and its performance is evaluated.*

Keywords: RAG, intelligent search, artificial intelligence, NLP, generative models, vector databases.

Вступ

У сучасних інформаційних системах актуальною є проблема підвищення точності пошуку даних та генерації релевантних відповідей. Традиційні пошукові системи обмежені у контекстному розумінні запитів, тоді як генеративні моделі можуть створювати неточну або вигадану інформацію.

Метою роботи є розробка програмного засобу інтелектуального пошуку на основі RAG, що поєднує пошук у базі знань із генерацією відповідей.

Для досягнення мети визначено такі завдання:

- проаналізувати сучасні підходи до інтелектуального пошуку;
- розробити архітектуру системи на основі RAG;
- реалізувати механізм обробки запитів користувача;
- оцінити ефективність роботи системи.

Основна частина

Запропонований програмний засіб базується на архітектурі Retrieval-Augmented Generation (RAG), що поєднує можливості інформаційного пошуку та генеративних моделей штучного інтелекту. Такий підхід дозволяє вирішити проблему обмеженості традиційних пошукових систем та зменшити кількість помилок, притаманних генеративним моделям.

На першому етапі роботи системи здійснюється попередня обробка запиту користувача, що включає нормалізацію тексту та перетворення його у векторне представлення (ембедінг). Після цього виконується пошук найбільш релевантних документів у векторній базі даних із використанням метрик семантичної подібності. Це дозволяє враховувати не лише ключові слова, а й зміст запиту.

Другий етап передбачає передачу знайдених документів до генеративної моделі, яка формує узагальнену відповідь на основі отриманого контексту. Важливою перевагою такого підходу є зменшення рівня так званих «галюцинацій», коли модель генерує недостовірну або вигадану інформацію.

Архітектура системи реалізована за модульним принципом і включає підсистеми обробки запитів, пошуку інформації та генерації відповідей. Така структура забезпечує гнучкість, масштабованість та можливість подальшого розширення функціоналу.

З метою оцінювання ефективності запропонованого підходу було проведено експериментальне дослідження, у якому здійснено порівняння класичного пошуку, генеративної моделі та RAG-системи за основними показниками якості обробки запитів.

Для аналізу використовувалися такі критерії, як точність відповіді, релевантність результатів, час обробки запиту, рівень галюцинацій та повнота відповіді. Отримані результати наведено в таблиці 1

Таблиця 1 – Порівняння ефективності підходів пошуку

Показник	Класичний пошук	Генеративна модель	RAG-система
Точність відповіді, %	65	75	90
Релевантність результатів, %	60	70	88
Час обробки запиту, с	0,3	1,2	0,8
Рівень галюцинацій, %	0	25	8
Повнота відповіді, %	50	85	92

Отримані результати демонструють, що використання RAG-підходу дозволяє досягти найкращого балансу між точністю, релевантністю та повнотою відповіді. У порівнянні з класичним пошуком система забезпечує більш змістовні результати, а у порівнянні з генеративними моделями – значно зменшує рівень галюцинацій.

Таким чином, застосування RAG є ефективним рішенням для побудови інтелектуальних пошукових систем.

Висновки

У результаті проведеного дослідження розроблено програмний засіб інтелектуального пошуку на основі підходу Retrieval-Augmented Generation, що поєднує механізми семантичного пошуку та генерації тексту. Експериментальні результати підтвердили доцільність використання RAG-архітектури для підвищення якості обробки запитів користувачів. Встановлено, що запропонований підхід забезпечує вищу точність та релевантність відповідей порівняно з класичними пошуковими методами, а також дозволяє суттєво зменшити рівень галюцинацій, характерних для генеративних моделей.

Реалізація системи на основі модульної архітектури забезпечує її масштабованість, гнучкість та можливість інтеграції з різними джерелами даних і сервісами обробки природної мови.

Таким чином, використання RAG-підходу є перспективним напрямом розвитку інтелектуальних інформаційних систем, зокрема у веб-застосунках, чат-ботах та системах підтримки прийняття рішень.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Lewis P. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. – 2020.
2. Jurafsky D., Martin J. Speech and Language Processing. – 2021.

Ткаченко Олександр Миколайович – к.т.н., доцент кафедри програмного забезпечення, Вінницький національний технічний університет, e-mail: alextk1960@gmail.com

Богонос Богдан Олександрович – студент групи ЗПІ-22б, Факультет інформаційних технологій та комп'ютерної інженерії, Вінницький національний технічний університет, м. Вінниця, e-mail: rihta14@gmail.com

Tkachenko Olexsandr Mykolaiovych – Candidate of Technical Sciences, Associate Professor of the Department of Software, Vinnytsia National Technical University, Vinnytsia.

Bohonos Bohdan Olexsandrovych – student of the group ZPI-22b, Faculty of Information Technology and Computer Engineering, Vinnytsia National Technical University, Vinnytsia.