УДК 004.41
В. Ямко; М. Петрик, д.ф.-м.н., професор
(Тернопільський національний технічний університет імені Івана Пулюя, Україна)

ВПРОВАДЖЕННЯ ВЕКТОРНОГО ПОШУКУ У СИСТЕМІ ДОКУМЕНТООБІГУ

UDC 004.41
V. Yamko; M. Petryk, Doctor of Physical and Mathematical Sciences, Professor

IMPLEMENTATION OF VECTOR SEARCH IN A DOCUMENT MANAGEMENT SYSTEM

Системи управління документами пройшли довгий шлях від фізичних архівів до складних цифрових рішень. Сьогодні вони активно використовуються у багатьох сферах для зберігання, пошуку та аналізу даних. Однією з важливих сучасних тенденцій у цій галузі є інтеграція штучного інтелекту, зокрема методів роботи з векторними просторами та контекстного аналізу тексту.

Звичайний пошук за ключовими словами у текстових полях, таких як назва документа чи ім'я клієнта, має суттєві обмеження. Він часто не враховує семантичну близькість слів. Наприклад, запити «договір оренди» та «контракт на оренду» можуть бути релевантними, але традиційні методи пошуку цього не розуміють.

Для розв'язання цієї проблеми використовуються текстові ембедінги – числові векторні представлення тексту, які зберігають його семантичний зміст. Завдяки алгоритмам, як-от Word2Vec, GloVe, або сучасним моделям типу BERT і GPT, кожен текстовий фрагмент можна представити у вигляді багатовимірного вектора [1]. Векторні пошуки дозволяють знаходити документи за їх схожістю у векторному просторі, навіть якщо слова у запиті не повністю збігаються зі словами у документі.

Наприклад, запит може бути перетворений на вектор, який порівнюється із векторами, що представляють документи у базі. Це відкриває нові можливості для гнучкого пошуку:

- Розуміння синонімів та контексту.
- Пошук релевантних документів, навіть якщо формулювання у запиті відрізняється.
- Аналіз зв’язків між документами для рекомендації пов’язаних матеріалів.

У випадках, коли документ містить великий обсяг тексту, наприклад, десятки сторінок, пряме використання ембедінгів для всього тексту може бути неефективним. Виникає ризик втрати важливої інформації або збільшення обчислювальної складності.

Контекстне чанкування є підходом, який дозволяє розбивати великий текст на менші логічні частини (чанки), кожна з яких аналізується окремо [2]. Наприклад, документ можна поділити за абзацами, розділами або навіть реченнями, залежно від поставленої задачі.

Чанки обробляються за допомогою моделей, які створюють ембедінги для кожного фрагмента тексту. Це дозволяє:

- Проводити більш точний пошук, фокусуючись лише на релевантних частинах документа.
- Підвищувати ефективність обчислень за рахунок паралельної обробки фрагментів.
- Зберігати контекстні зв’язки між частинами документа, що важливо для відновлення повної картини під час пошуку.

Для зберігання та пошуку векторів у реальному часі використовуються спеціалізовані рішення, такі як Pinecone, Weaviate або MongoDB Atlas Search [3]. Вони підтримують:

- Зберігання багатовимірних векторів у базах даних.
- Пошук найближчих сусідів (nearest neighbor search) для швидкого знаходження релевантних записів.
- Інтеграцію з традиційними SQL або NoSQL системами для гібридного підходу до пошуку.

Такі системи можуть стати незамінними у великих організаціях, де пошук і аналіз документів є ключовою задачею. Завдяки ембедінгам і контекстному чанкуванню досягається новий рівень точності та швидкості обробки інформації.

Література

1. Hugging Face Transformers [Електронний ресурс] – Режим доступу до ресурсу: https://huggingface.co/docs/transformers/index
2. LangChain Documentation [Електронний ресурс] – Режим доступу до ресурсу: https://python.langchain.com/docs/introduction/
3. MongoDB Atlas Search Documentation [Електронний ресурс] – Режим доступу: https://www.mongodb.com/docs/atlas/search/.