УДК 004.41

Л. Б. Ліщинська
М. Р. Слободяник

ЗАСТОСУВАННЯ ВЕЛИКИХ МОВНИХ МОДЕЛЕЙ ДЛЯ АВТОМАТИЗАЦІЇ АНАЛІЗУ ТАБЛИЧНИХ ДАНИХ

Vinnytsia National Technical University

Анотація

У роботі Розглянуто підходи до інтеграції великих мовних моделей у системи аналізу табличних даних. Проаналізовано методи профілювання структури даних, LLM-планування аналітичних операцій та автоматичної генерації текстових звітів. Подано порівняльну характеристику підходів до визначення набору аналітичних функцій для довільного набору даних.

Ключові слова: великі мовні моделі; аналіз даних; автоматична генерація звітів; LLM-планування; Spring AI.

Abstract

Approaches to integrating large language models into tabular data analysis systems are considered. Methods for data structure profiling, LLM-based analytical planning, and automatic text report generation are analyzed. A comparative characterization of approaches to determining the set of analytical functions for an arbitrary dataset is provided.

Keywords: large language models; data analysis; automatic report generation; LLM planning; Spring AI.

Вступ

Аналіз табличних даних є одним із найпоширеніших завдань у бізнесі, освіті та науковій діяльності. Традиційно цю задачу вирішували або спеціалізованими статистичними інструментами, що вимагають технічних знань, або BI-платформами, орієнтованими на роботу з підключеними базами даних. Жоден із цих підходів не дозволяє отримати повноцінний текстовий аналітичний висновок для довільного файлу без участі кваліфікованого спеціаліста. Поява великих мовних моделей відкриває принципово новий підхід до вирішення цієї задачі [1].

Основна частина

Ключовою проблемою при автоматизованому аналізі табличних даних є визначення того, які саме обчислення доцільно виконати для конкретного набору. Традиційні підходи базуються на правилових системах, де для кожного типу даних прописується фіксований набір аналітичних функцій. Такі системи передбачувані, але не здатні враховувати семантику даних – для таблиці продажів і для медичних показників доцільний набір метрик принципово різний.

Альтернативний підхід полягає у передачі цієї задачі великій мовній моделі. LLM отримує опис структури даних: назви колонок, їх типи, семантичні ролі та прев'ю значень. На основі цього модель формує план аналізу у вигляді структурованого JSON із переліком аналітичних функцій, цільових колонок та обґрунтуванням вибору. Перевага такого підходу полягає у здатності моделі розрізняти, наприклад, числову колонку з датами і числову колонку з цінами, та обирати відповідні методи аналізу [2].

Важливим аспектом є розмежування між задачами, які виконує LLM, і задачами, що вирішуються детерміновано. Самі обчислення – агрегації, розподіли, виявлення аномалій методом Z-score, кореляційний аналіз, часові тренди – виконуються на сервері без участі моделі. LLM залучається лише на двох етапах: планування та генерація тексту. Це гарантує точність числових результатів і водночас забезпечує гнучкість у виборі методів аналізу та якість текстових висновків [3].

Окремою задачею є профілювання структури вхідних даних. До передачі метаданих у LLM система повинна самостійно визначити типи колонок і їх семантичні ролі. Статистичний підхід до

профілювання передбачає аналіз вибірки значень кожної колонки: оцінюється частка числових записів, перевіряється відповідність форматам дат, визначається кількість унікальних значень відносно загальної кількості рядків. На основі цього колонкам присвоюються ролі числової міри, категоріального виміру, часового ряду або ідентифікатора. Точність профілювання безпосередньо впливає на якість подальшого планування – модель не може запропонувати доречний аналіз, якщо отримала некоректний опис структури даних.

Порівнюючи два підходи, правилові системи вирізняються передбачуваністю та відсутністю залежності від зовнішніх сервісів, проте не здатні враховувати семантику даних і пропонують однаковий набір метрик незалежно від предметної області. LLM-планування, навпаки, адаптує набір аналітичних операцій під конкретний контекст — модель розрізняє числову колонку з цінами і числову колонку з температурами та обирає відповідні методи. Недоліком такого підходу є залежність від доступності зовнішнього API та певна непередбачуваність формату відповіді, що вирішується механізмом повторних спроб парсингу. З точки зору якості текстових висновків перевага LLM-підходу є беззаперечною: шаблонні тексти не дають аналітичної цінності, тоді як мовна модель здатна інтерпретувати числові результати у контексті предметної області.

Генерація текстового звіту є фінальним етапом, де LLM отримує числові результати розрахунків і формує структурований документ із секціями загального огляду, ключових показників, детального аналізу, ризиків та рекомендацій. Підхід із шаблонними текстами тут не працює, оскільки не дає аналітичної цінності. Натомість модель здатна інтерпретувати числові результати у контексті предметної області та формулювати змістовні висновки природною мовою [1].

Висновки

Інтеграція великих мовних моделей у системи аналізу табличних даних дозволяє автоматизувати два принципово різні за природою завдання: вибір методів аналізу та інтерпретацію результатів. Поєднання LLM-планування з детермінованим виконанням розрахунків на стороні сервера забезпечує баланс між точністю обчислень і гнучкістю аналітичного підходу. Такі системи відкривають можливість отримання структурованого аналітичного висновку для довільного табличного файлу без технічних знань з боку користувача.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Ozdemir S. Quick Start Guide to Large Language Models. — Addison-Wesley Professional, 2023. — 384 p.
2. Gemini API Documentation [Електронний ресурс]. – Режим доступу: https://ai.google.dev/gemini-api/docs.
3. Spring AI [Електронний ресурс]. – Режим доступу: https://spring.io/projects/spring-ai.

Ліщинська Людмила Броніславівна **–** д-р техн. наук, професор, професор кафедри програмного забезпечення, Вінницький національний технічний університет, м. Вінниця, e-mail: llb@vntu.edu.ua

Слободяник Михайло Романович **–** студент групи 5ПІ-22б, Факультет інформаційних технологій та комп'ютерної інженерії, Вінницький національний технічний університет, м. Вінниця, mishaslobodianik26@gmail.com

Lyudmila Bronislavivna Lishchynska **–** Dr. Sc. (Eng.), Full Professor, Professor of Program Engineering, Vinnytsia National Technical University, Vinnytsia, e-mail: llb@vntu.edu.ua.

Slobodianyk Mykhailo Romanovych **–** student of group 5PI-22b, Faculty of Information Technology and Computer Engineering, Vinnytsia National Technical University, Vinnytsia.