УДК 004.8:004.89

ВИКОРИСТАННЯ ВЕЛИКИХ МОВНИХ МОДЕЛЕЙ ДЛЯ ГЕНЕРАЦІЇ СИНТЕТИЧНИХ ДАНИХ У РЕКОМЕНДАЦІЙНИХ СИСТЕМАХ

Фарафонов Є.Ю.
e-mail: yevhenii.farafonov@nure.ua
Науковий керівник – к.т.н., проф. Рябова Н.В.
Харківський національний університет радіоелектроніки, каф. ШІ
м. Харків, Україна

One of the main challenges in developing recommendation algorithms is the sparsity of user–item interaction data, which limits the effectiveness of traditional methods such as collaborative filtering. Synthetic data generation is considered a promising approach for expanding training datasets and improving the robustness of recommendation models. The proposed approach involves using a large language model to generate synthetic interactions between users and items based on descriptions of user preferences and item characteristics. Such generated data can simulate realistic usage scenarios and supplement existing datasets. The study proposes to investigate methods for generating synthetic interactions, modeling different types of user behavior, and evaluating the impact of generated data on recommendation quality.

У сучасних інформаційних системах рекомендаційні алгоритми широко використовуються для персоналізації контенту, товарів та послуг. Такі системи активно застосовуються у електронній комерції, мультимедійних сервісах, соціальних мережах та інформаційних платформах. Ефективність рекомендацій значною мірою залежить від обсягу та якості доступних даних про взаємодію користувачів з об’єктами системи. Проте на практиці часто виникає проблема розрідженості даних (data sparsity), коли кількість доступних взаємодій є недостатньою для побудови точних моделей рекомендацій [1].

Існуючі рекомендаційні системи базуються на різних підходах, серед яких виділяють колаборативну фільтрацію, контентно-орієнтовані методи та гібридні моделі. Незважаючи на їх ефективність, ці методи потребують значних обсягів історичних даних для навчання моделей. У випадках, коли кількість взаємодій обмежена або система тільки починає функціонувати, виникає проблема “холодного старту”, що призводить до зниження якості рекомендацій .

Одним із перспективних напрямів вирішення цієї проблеми є використання синтетичних даних, які генеруються штучно для розширення навчальних наборів. Синтетичні дані можуть моделювати поведінку користувачів, їхні інтереси та можливі сценарії взаємодії з об’єктами системи. Такий підхід дозволяє збільшити різноманітність навчальних прикладів та зменшити залежність від обмежених реальних даних [2].

Останніми роками значного розвитку набули великі мовні моделі, які здатні аналізувати та генерувати складні текстові структури, моделювати контекст та відтворювати закономірності у великих масивах інформації. Завдяки цьому вони можуть бути використані для створення синтетичних описів користувачів, сценаріїв взаємодії або навіть повних наборів взаємодій між користувачами та об’єктами рекомендаційної системи. Такі моделі дозволяють формувати більш реалістичні та різноманітні дані, які можуть доповнювати існуючі набори даних [3].

Застосування великих мовних моделей для генерації синтетичних даних має ряд переваг. По-перше, це можливість швидкого створення великого обсягу даних без необхідності збору інформації від реальних користувачів. По-друге, такий підхід може сприяти зменшенню проблем конфіденційності, оскільки синтетичні дані не містять прямої інформації про реальних користувачів. По-третє, генеративні моделі можуть створювати різні сценарії поведінки, що дозволяє підвищити стійкість алгоритмів рекомендацій до нових або нестандартних ситуацій .

Запропоноване дослідження передбачає використання великої мовної моделі як генератора синтетичних взаємодій між користувачами та об’єктами системи. На основі опису предметної області, характеристик користувачів та властивостей об’єктів модель може формувати нові записи, що імітують реальні сценарії використання системи. Згенеровані дані можуть містити інформацію про уподобання користувачів, оцінки об’єктів або інші типи взаємодій, які використовуються у рекомендаційних алгоритмах.

Базова рекомендаційна система може бути реалізована на основі колаборативної фільтрації, де використовується матриця взаємодій користувачів та об’єктів.

На першому етапі формується початковий набір даних, які використовуються для аналізу основних характеристик користувачів, властивостей об’єктів та типових сценаріїв їх взаємодії.

Далі на основі отриманих характеристик формується структурований запит до мовної моделі, який визначає правила генерації синтетичних взаємодій. У такому запиті задається опис предметної області, що включає перелік можливих типів користувачів, їх поведінкові характеристики та інтереси, а також опис об’єктів рекомендації, зокрема їх категорії, основні властивості та рівень популярності. Крім цього у запиті визначається формат вихідних даних, який повинен відповідати структурі, що використовується у рекомендаційних алгоритмах. Наприклад, моделі пропонується згенерувати набір записів у вигляді структурованих взаємодій типу «ідентифікатор користувача – ідентифікатор об’єкта – оцінка», «користувач – об’єкт – перегляд», або «користувач – об’єкт – клік».

Для забезпечення більшої різноманітності синтетичних даних у запиті також можуть задаватися додаткові обмеження та параметри генерації, такі як кількість взаємодій, можливий діапазон оцінок, розподіл популярності об’єктів або співвідношення між різними типами взаємодій. Генерація виконується окремо для різних типів користувачів або категорій об’єктів, що дозволяє моделі формувати взаємодії, які відображають різні сценарії поведінки користувачів у системі. У результаті формується набір синтетичних записів, що імітують можливі взаємодії користувачів з об’єктами рекомендаційної системи та можуть бути використані для подальшого розширення навчальних даних.

Згенеровані записи приводяться до табличного формату, проводиться перевірка отриманих даних, яка включає видалення дубльованих або некоректних записів, перевірку допустимих значень оцінок та аналіз розподілу взаємодій між користувачами і об’єктами. Додатково може виконуватись статистичне порівняння основних характеристик синтетичного набору даних з відповідними характеристиками реального набору з метою оцінювання їх подібності та реалістичності.

Після цього сформований набір синтетичних даних об’єднується з наявними реальними даними та використовується для навчання і тестування рекомендаційних алгоритмів.

У межах дослідження доцільно провести порівняльний аналіз результатів роботи моделей, навчених на різних варіантах наборів даних: лише на реальних даних та на розширених наборах, що включають синтетично згенеровані записи.

Окрему увагу можна приділити оцінюванню впливу різних варіантів запитів до мовної моделі, використаних під час генерації даних, на якість отриманих рекомендацій. Це дозволить визначити, які підходи до формування запитів забезпечують найбільш корисні синтетичні дані для навчання рекомендаційних алгоритмів.

Список використаних джерел:

1. Aggarwal C. C. Recommender systems: the textbook. Springer, 2018. 519 p.

2. Bengio Y., Courville A., Goodfellow I. Deep learning. MIT Press, 2016. 800 p.

3. Leveraging large language models for pre-trained recommender systems. arXiv.org e-Print archive. URL: https://arxiv.org/pdf/2308.10837 (date of access: 11.03.2026).