

ІНТЕГРАЦІЯ КОНТЕНТНИХ ОЗНАК ПРИ ІНІЦІАЛІЗАЦІЇ ЛАТЕНТНИХ ФАКТОРІВ В СИСТЕМАХ ПЕРСОНАЛІЗОВАНИХ РЕКОМЕНДАЦІЙ

¹ Вінницький національний технічний університет

Анотація

В роботі розглянуто підхід до вдосконалення матричної факторизації шляхом інтеграції контентних ознак під час ініціалізації латентних факторів. Проаналізовано обмеження стандартного методу SVD, зокрема проблему «холодного старту», та запропоновано інтеграцію контентної та контекстної моделей, що об'єднує факторизацію та векторизацію контентних характеристик. Показано, що використання методу TF-IDF як початкової параметризації дозволяє покращити якість прогнозування та прискорити збіжність процедури навчання.

Ключові слова: системи рекомендацій, матрична факторизація, SVD, TF-IDF, гібридна модель, персоналізація.

Abstract

The study examines an approach to improving matrix factorization by integrating content-based features during the initialization of latent factors. The limitations of the standard SVD method, particularly the “cold start” problem, are analyzed, and an integration of content-based and context-based models is proposed, combining factorization with vectorization of content characteristics. It is shown that using the TF-IDF method for initial parameterization enhances prediction accuracy and accelerates the convergence of the training procedure.

Keywords: recommender systems, matrix factorization, SVD, TF-IDF, hybrid model, personalization.

Вступ

Розвиток сучасних інформаційних технологій сприяє постійному збільшенню обсягів інформації, яка постійно оновлюється та змінюється. Користувачу потрібно витрачати все більше часу для знаходження цікавої та корисної для нього інформації. В цьому допомагають системи персональних рекомендацій, які отримують все більше поширення та допомагають користувачам знаходити відповідний контент, що відповідає їх вподобанням або інтересам. Такі системи прогнозують уподобання користувачів на основі зібраних даних та пропонують персоналізований контент. Їх робота реалізується через використання двох ключових підходів — колаборативної та контентної фільтрації [1].

Класичні методи колаборативної фільтрації демонструють високу точність прогнозування для елементів із великою кількістю історичних взаємодій, проте виявляються неефективними для нового контенту. Матриця рейтингів у реальних веб-сервісах є високорозрідженою, а випадкова ініціалізація латентних факторів у моделі сингулярного розкладання матриць SVD призводить до погіршення якості рекомендацій та тривалого процесу навчання. Тому розробка підходу, здатного ефективно працювати у сценаріях нестачі даних, є актуальною задачею.

Результати досліджень

Метод сингулярного розкладання матриці SVD (Singular Value Decomposition) демонструє низьку похибку RMSE (Root Mean Squared Error) та достатню масштабованість, однак виявляється неспроможним формувати коректні прогнози для елементів, що не мають історичних рейтингів [2]. У таких випадках нові об'єкти отримують випадково згенеровані латентні вектори, і це призводить до неможливості використовувати їх на ранніх етапах рекомендацій, що суттєво погіршує якість системи.

Для розв'язання цієї проблеми запропоновано підхід, у якому текстові описи контенту попередньо перетворюються у числові ознаки за допомогою методу Term Frequency–Inverse Document Frequency (TF-IDF), і ці вектори використовуються для ініціалізації матриці латентних факторів Q перед запуском SVD. Такий підхід дозволяє надати латентним представленням змістовну структуру, що відповідає реальним характеристикам елементів, і забезпечує значно вищу точність прогнозів на початкових етапах роботи моделі. Гібридна факторизація поєднує класичне розкладання матриці рейтингів із контентним представленням елементів, створюючи початкову точку для навчання та підвищуючи як точність передбачень, так і швидкість збіжності алгоритму матричної факторизації.

Матриця рейтингів визначається за формулою:

$$R \approx P \times Q^T,$$

де $P \in R^{U \times K}$ — включає латентні вектори користувачів; $Q \in R^{I \times K}$ — включає латентні вектори контенту; U — кількість користувачів у системі; I — кількість елементів контенту; K — кількість латентних факторів; T — транспонована матриця.

Використання контентних векторів дає можливість зменшити тривалість процесу навчання та покращити початкову якість прогнозів.

Експериментальне дослідження, проведене на датасеті *goodbooks-10k*, виявило суттєве покращення як кількісних, так і якісних характеристик моделі. Зокрема, значення RMSE для базової моделі SVD становило 0,867, тоді як при поєднанні контентної та контекстної моделей — 0,851, що відповідає зменшенню похибки на 0,016 (приблизно 1,8 %). Підвищення якості рекомендацій також підтверджується зростанням метрик точності Precision з 0,312 до 0,335 (збільшення на 7,4 %) та метрик повноти Recall з 0,227 до 0,243 (зростання на 7,0 %). Це свідчить не лише про зниження середньої помилки прогнозу, але й про підвищення релевантності сформованих списків рекомендацій для кінцевого користувача. Встановлено, що поєднання контентної та контекстної моделей значно ефективніше працює з об'єктами, які раніше не мали жодної історії взаємодій, що підтверджує успішне вирішення проблеми «холодного старту».

Поєднання контентної та контекстної моделей показало кращу стійкість до зміни параметрів навчання та менш виражену варіативність результатів між різними запусками моделі, що вказує на її високу здатність стабільно відтворювати результати на нових даних. Крім того, після побудови TF-IDF-векторів оновлення інформації для нового контенту не потребує повного перенавчання моделі: достатньо лише додати новий вектор до матриці Q , що значно підвищує придатність системи до використання у реальних високоінтенсивних сервісах.

Висновки

Інтеграція контентних ознак у процес ініціалізації латентних факторів дозволяє компенсувати слабкі сторони колаборативних методів і забезпечити стабільну якість прогнозування навіть при низькій кількості доступних поведінкових даних.

Запропонований підхід усуває проблему «холодного старту» для нового контенту, забезпечуючи наявність осмислених початкових представлень елементів. Крім того, використання контентних векторів сприяє прискоренню збіжності стохастичного градієнтного спуску, що позитивно впливає на ефективність навчання.

Запропонований підхід також забезпечує підвищення точності рекомендацій у випадках розріджених даних і приводить до покращення ключових метрик, зокрема RMSE, Precision та Recall, у порівнянні з базовими модельними підходами.

Отримані результати підтверджують доцільність подальшого розвитку комбінованих методів та їх впровадження у реальні веб-сервіси.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Aggarwal C. C. *Recommender Systems: The Textbook*. 2nd ed. Springer, 2022. p. 115–165.
2. Burke R. Hybrid Recommender Systems: Past, Present, and Future. *User Modeling and User-Adapted Interaction*, 2020. p. 30–40. doi:10.1023/A:1021240730564.
3. Wang X., He X., Wang Y., Feng F., Chua T.-S. Neural Graph Learning for Recommendation. *IEEE Transactions on Knowledge and Data Engineering*, 2023. p. 12–18.

Воронов Владислав Володимирович — студент групи 2KI-24м, факультет інформаційних технологій та комп'ютерної інженерії, Вінницький національний технічний університет, м. Вінниця, email: krasina952@gmail.com

Войцеховська Олена Валеріївна — к.т.н, доцент, доцент кафедри комп'ютерної техніки, Вінницький національний технічний університет, м. Вінниця

Voronov Vladyslav V. — student of group 2KI-24m, Faculty of Information Technologies and Computer Engineering, Vinnytsia National Technical University, Vinnytsia, email: krasina952@gmail.com

Voitsekhovska Olena V. — Ph.D., Associate Professor, Associate Professor of the Department of Computer Techniques, Vinnytsia National Technical University, Vinnytsia