УДК 004.85 DOI: https://doi.org/10.30837/IYF.IIS.2024.064

ДОСЛІДЖЕННЯ ТА ЗАСТОСУВАННЯ МЕТОДІВ NLP ДЛЯ ВИРІШЕННЯ ПРОБЛЕМИ ХОЛОДНОГО СТАРТУ В РЕКОМЕНДАЦІЙНИХ СИСТЕМАХ

Грішаєва А. М.
Науковий керівник – проф. Рябова Н. В.
Харківський національний університет радіоелектроніки, каф. ШІ
м. Харків, Україна
e-mail: anastasiia.hrishaieva@nure.ua

This work explores the application of Natural Language Processing (NLP) methods to address the cold-start problem in recommendation systems. Specifically, it investigates the use of text vectorization and clustering techniques to analyze technical articles, aiming to enhance the recommendation accuracy for new users or items with no prior interactions. Potential challenges include managing large datasets and optimizing clustering algorithms to capture the nuances of technical texts accurately. This study promises to offer valuable insights into refining recommendation systems through sophisticated text analysis methodologies.

В епоху цифровізації, коли обсяги даних зростають експоненційно, рекомендаційні системи відіграють ключову роль у навігації користувачів через масивні потоки інформації, допомагаючи їм виявляти релевантний контент. Однак, проблема холодного старту залишається значним викликом, особливо при інтеграції нових користувачів або введенні нових елементів до системи. Ця робота зосереджена на застосуванні методів обробки природної мови (NLP) для покращення ефективності рекомендаційних систем у контексті холодного старту.

Проблема холодного старту існує для двох типів сутностей – нові відвідувачі, інформація про вподобання яких відсутня, а також для нових товарів, у яких немає взаємодій [1].

Основна мета дослідження полягає у розробці стратегій, які використовують NLP для мінімізації проблем холодного старту в рекомендаційних системах [2].

Існуючі підходи до вирішення цієї проблеми можна класифікувати на кілька основних категорій. Контент-орієнтовані методи використовують інформацію про самі елементи (наприклад, описи, теги, категорії) для рекомендацій. Деякі варіанти колаборативної фільтрації використовують гібридні моделі, що поєднують контент-орієнтовані та колаборативні підходи. Підходи, що базуються на метаданих, використовують додаткову інформацію про користувачів або елементи, таку як вік, стать, географічне розташування користувачів або категорії та теги елементів. Кожен з цих підходів має свої переваги та недоліки, і часто найкращі результати

досягаються за допомогою їх комбінації, щоб врахувати різноманіття сценаріїв взаємодії користувачів і елементів у рекомендаційних системах.

Методи обробки природної мови (NLP) відіграють ключову роль у вдосконаленні рекомендаційних систем, особливо при вирішенні проблеми холодного старту. Вони дозволяють системам краще розуміти інтереси та потреби користувачів за допомогою аналізу текстової інформації, такої як описи продуктів, відгуки користувачів, а також інші текстові дані.

Основні методи NLP, які можуть бути застосовані у контексті вирішення проблеми холодного старту, включають в себе наступні пункти. Аналіз настроїв може допомогти ідентифікувати продукти або послуги, які користуються популярністю або несприйняттям, і відповідно адаптувати рекомендації. Векторні представлення, такі як Word2Vec, GloVe або FastText, можуть допомогти знайти зв'язки між новими елементами або користувачами та існуючими в системі, навіть без історичних даних про взаємодії. Кластеризація може сприяти кращому розумінню різноманітності контенту або інтересів користувачів і, відповідно, покращити точність рекомендацій.

Обраною предметною областю для дослідження та застосування методів NLP стали технічні статті. Технічні статті є важливим джерелом інформації для професіоналів у сферах науки, технологій та інженерії. Вони містять описи новітніх технологій, методів досліджень та практичних рішень, які впливають на розвиток та прогрес відповідних галузей. Враховуючи обсяг та різноманітність інформації, яка міститься у таких статтях, використання методів NLP стає ключовим для автоматизації аналізу, класифікації та рекомендації відповідно до індивідуальних потреб користувачів.

В нашій роботі ми обрали використання методів векторного представлення слів (Word Embeddings) та моделей класифікації тексту для вирішення проблеми холодного старту в рекомендаційних системах [3]. Цей вибір є обґрунтованим з декількох причин.

По-перше, методи векторного представлення слів, такі як Word2Vec, дозволяють перетворити слова або фрази у вектори числового представлення у векторному просторі. Це дозволяє моделі отримувати інформацію про семантичні зв'язки між словами та їхнім контекстом в тексті [4]. Завдяки цьому, ми можемо врахувати семантичну схожість між об'єктами, навіть якщо немає даних про їхню взаємодію.

По-друге, моделі класифікації тексту дозволяють аналізувати та класифікувати тексти за їхнім змістом або семантикою. Це може бути корисним для визначення схожості між новими об'єктами та існуючими у системі, що допоможе у побудові рекомендацій [5].

В рамках нашого дослідження ми плануємо використати методи векторизації та кластеризації для аналізу технічних статей, що дозволить рекомендаційним системам ефективно вирішувати проблему холодного

старту. Таким чином, в роботі ми найбільше приділимо увагу саме холодним товарам. Очікується, що використання цих методів допоможе виявляти схожості між технічними документами та користувацькими інтересами, навіть коли експліцитні дані відсутні. Це, в свою чергу, має сприяти підвищенню точності рекомендацій для нових користувачів або продуктів, покращуючи користувацький досвід з перших кроків взаємодії з системою.

Під час реалізації цих методів ми можемо зіткнутися з декількома викликами, включаючи необхідність обробки великих обсягів текстових даних та визначення оптимальних параметрів для алгоритмів кластеризації. Існує також ризик того, що векторизація може не вловлювати всі нюанси технічних текстів, що може вплинути на якість кластеризації та, відповідно, на точність рекомендацій. Ми плануємо використовувати передові техніки обробки природної мови для мінімізації цих ризиків та оптимізації процесу векторизації.

Також важливим обмеженням є забезпечення приватності та безпеки даних користувачів під час обробки їхньої інформації. У нашому випадку ми будемо аналізувати відкриті дані. Але якщо до цього додати аналіз коментарів та відгуків, або реакцій користувачів на рекомендації, то необхідно анонімізувати дані і забезпечити високий рівень захисту інформації.

Застосування векторизації та кластеризації до аналізу технічних статей відкриває нові можливості для вирішення проблеми холодного старту в рекомендаційних системах. Ці методи дозволяють створити більш глибоке та точне розуміння інтересів користувачів та схожості продуктів, навіть коли експліцитні взаємодії відсутні. Наше дослідження має потенціал значно покращити взаємодію між користувачами та рекомендаційними системами, підвищуючи ефективність та задоволення користувачів від першого досвіду користування.

Список використаних джерел:

1. Falk K. Practical Recommender Systems. Manning Publications Co. 2019.

2. Ricci F., Rokach L., Shapira B. Recommender Systems Handbook. Third Edition. Springer, 2022.

3. Singh P. Machine Learning with PySpark, with Natural Language Processing and Recommender Systems. APRESS, 2019.

4. Sineglazov V., Savenko I. Comparative Analysis of Text Vectorization Methods. Electronics and Control Systems. 2023. Vol. 2, № 76. P. 21–27.

5. K-means Clustering Algorithms: A Comprehensive Review, Variants Analysis, and Advances in the Era of Big Data / A. M. Ikotun et al. Information Sciences, 2022.