УДК 004.852
DOI https://doi.org/10.32782/2663-5941/2025.2.2/31

Шерстюк І. О.
Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського»

Голубєв Л. П.
Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського»

РЕКОМЕНДАЦІЙНА СИСТЕМА НА ОСНОВІ УДОСКОНАЛЕНОГО АЛГОРИТМУ КОРЕЛЯЦІЇ

У статті розглядається розробка рекомендаційної системи, яка використовує кореляційний аналіз та параметри регуляризації для покращення точності рекомендацій. Рекомендаційні системи є важливим інструментом у сфері електронної комерції, потокових сервісів та соціальних мереж, оскільки вони допомагають користувачам знаходити цікаві для них товари чи контент. У роботі було розглянуто основні методи рекомендацій, зокрема колаборативну фільтрацію та кореляційний підхід. Колаборативна фільтрація є одним з найбільш популярних методів побудови рекомендацій. В її основі лежить матриця оцінок користувачами тих чи інших елементів. Кореляційний метод направлений на визначення оцінок для пропущених значень на основі схожості уподобань користувачів. В статті висвітлено проблеми, пов'язані з розрідженістю даних, і шляхи їх подолання. Основною проблемою традиційних методів є розрідженість матриці оцінок, що ускладнює знаходження подібних користувачів та об'єктів. Для вирішення цієї проблеми застосовано регуляризацію, яка допомагає зменшити вплив розрідженості на точність рекомендацій. У ході експериментального дослідження було перевірено ефективність регуляризації на синтетичних наборах даних. Аналіз показав, що застосування регуляризації покращує якість рекомендацій, особливо для даних із великою кількістю пропущених значень. Отримані результати показують, що, чим більш розріджена матриця, тим значніший позитивний ефект від регуляризації. Також було проведене дослідження залежності якості рекомендацій від значення параметра регуляризації. Реалізовано алгоритми обчислення подібності між користувачами та рекомендації на основі їх уподобань. Отримані результати підтверджують доцільність використання регуляризованого методу для побудови рекомендаційних систем. Запропоноване рішення дозволяє підвищити точність рекомендацій без значного ускладнення обчислювальної моделі.

Ключові слова: *рекомендаційна система, кореляційні методи, регуляризація, колаборативна фільтрація.*

Постановка проблеми. Рекомендаційні системи є важливим інструментом для персоналізації контенту та підвищення зацікавленості користувачів. Однак традиційні методи, такі як колаборативна фільтрація, стикаються з проблемою розрідженості матриці оцінок, що ускладнює точність прогнозів. Методи машинного навчання можуть обробляти такі дані, але їхня висока обчислювальна складність обмежує застосування в малих бізнесах. Кореляційні методи є більш доступними та простими у використанні, проте їхня точність потребує вдосконалення.

Також однією з переваг методів колаборативної фільтрації є те, що йому не потрібно для розрахунків додаткової інформації про користувача, оскільки ця інформація не змінює та не впливає на результати рекомендацій. Це робить колаборативну фільтрацію більш гнучким методом, ніж інші.

Основним недоліком методу є те, що для генерації результату прогнозування використовується загальна матриця оцінки елементів. У реальних рекомендаціях зазвичай найактивніші користувачі оцінюють дуже обмежений відсоток елементів у порівнянні з доступною загальною кількістю. Це призводить до розрідженості матриць елементів користувача, неможливості знайти успішних сусідів і, нарешті, створення нерелевантних рекомендацій

Враховуючи ці аспекти, розробка ефективної, зручної та продуктивної рекомендаційної системи є актуальним завданням.

Аналіз останніх досліджень і публікацій. Техніка колаборативної фільтрації спирається на описи продуктів і профілі користувачів. Даний підхід включає різні методики розрахунків, а його результати застосовуються для побудови рекомен-

дації для фільмів, туристичних пам'яток тощо [1], [2]. Однак в більшості досліджень вказується, що відсутність великої кількості оцінок може суттєва погіршити рекомендації. Для цього пропонується використовувати різні техніки. В роботі [3] застосовується додаткова інформація щодо переваг користувачів. В роботах [4, 5] застосовуються різні методи машинного навчання або графові методи, які можуть забезпечити кращу точність рекомендацій. Однак, в будь-якому разі, ці методи потребують додаткової інформації та/або додаткових обчислювальних ресурсів. Тому задача більш простих методів без залучення додаткової інформації потребує подальшого дослідження.

Постановка завдання. Метою дослідження є вдосконалення алгоритму рекомендацій за рахунок використання регуляризації для коригування впливу відсутності значної кількості оцінок.

Виклад основного матеріалу. Рекомендаційні системи – це програмні рішення, що прогнозують, які об'єкти (фільми, музика, книги, новини, веб-сайти, товари) можуть зацікавити користувача, спираючись на інформацію про його вподобання. Вони надають рекомендації в різних ситуаціях, наприклад, коли користувач обирає серед великої кількості варіантів або прагне отримати персоналізовані пропозиції [6].

Виділяють чотири ключові функції рекомендаційних систем [7]:

– підтримка у прийнятті рішень шляхом прогнозування оцінок користувача для певних товарів;

– допомога у порівнянні, що забезпечує персоналізоване ранжування списку доступних елементів;

– сприяння у пошуку, надаючи користувачеві рекомендації щодо нових об'єктів, які можуть його зацікавити;

– розширення можливостей дослідження шляхом пропонування елементів, подібних до обраного користувачем.

Система рекомендацій на основі колаборативної фільтрації використовують міру схожості між користувачами. В основі даного підходу лежить така ідея: пошук групи або колекції користувачів X, чиї вподобання та антипатії подібні до уподобань користувача A. Нові елементи, які подобаються більшості користувачів у X, рекомендуються користувачеві A (рис. 1) [8].

Ефективність побудови рекомендацій залежить від того, на скільки точно алгоритм може знайти околиці цільового користувача, тобто визначити групи користувачів зі спільними уподобаннями. В основу методу покладені три правила [6]:

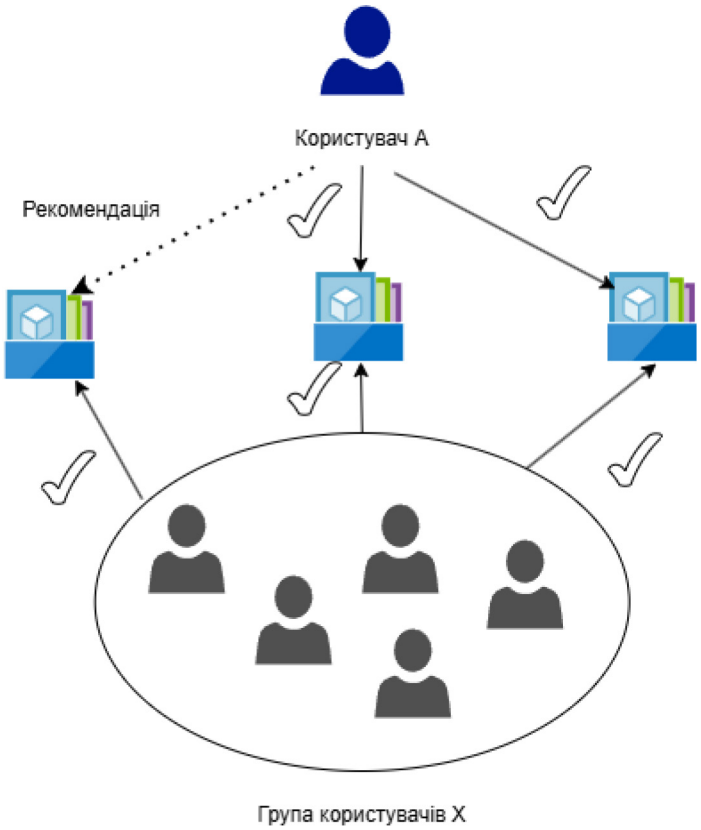

Рис. 1. Принцип роботи алгоритму колаборативної фільтрації

– люди мають однакові переваги та об'єкти своєї уваги,

– їхні переваги та об'єкти уваги є стабільними,

– їхній вибір можна визначити, досліджуючи їх історичні переваги.

Таким чином, алгоритм пошуку рекомендацій, побудований на діях користувачів, займається пошуком сусідів для кожного користувача та прогнозуванні його інтересів згідно інтересів чи переваг його сусідів.

Розглянемо загальну постановку задачі колаборативної фільтрації. Нехай рекомендаційна система визначається такими множинами [9]:

множина користувачів $U = \{u_1, u_2, \ldots, u_N\}$,

множина елементів $V = \{v_1, v_2, \ldots, v_M\}$,

матриця оцінок $R = \{r_{ij}\}_{N \cdot M}$, $i = 1, 2, \ldots, N, j = 1, 2, \ldots, M$. Потрібно зауважити, що не всі елементи даної матриці заповнені.

Алгоритм рекомендаційної системи повинен спрогнозувати значення пропущених елементів матриці на основі вже існуючих елементів матриці.

Загалом, колаборативна фільтрація використовує шкалу схожості для пошуку активних сусідів користувачів і кандидатів у спільні елементи. Для цього використовуються різні методи схожості: кореляційні моделі та методи на основі відстані між елементами [10,11].

Кореляційний метод у рекомендаційних системах ґрунтується на виявленні залежностей між вподобаннями користувачів або характеристиками об'єктів на основі статистичних показників. Основна ідея цього підходу полягає у визначенні схожості між користувачами шляхом обчислення коефіцієнта кореляції, такого як коефіцієнт Пірсона або косинусна міра подібності.

Кореляція Пірсона обчислює кореляцію між двома елементами та визначає їхню схожість на основі моделей їх сумісної появи. Коефіцієнт кореляції Пірсона (r) є мірою міцності лінійного зв'язку між двома змінними. Він коливається від –1 до 1, де 1 означає ідеальну позитивну кореляцію, –1 означає ідеальну негативну кореляцію, а 0 означає відсутність кореляції. Коефіцієнт кореляції Пірсона використовується для порівняння схожості між двома об'єктами.

$$P(u,v) = \frac{\sum_{i \in I} (r_{ui} - \overline{r}_u)(r_{vi} - \overline{r}_v)}{\sqrt{\sum_{i \in I} (r_{ui} - \overline{r}_u)^2 (r_{vi} - \overline{r}_v)^2}} \qquad (1)$$

У випадку колаборативної фільтрації метод дозволяє визначати користувачів зі схожими уподобаннями та рекомендувати об'єкти, які вони оцінили позитивно, іншим користувачам зі схожим профілем.

При застосуванні кореляційного методу потребує врахування низки викликів, таких як проблема розрідженості матриці оцінок та можливий вплив шуму в даних.

Для побудови рекомендаційної системи пропонується використовувати регуляризацію. Удосконалення кореляційного методу за допомогою регуляризації дозволяє підвищити його стабільність та точність. Регуляризація дає змогу зменшити вплив проблем кореляційного методу, додаючи штрафний елемент до розрахунку кореляції, що забезпечує стабільність коефіцієнта, особливо за наявності малої кількості даних.

Формула регуляризованого коефіцієнта кореляції Пірсона має вигляд

$$r_{reg} = \frac{\sum_{i=1}^{n} w_i (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n} w_i (x_i - \overline{x})^2} \sqrt{\sum_{i=1}^{n} w_i (y_i - \overline{y})^2} + \lambda}, \qquad (2)$$

де λ –параметр регуляризації, який контролює рівень штрафу за невелике число спостережень; n – кількість спільних об'єктів (наприклад, рейтингів користувачів).

Регуляризація дозволяє «згладжувати» значення кореляції для пар із невеликою кількістю спостережень, зменшуючи їх вплив на рекомендації.

Значення λ підбирається експериментально, залежно від характеру даних.

Для наборів даних із високим рівнем шуму рекомендується більший λ, щоб зменшити вплив екстремальних значень. У випадках, коли більшість користувачів взаємодіє з великою кількістю продуктів, можна використовувати менший λ. Уявімо двох користувачів, які оцінюють лише кілька спільних продуктів. У класичному підході їхній коефіцієнт кореляції може бути високим, навіть якщо ці оцінки випадкові. Регуляризація зменшує цей коефіцієнт, відображаючи недостатність даних для точного визначення схожості.

Регуляризація має певні переваги. Розглянемо їх детальніше.

Вона забезпечує стабільність результатів, тобто зв'язки для пар із малою кількістю спостережень мають зменшену оцінку. Зменшується вплив аномалій, штрафний параметр компенсує вплив екстремальних значень.

Регуляризація може легко поєднуватися з іншими методами вдосконалення, такими як зважування чи кластеризація.

Удосконалений кореляційний метод із регуляризацією забезпечує кращу точність рекомендаційних систем, особливо для даних із малою кількістю спостережень чи високою варіативністю. Таке вдосконалення дозволяє уникати недоліків класичного підходу, зберігаючи його простоту та інтуїтивність.

Для дослідження запропонованого методу було проведено експеримент на основі згенерованих даних. Це дало б змогу з'ясувати, чи покращується рекомендація при використанні регуляризованого методу. При проведенні експерименту досліджувалось різні вхідні параметри: розмір матриці рекомендацій, заповненість даної матриці даними.

Для оцінки рекомендацій використовувались такі міри якості [10, 12].

MAE (Mean Absolute Error) визначає середню похибку між передбаченими та реальними рейтингами.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|r_i - \hat{r}_i\right|, \qquad (3)$$

де N – кількість прогнозованих рейтингів; r_i – реальний рейтинг; $\hat{r}_i$ – передбачений рейтинг

Менші значення MAE означають кращу якість рекомендацій.

RMSE (Root Mean Squared Error) вимірює середньоквадратичну похибку передбачених рейтингів.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(r_i - \hat{r}_i)^2}. \qquad (4)$$

Даний показник є більш чутливим до великих помилок на відміну від MAE.

Перш за все досліджувалось вплив загальної кількості користувачів на точність рекомендацій. У рамках експерименту було обрано різні розміри матриці оцінок, змінюючи кількість користувачів у діапазоні від 15 до 100. Кількість товарів змінювалася пропорційно, щоб підтримувати прийнятне співвідношення між користувачами та об'єктами.

Збільшення кількості користувачів дає змогу оцінити роботу алгоритму за умов підвищеного навантаження та перевірити, чи зберігається стабільність рекомендацій. Водночас аналіз при малих значеннях U дозволяє визначити ефективність методу в ситуаціях із обмеженою кількістю даних. Це має ключове значення для порівняння різних підходів та вибору оптимальних параметрів моделі.

Результати експерименту, проведеного за співвідношенням користувачів до об'єктів оцінювання 2:1, наведені на рис. 2 (за умови відсутності 30 % рекомендацій).

Отримані результати показують, що зі збільшенням кількості користувачів, але збереженням співвідношення між користувачами та об'єктами оцінювання, регуляризований метод демонструє дещо кращу ефективність порівняно з методом Пірсона. Значення критерію MAE для удосконаленого методу знаходиться в діапазоні від 1,35 до 1,66, а RMSE змінюється в межах від 1,75 до 2,15. Це свідчить про стабільність методу при формуванні рекомендацій.

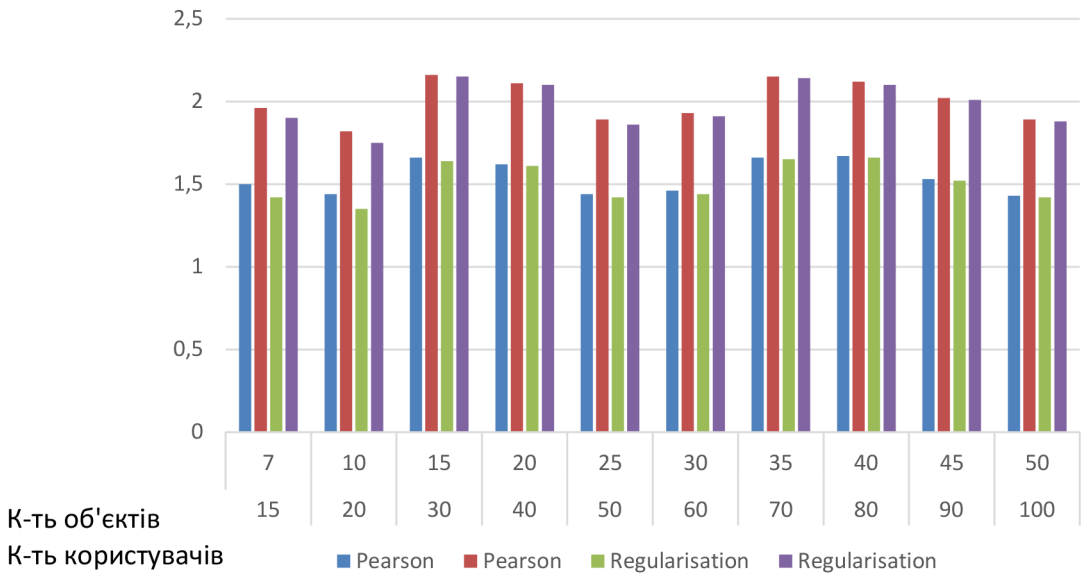

Рис. 2. Порівняння критеріїв оцінки рекомендацій для кореляції Пірсона та удосконаленого методу для різної кількості користувачів

Дослідження роботи методів при зміні співвідношення між кількістю користувачів та кількістю об'єктів оцінювання (рис. 3) показало, що вдосконалений метод перевершує метод Пірсона в середньому на 1,27 в.п. Крім того, варто зазначити, що зі збільшенням кількості оцінюваних об'єктів похибка рекомендаційної системи зменшується. Це пояснюється тим, що система має ширший набір даних, що дає змогу точніше визначати схожість між користувачами та формувати більш релевантні рекомендації.

Подальший аналіз удосконаленого методу був спрямований на зміну загального відсотку пропущених оцінок (рис. 4). Експеримент відбувався при кількості користувачів 50, та предметів для оцінювання – 10.

Проведений експеримент показує, що удосконалений метод покращує визначення рекомендацій у всіх випадках. Найкращі результати узагальнений метод показує при частці пропущених оцінок від 0,3 до 0,55. При майже заповненій матриці рекомендацій обидва методи працюють достатньо добре. Також якість рекомендації знижується при збільшенні частки пропущених оцінок до 60 %.

Отже, удосконалений метод ефективніше відновлює пропущені значення та враховує додаткову інформацію, що сприяє точнішому формуванню рекомендацій. Натомість метод Пірсона виявляється менш стійким до неповноти даних, оскільки класичний підхід до кореляції не дозволяє ефективно обробляти пропущені значення.

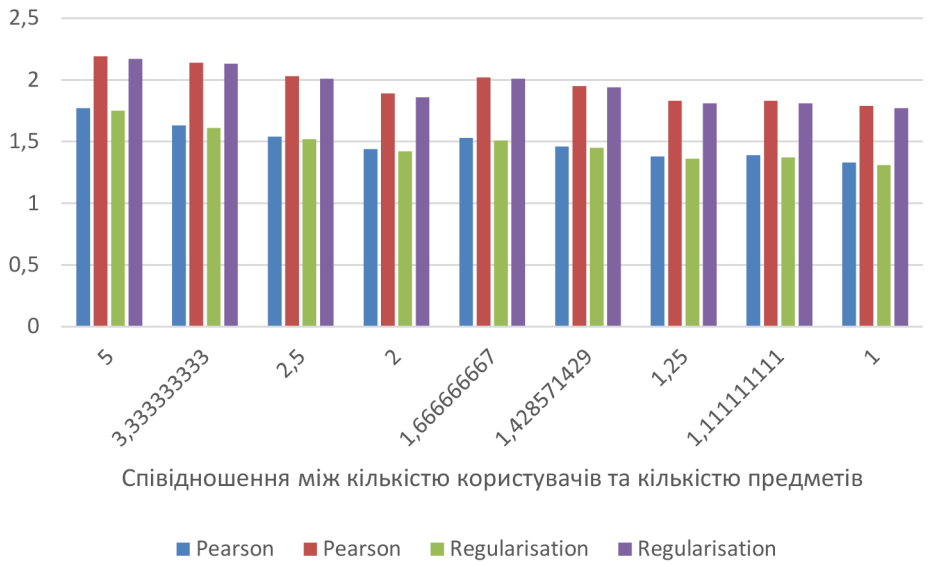

Рис. 3. Порівняння критеріїв оцінки рекомендацій для кореляції Пірсона та удосконаленого методу для різних співвідношень між користувачами та об'єктами для оцінки

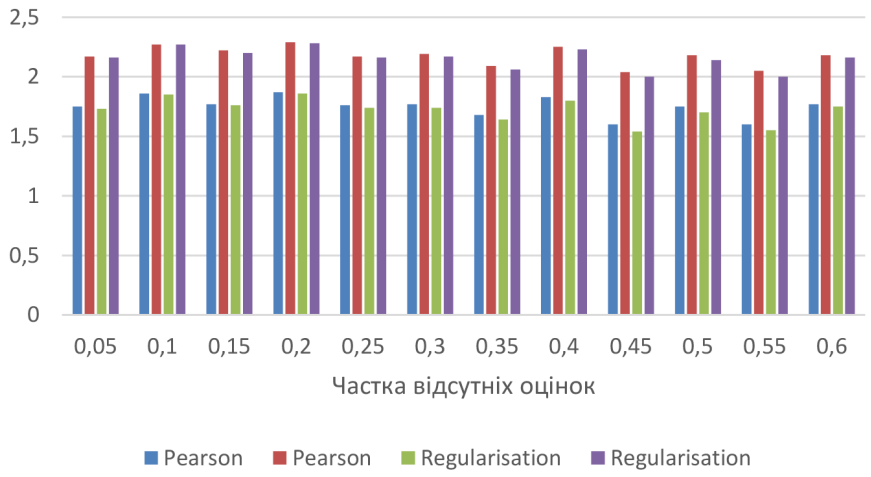

Рис. 4. Порівняння критеріїв оцінки рекомендацій для кореляції Пірсона та удосконаленого методу для різної частки відсутніх оцінок

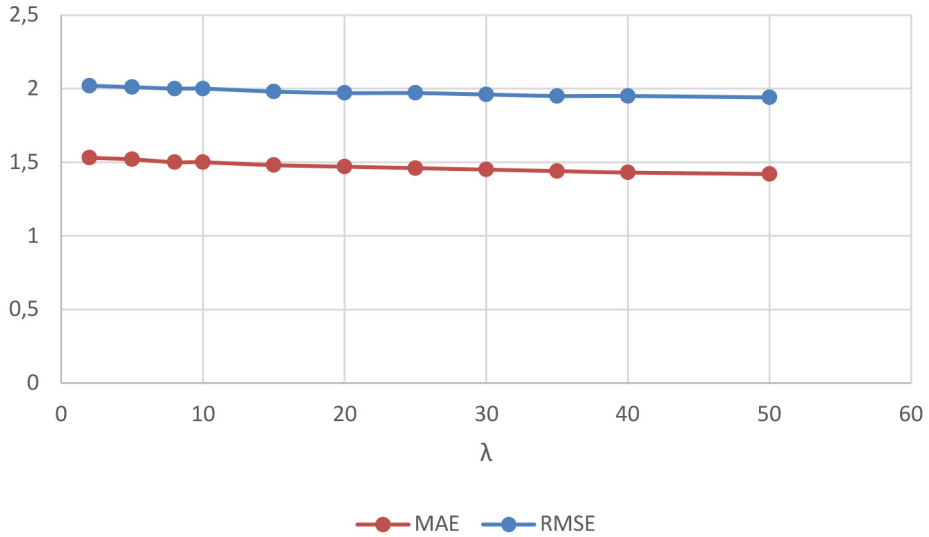

Рис. 5. Залежність похибок прогнозу від значення коефіцієнта регуляризації

При майже заповненій матриці обидва методи демонструють схожу ефективність, оскільки наявність достатньої кількості даних забезпечує точніші обчислення. У такій ситуації різниця між підходами мінімізується, адже обидва методи оперують майже повною інформацією для визначення подібності між користувачами чи товарами. Однак, коли частка пропущених оцінок перевищує 55 %, якість рекомендацій значно погіршується. Це пов'язано з тим, що розріджена матриця містить недостатньо інформації для точних прогнозів. Навіть удосконалені методи в таких умовах стикаються з проблемою обмеженої кількості зв'язків між користувачами та товарами, що ускладнює побудову надійних рекомендацій.

Порівняємо результати роботи обох методів для різних значень параметра регуляризації. Порівняння будемо здійснювати для кількості пропущених значень 30 % та для 50 користувачів з 20 об'єктами для рекомендацій (рис. 5).

Отримані результати свідчать про те, що коефіцієнт регуляризації покращує результати оцінки рекомендацій. В той же час при великих значеннях даного коефіцієнту результати роботи методу практично не покращуються.

Таким чином, проведені експерименти засвідчують, що удосконалений метод в усіх випадках показує кращі результати порівняно з класичним методом кореляції Пірсона. В той же час найкращі результати досягаються при частці відсутніх оцінок від 0,3 до 0,55 та коефіцієнті регуляризації від 10 до 40. Для інтернет-магазинів має сенс використовувати регуляризовану кореляцію, оскільки вона більш точно передбачає рекомендації при відсутності значної кількості оцінок.

Висновки. У ході дослідження було проаналізовано та експериментально перевірено ефективність удосконаленого кореляційного методу. Отримані результати показали, що регуляризація дозволяє підвищити якість рекомендацій, особливо у випадку розріджених матриць. Це зумовлено тим, що класичний метод Пірсона чутливий до відсутніх значень, тоді як удосконалений підхід враховує додаткову інформацію, що забезпечує стабільність результатів. Незважаючи на загальну ефективність удосконаленого методу, при значній кількості пропущених значень (понад 55 %) спостерігається помітне зниження точності рекомендацій. Власне усе зазначене вказує на необхідність додаткових механізмів обробки розріджених даних, таких як комбінування різних методів або застосування нейронних мереж.

Список літератури:

1. Anwar T., Uma V. Comparative study of recommender system approaches and movie recommendation using collaborative filtering. *Int. J. Syst. Assurance Eng. Manage.*, vol. 12, no. 3, pp. 426–436, Jun. 2021, doi: 10.1007/s13198-021-01087-x

2. Campos P., Pinto E., Torres A. Rating and Perceived Helpfulness in a Bipartite Network of Online Product Reviews. Cham, Switzerland:Springer, 2023, doi: 10.1007/s10660-023-09725-1

3. Haruna K., Akmar I. M., Damiasih D., Sutopo J., Herawan T. A Collaborative Approach for Research. *Paper Recommender System. PLoS ONE*. 12 (10), e0184516 (2017). https://doi.org/10.1371/journal.pone.0184516

4. Mallouk I., Abou el Majd B., Sallez Y.A generic model of the information and decisional chain using Machine Learning based assistance in a manufacturing context. *Mathematical Modeling and Computing*. 10(4), 1023–1036 (2023).

5. Lin K., Sonboli N., Mobasher B., Burke R. Calibration in collaborative filtering recommender systems: a user-centered analysis. *Proceedings of the 31st ACM Conference on Hypertext and Social Media*, pp. 197–206, 2020.

6. Meyer F. Recommender systems in industrial contexts. Thèse de Doctorat, University of Grenoble, France. 6, 21. 2012.

7. Le W., Cao J., Wu J. Huang C., Buyya R. A collaborative filtering recommendation method based on discrete quantum-inspired shuffled frog leaping algorithms in social networks. *Future Generation Computer Systems*. Volume 88, November 2018. https://doi.org/10.1016/j.future.2018.05.008

8. Li Z., Huang M., Zhang Y. A collaborative filtering algorithm of calculating similarity based on item rating and attributes. *Proc. – 2017 14th Web Inf. Syst. Appl. Conf*. WISA 2017, vol. 2018–Janua, pp. 215–218, 2018.

9. Mahesh T.R., Vivek V., Kumar V.V. Recommendation Systems: Different Techniques, Challenges and Future Directions. *International Journal of Information Technology, Research and Applications*, 2022. vol. 1, no. 2, pp. 1–8.

10. Hassanieh L., Abdo J.B., Demerjian J. Similarity Measures for Collaborative Filtering Recommender Systems, 2018.

11. Wang Y., Deng J., GaoJ., Zhang P. A hybrid user similarity model for collaborative filtering. *Inf. Sci. (Ny)*., 2017. vol. 418–419, pp. 102–118. https://doi.org/10.1016/j.ins.2017.07.040

12. Kumar P., Gupta M., Rao C., Bhavsingh M., Srilakshmi M. A Comparative Analysis of Collaborative Filtering Similarity Measurements for Recommendation Systems. *International Journal on Recent and Innovation Trends in Computing and Communication*. 2023. 11. 184-192. 10.17762/ijritcc.v11i3s.6180

Sherstiuk I.O., Holubiev L.P. RECOMMENDATION SYSTEM BASED ON AN IMPROVED CORRELATION ALGORITHM

The article discusses the development of a recommendation system that uses correlation analysis and regularization parameters to improve the accuracy of recommendations. Recommender systems are an important tool in the field of e-commerce, streaming services, and social networks, as they help users find products or content that they are interested in. The paper considered the main methods of recommendations, in particular, collaborative filtering and the correlation approach. Collaborative filtering is one of the most popular methods for building recommendations. It is based on a matrix of user ratings of certain elements. The correlation method is aimed at determining ratings for missing values based on the similarity of user preferences. The article highlights the problems associated with data sparsity and ways to overcome them. The main problem of traditional methods is the sparsity of the rating matrix, which makes it difficult to find similar users and objects. To solve this problem, regularization was used, which helps to reduce the impact of sparsity on the accuracy of recommendations. During the experimental study, the effectiveness of regularization was tested on synthetic datasets. The analysis showed that the use of regularization improves the quality of recommendations, especially for data with a large number of missing values. The results obtained show that the more sparse the matrix, the greater the positive effect of regularization. The dependence of the quality of recommendations on the value of the regularization parameter was also studied. Algorithms for calculating the similarity between users and recommendations based on their preferences were implemented. The results obtained confirm the feasibility of using the regularized method for building recommendation systems. The proposed solution allows you to increase the accuracy of recommendations without significantly complicating the computational model.

Key words: *recommendation system, correlation methods, regularization, collaborative filtering.*