8.**Козаков А.**ADEM – CAD/CAM-интеграциявысокогоуровня / А.Козаков, А.Красильников // САПР і графіка. – 2003 – №7 – С. 38-44.

9.**Норенков І.П.** Основы САПР [Електронний ресурс]: електронний підручник / І.П. Норенков, В.А. Трудоношин, М.Ю. Уваров, Е.В. Федорук. – Москва: МГТУ, 2010. – Режим доступу: http://bigor.bmstu.ru/.

10.**Квєтний Р.Н.** Комп'ютерне проектування систем і процесів [Електронний ресурс]: посібник / Р.Н. Квєтний, І.В. Богач, О.Р. Бойко, О.Ю. Софіна, О.М. Шушура. – Режим доступу: posibnyky.vntu.edu.ua/k_m/t1/173.htm.

11.**Ступницький В.В.** Ефективність впровадження CALS-технологій на машинобудівних підприємствах України / В.В. Ступницький // Вісник Національного університету «Львівська політехніка» «Оптимізація виробничих процесів і технічний контроль у машинобудуванні та приладобудуванні». - 2009.- №642. - С. 80-84.

12.**Криськов О.Д.** САПР операцій механічної обробки. Математичне моделювання технологічних процесів / Криськов О.Д. – Кіровоград: КНТУ, 2004. – 75с

УДК 004.852

В. О. МІТРОШИН*, студ., Н. Н. ШАПОВАЛОВА, І. О. ДОЦЕНКО, ст. викладачі,
Н. Х. САІТГАРЕЄВ, доц.
Криворізький національний університет

МОДЕЛЬ ПЕРСОНАЛІЗАЦІЇ РЕКОМЕНДАЦІЙ КОНТЕНТУ НА ОСНОВІ КОЛАБОРАТИВНОЇ ФІЛЬТРАЦІЇ

Мета роботи – розробити і теоретично обґрунтувати ефективність застосування персоналізованої рекомендаційної системи товарів, послуг або контенту на основі технології машинного навчання, яка поєднує в собі такі підходи до персоналізації рекомендацій, як колаборативна фільтрація та фільтрація контенту. Розробити механізм визначення доцільності використання певної метрики виявлення подібності користувачів і реалізувати його у режимі «реального часу» на розроблювальній системі. Створити безпечний сервіс з персоналізованою системою рекомендацій товарів, послуг або контенту, забезпечивши захист особистих даних користувачів.

Методи дослідження. У роботі використано наступні методи дослідження: аналіз джерел з досліджуваної теми, метричні методи визначення приналежності об'єктів до певної групи за їх схожістю, методи теорії штучного інтелекту, моделювання процесу навчання алгоритмів класифікації, формалізація побудованих моделей, методи проектування програмного забезпечення для розробки програмної моделі, емпіричні методи обґрунтування оптимальних параметрів навчання моделі, методи об'єктно-орієнтованого проектування та програмування.

Наукова новизна полягає в тому, що розроблена модель рекомендаційної системи дозволяє на початковому етапі роботи системи оцінити вподобання користувача, майже не володіючи інформацією про його прихильності. В роботі запропоновано механізм підбору певної метрики для визначення групи приналежності користувача, що в цілому має забезпечити точність наданої рекомендації.

Практична значимість виконаної роботи полягає в тому, що розроблено веб-сервіс, який надає персоналізовані рекомендації товарів, послуг або контенту щодо уподобань користувача, з забезпеченням вимог захисту персональної інформації, високої точності виконаних рекомендацій. Модуль рекомендаційної системи оформлено у вигляді прикладного програмного інтерфейсу, що дозволяє його застосування на будь-якій веб-платформі.

Результати. Запропоновано особистісно-орієнтований підхід до рекомендацій товарів, послуг або контенту, розроблено універсальну рекомендаційну систему, яка поєднує в собі і колаборативну, і фільтрацію контенту, а також розроблено веб-сервіс, з урахуванням вимог безпеки персональних даних користувачів.

Ключові слова: Рекомендаційна система, машинне навчання, метричний метод, метрика, колаборативна фільтрація, фільтрація контенту.

doi: 10.31721/2306-5451-2021-1-52-142-146

Проблема та її зв'язок з науковими і практичними задачами. В даний час кожна сучасна людина щодня стикається з різноманітними веб-сайтами та соціальними мережами. Будь-які онлайн-сервіси, зокрема постачальники фільмів і серіалів, товарів та послуг, надаючи свій контент, прагнуть максимально догодити користувачам. Такі ресурси автоматично редагують свої зміст і наповнення, налаштовуючись до вподобань кожного конкретного користувача. Одним з рішень задач персоналізації контенту є рекомендаційні системи. Рекомендаційні системи – це програмні модулі, які аналізують інтереси користувачів і намагаються передбачити, який саме набір фільмів або товарі і послуг буде найцікавіший конкретному користувачеві в даний момент часу, тим самим підвищуючи рівень його лояльності.

Вдалі рекомендації збільшують рівень задоволення користувачів, що в свою чергу підвищує їх лояльність, що в кінцевому рахунку призводить до зростання прибутку. Таким чином, одним з найбільш частих запитів на розробку заснованих на машинному навчанні маркетингових рішень, є розробка рекомендаційних моделей. Яскравою ілюстрацією актуальності розробки рекомендаційних систем є історія компанії Netflix і її конкурс Netflix Prize [1]. У 2006 році компанія Netflix володіла унікальним алгоритмом Cinematch [2]. Він дозволяв рекомендувати клієнтам фільми не за загальним середнім рейтингом, а робити це індивідуально. Але в Netflix хотіли піти ще далі.

Конкурс Netflix Prize було розпочато в 2006 році і тривав він протягом майже трьох років. Його суть полягала в підвищенні якості існуючого алгоритму на 10%. Учасниками могли бути всі бажаючі розробники з усього світу, а призовий фонд становив 1000000 $. До тих же пір, поки не був завойований головний приз, щорічно кращій команді присуджувався приз за прогрес в розмірі 50000 $. Остаточну перемогу в 2009 році здобула команда BellKor's Pragmatic Chaos, розділивши перше місце з командою The Ensemble, але випередивши її у відправці рішення по часу. Випередження склало всього 20 хвилин [3].

Аналіз досліджень і публікацій. Для створення рекомендаційних систем існують дві основні стратегії: фільтрація контенту і колаборативна фільтрація. У разі фільтрації контенту створюються профілі користувачів і об'єктів. Профілі користувачів і об'єктів представляють собою суху характеристику користувачів і товарів. Даний підхід є найпростішим, але він формує тільки неперсоналізовані рекомендації [4].

У разі ж колаборативної фільтрації, використовується інформація про минулу поведінку користувачів. Наприклад, інформація про вподобання чи оцінки будь-якого товару, фільму тощо. У такому випадку не має значення, з якими типами об'єктів ведеться робота, але при цьому до уваги беруться неявні характеристики, які складно враховувати в момент створення профілю. Колаборативна фільтрація є одним з найефективніших засобів створення рекомендацій, що складається з трьох етапів: збір інформації про користувачів, побудова матриці для розрахунку асоціацій і формування імовірної рекомендації [5]. Саме на неї і впав остаточний вибір для вирішення поставленого завдання, а для її реалізації був використаний один з методів машинного навчання – метод *k*-найближчих сусідів.

Постановка завдання. Проектування веб-сервісу з рекомендаційною системою складається з реалізації наступних завдань: розробка веб-сайту й наповнення его контентом; розробка модуля рекомендаційної системи, яка в свою черга включає етап автоматизованого аналізу даних, підбору оптимального типу метрики для реалізації методу визначення групи вподобань певного користувача, безпосередньо реалізація алгоритму машинного навчання для визначення прогнозованої рейтингової оцінки фільму, захист персональних даних користувачів; підбір контенту з урахуванням прогнозованого рейтингу.

Викладення матеріалу та результати. Метод *k*-найближчих сусідів (*k* Nearest Neighbor, *k*-NN) є метричними методом, тобто для прогнозування цільового значення об'єкта використовується поняття метрики – відстані від об'єкта до найближчих його сусідів по їх ознаковому опису. Ідея методу полягає в пошуку таких найближчих об'єктів. Після їх виявлення новий об'єкт буде віднесений до того ж класу, що і його сусіди. У разі реалізації завдання регресії, значення прогнозованої величини розраховується як середнє значення сусідів**.** Зважений метод *k* сусідів дає більш точні оцінки. В якості параметрів (ваг) такого методу можна використовувати функцію від відстані до об'єктів або функцію від номера найближчих сусідів**.**

Виникає питання як визначити ступінь близькості між об'єктами. Сусідство об'єктів визначається за допомогою метрики – функції, яка задає відстань в метричному просторі. Таких функцій існує безліч, найбільш відомими з них є Евклідова метрика, метрика Мінковського і манхеттенська відстань. На площині і в просторах невеликої розмірності доцільно використовувати Евклідову метрику. Однак, із зростанням розмірності простору ознак опису об'єктів, визначити відстань стає все складніше, тому що простори з високою розмірністю схильні до так званого «прокляття розмірності». Цей термін характеризує сукупність труднощів, що виникають при обробці інформації в просторах багатьох вимірів. А саме з такими даними, як правило, доводиться працювати в реальних умовах. Якщо виходити з поставленого завдання, то опис кожного користувача становить кількість вимірів, рівну кількості представлених на платформі фільмів. Не важко здогадатися, що це дуже велике число вимірювань, причому, матриця характери-

стик в достатній мірі розріджена. Таким чином, необхідно визначити яку метрику доцільно використовувати для вирішення поставленого завдання.

Оскільки нас цікавить не стільки відстань або ступінь віддаленості об'єктів один від одного, скільки міра їх схожості, близькості по ознаковому опису, будемо використовувати показник, що описує силу статистичного зв'язку – коефіцієнт кореляції Пірсона (1). Очевидно, що якщо зв'язок між випадковими користувачами значний, то можна прогнозувати значення рейтингових оцінок фільмів для того користувача, який ці фільми ще не дивився

$$w_{i,j} = \frac{\sum_a (r_{i,a} - \bar{r}_i)(r_{j,a} - \bar{r}_j)}{\sqrt{\sum_a (r_{i,a} - \bar{r}_i)^2} \sqrt{\sum_a (r_{j,a} - \bar{r}_j)^2}}, \quad (1)$$

де i – індекс користувача, для якого розраховується рейтингова оцінка; j – індекс користувача, по оцінках якого розраховується рейтингова оцінка; a – індекс фільму; $r_{i,a}$ – оцінка i-го користувача за a-ий фільм; $\bar{r}_i = \frac{1}{N_i} \sum_a r_{i,a}$ – середні оцінки i-го користувача; N – кількість фільмів, що продивився i-ий користувач.

Для реалізації колаборативної фільтрації використовуються два різних підходи: Item-Based і User-Based. Item-Based заснований на пошуку схожих об'єктів – в нашому випадку фільмів (в датасеті вони розташовані у колонках) [6]. На противагу цьому User-Based заснований на пошуку схожих користувачів, тобто по рядках. Підсумкове підсумовування ведеться лише за тими фільмами, які дивилися обидва користувачі.

Узагальнюючи вищесказане, можемо визначити послідовність процесу формування рейтингу фільму і прийняття рішення про рекомендацію його користувачеві: перетворити дані в векторну модель User-Based, отримуючи тим самим вподобання всіх зареєстрованих користувачів у вигляді дійсних значень оцінок; розрахувати коефіцієнти кореляції з кожним користувачем; сортувати користувачів по спадаючій міри близькості; використовуючи метод k-NN, знаходимо рейтингову оцінку для кожного користувача по заданому фільму за формулою (2). Тут сума у чисельнику і знаменнику підраховується лише по кількості обраних сусідів

$$\hat{r}_{i,a} = \bar{r}_i + \frac{\sum_{j \in kNN(i)} (r_{i,a} - \bar{r}_j) w_{i,j}}{\sum_{j \in kNN(i)} |w_{i,j}|}. \quad (2)$$

Навчання методу полягає в налаштуванні параметра k – кількості сусідів, за якими буде будуватися прогноз [8]. Помилка отриманої моделі оцінюється метрикою RMSE [7] – корінь з суми квадратів різниць між прогнозованою і істиною оцінкою (3)

$$RMSE = \sqrt{\sum_{i=1}^{N} (\hat{r}_{i,a} - r_{i,a})^2}. \quad (3)$$

Щоб уникнути явища перенавчання [8], коли модель буде підлаштовуватися під дані, що знаходяться в навчанні, і робити невірні прогнози для об'єктів, яких в навчанні не було, використовується стратегія крос-валідації. Вибірка об'єктів ділиться на навчальну і тестову в пропорції 3:1 (75% і 25% вибірки відповідно). У свою чергу, на навчальній вибірці проводиться процедура крос-валідації [9]. Ця процедура складається з наступних етапів: навчальна вибірка ділиться на n рівних частин (n фолдів), і кожна з частин в процесі n ітерацій буває тестовою, інші ж – формують навчальну вибірку. На кожній з n ітерацій оцінюється якість моделі, а потім усереднюється. Отримуємо середнє значення якості моделі на n підвибірках. Контрольне ж визначення якості виконується на відкладених 25% вихідної вибірки. Таким чином підбирається параметр методу k – кількість сусідів. В ході навчання було підібрано k = 11 на вибірці з 1000 об'єктів (рис. 1). При цьому якість моделі складає 92%.

Основною технологією для реалізації системи було використаний мову програмування Python і одна з найпопулярніших бібліотек машинного навчання Scikit-learn [10].

Отримавши модуль рекомендацій і реалізувавши його для розроблюваного веб-сервісу, стикаємося з однією важливою проблемою – так званою проблемою холодного старту.

Холодний старт – це стандартна ситуація, яка виникає, коли ще не накопичено достатню кількість даних для коректної роботи рекомендаційної системи. Це може відбуватися при появі нових фільмів або фільмів, які мало переглядаються, а також при реєстрації нового користувача. Для вирішення цієї проблеми буде використаний підхід Item-Based. Ідея полягає в тому, що

при реєстрації на веб-сервісі, перші рекомендації будуть ґрунтуватися не на перевагах користувача, а на першому вподобаному фільмі, який він переглянув або за результатами невеликого опитування про улюблені фільми, який буде запропонований користувачеві. Таким чином, рекомендація буде формуватися на основі загальної інформації по характеристикам фільму: режисер, акторський склад, жанр, рік виробництва тощо. В якості джерела подібної інформації була обрана найбільша в світі база даних про кінематограф IMDb.

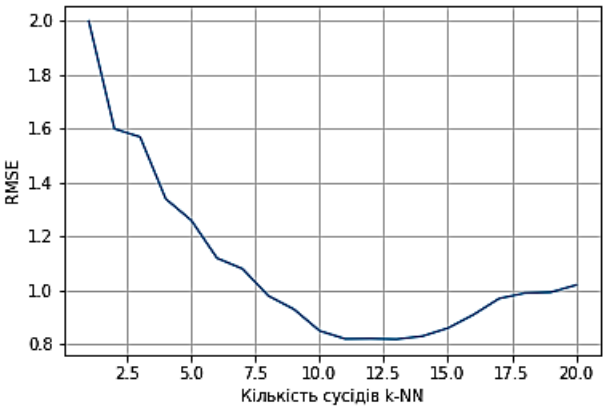

Рис. 1 Помилка моделі в залежності від кількості сусідів *k*-NN

IMDb функціонує на основі вільного програмного забезпечення [11]. Деякий час назад даний веб-сайт вивантажив набори своїх датасетів у вільний доступ на платформу Kaggle [12], ці дані були використані в роботі.

Рішення проблеми холодного старту також реалізовано метричним методом, проте в якості метрики використовується Евклідова відстань. Оскільки розмірність простору ознак не настільки велика, як в даних, які використовуються в основному модулі, це дозволяє застосувати інтуїтивно зрозумілий підхід, який легко інтерпретується. Метод зваженого *k*-NN також пройшов навчання, отримано оптимальну кількість сусідів, рівне п’яти, параметри моделі обчислювалися як функція від номера найближчого сусіда. Валідація моделі проводилася на п’яти фолдах і показала якість на тестовому наборі 86,6%. Треба зазначити, що авторами не ставилася мета домогтися максимальної якості моделі, оскільки перші рекомендації новому користувачеві не повинні складатися виключно з фільмів його «улюбленого» режисера з участю «улюбленого» актора: автори залишили за собою право привнести в рекомендацію елемент «спонтанного» вибору, дотримуючись основних ознак вподобань конкретного користувача.

Таким чином для вирішення проблеми холодного старту був розроблений додатковий модуль, який формує рекомендації за допомогою фільтрації контенту. Надалі, коли буде отримана значна власна база даних, буде проведений перехід на основний модуль, який формує рекомендації з допомогою колаборативної фільтрації.

В розробці приділено чималу увагу безпеці збереження користувацьких даних. Будемо виконувати шифрування даних на стороні клієнта, тобто у базу даних вони надходитимуть вже у зашифрованому вигляді [13]. Таким чином не потрібно окремо піклуватися про захист каналу передачі даних: якщо зловмисник перехопить пакет, то отримає повністю зашифровані дані. Недолік цього методу полягає у зниженні швидкодії обробки запиту, але, оскільки обсяг передачі не значний (лише логін і пароль), то на продуктивність веб-сервісу такий підхід не вплине. Реалізація шифрування на стороні застосунку використовує готове рішення на сервері CryptDB MySQL. Ідея полягає у використанні детермінованих шифрів (клас шифрів DET), тобто таких шифрів, які при однаковому ключі той самий текст шифрують однаково [14]. При використанні DET не виключена вразливість перед деякими видами відомих атак, тому пропонується використовувати асиметричне шифрування. Асиметричне шифрування, звісно програє по швидкодії симетричному, але при невеликих обсягах даних, можемо знехтувати цим фактом на користь більшої безпеки.

Висновки та напрямок подальших досліджень. У процесі дослідження проблеми розробки веб-сервісу з рекомендаційною системою було проаналізовано форми і підходи до реалізації фільтрації контенту, обґрунтовано вибір метрик у методах машинного навчання, що дають змогу віднайти близьких за вподобаннями користувачів і схожі за певними ознаками фільми. В результаті навчання моделей підібрано оптимальні значення гіперпараметрів методу *k*-NN: кількість сусідів та функцію ваг, які дають змогу робити рекомендації для певного користувача з достовірністю у 92%. Цей рівень якості моделі цілком задовольняє рівень достовірності того факту, що запропонований контент користувачеві сподобається. В роботі враховано умови безпечного збереження даних і конфіденційності користувачів сервісу. В подальшому планується

вдосконалити захист даних, розширити базу контенту сервісу і провести дослідження щодо налаштування рекомендацій нейромереживими методами.

Список літератури

1. The Netflix Prize Rules [Електронний ресурс] / – 2006. – Режим доступу до ресурсу: https://www.netflixprize.com/assets/rules.pdf.

2. Stephanie Crawford. How Netflix Works [Електронний ресурс] / Stephanie Crawford, Tracy V. Wilson. – 2010. – Режим доступу до ресурсу: https://electronics.howstuffworks.com/netflix2.htm.

3. Michael Jahrer. The BigChaos Solution to the Netflix Prize 2008 [Електронний ресурс] / Michael Jahrer, Andreas T¨oscher. – 2008. – Режим доступу до ресурсу: https://www.netflixprize.com/assets/ProgressPrize2008_BigChaos.pdf.

4. Bharat Bhasker. Recommender Systems in e-Commerce / Bharat Bhasker, K Srikumar. // EC '99: Proceedings of the 1st ACM conference on Electronic commerce. – 2010. – С. 158–166.

5. Fleder D., Hosanagar K. Blockbuster Culture's Next Rise or Fall: The Impact of Recommender Systems on Sales Diversity (журнал) // Management Science, Vol. 55, No. 5, May 2009, pp. 697-712. — 2009. — P. 1 - 49.

6. Muffaddal Qutbuddin Comprehensive Guide on Item Based Collaborative Filtering [Електронний ресурс] / Muffaddal. – 2020. – Режим доступу до ресурсу: https://towardsdatascience.com/comprehensive-guide-on-item-based-recommendation-systems-d67e40e2b75d.

7. Гудфеллоу Я., Бенджио И., Курвилль А. Глубокое обучение / пер. с анг. А. А. Слинкина. – 2-е изд., испр. – М.: ДМК Пресс, 2018. – 652 с.: цв. ил.

8. Сегаран Т. Программируем коллективный разум. – Пер. с англ. – СПб.: Символ-Плюс. 2008. – 368 с.

9. Введение в кросс-валидацию k-fold [Електронний ресурс] – Режим доступу до ресурсу: https://codecamp.ru/blog/cross-validation-k-fold/.

10. Introduction to scikit-learn [Електронний ресурс] - Режим доступу до ресурсу: https://neurohive.io/ru/osnovy-data-science/vvedenie-v-scikit-learn/

11. Сайт компанії IMDb – Режим доступу до ресурсу: https://www.imdb.com/

12. Kaggle [Електронний ресурс] / Режим доступу до ресурсу: https://www.kaggle.com/ashirwadsangwan/imdb-dataset

13. Luc Bouganim, Yanli Guo. Database Encryption // Encyclopedia of Cryptography and Security / Ed. by Henk C. A. van Tilborg and Sushil Jajodia. — Springer, 2011. — P. 307—312.

14. Rakesh Agrawal, Jerry Kiernan, Ramakrishnan Srikant, Yirong Xu. Order preserving encryption for numeric data (англ.) // Proceedings of the 2004 ACM SIGMOD international conference on Management of data - SIGMOD '04. — Paris, France: ACM Press, 2004. — P. 563

УДК 52-1:524.382

Ю. Д. ТКАЧЕНКО*, студ., Д. В. ШВЕЦЬ, Н.О. КАРАБУТ, старші викладачі
Криворізький національний університет

АНАЛІЗ МЕТОДІВ ВИЯВЛЕННЯ ТА СПОСТЕРЕЖЕННЯ ПОДВІЙНИХ ЗІРОК

Мета. Метою дослідження є розгляд та класифікація існуючих методів реєстрації та спостереження подвійних зірок.

Методи дослідження. Проведено системний аналіз існуючих способів виявлення подвійних систем, проаналізовано наявні на сьогодні підходи та розглянуто їх особливості.

Наукова новизна. Розглянуто типи подвійних зірок, які можна класифікувати як візуальні, що виглядають як два окремих компоненти, відносне положення яких змінюється через рух вздовж своїх орбіт, астрометричні, в яких візуально помітна тільки одна компонента зі змінним рухом, який спричиняється гравітаційним впливом другої компоненти, спектроскопічні, спектри яких виявляють регулярні зміни та фотометричні, періодичні коливання повної яскравості яких викликані рухом компонентів подвійної системи.

Практичне значення. Фізичні подвійні зірки представляють для астрономії як науки в цілому фундаментальний інтерес, який здебільшого визначається тим, що саме вивчення подвійних зірок дозволило однозначно встановити єдність закону всесвітнього тяжіння Ньютона у Всесвіті і отримати, спираючись на спостереження, фундаментальні знання про маси зірок, їх світності і еволюції.

Результати. Використовуючи існуючі на сьогодні методи виявлення подвійних зірок можливо з'ясувати велику кількість їх параметрів незалежно від типу подвійних систем. Відповідно до методу спостереження і виявлення тієї або іншої системи зірок, можна скористатись різними варіантами знаходження параметрів світил. Зміна блиску тіла системи зазвичай викликана зміною взаємного положення тіл, внаслідок їх руху по орбітах, а також обертанням тіла навколо власної осі. В останньому випадку крива блиску дозволяє встановити період обертання тіла на час спостереження. У змінних зірок зміна блиску може бути пов'язана з рухом навколо неї менш яскравої зірки-компаньйона, а також може свідчити про наявність планет коло неї. Зміни зсувів або роздвоєнь спектральних ліній спектрально-