УДК 004.62

ДОСЛІДЖЕННЯ МЕТОДІВ КЛАСТЕРИЗАЦІЇ ДЛЯ РЕАЛІЗАЦІЇ РЕКОМЕНДАЦІЙНОЇ ФУНКЦІЇ НА ОСНОВІ СУМІСНОЇ ФІЛЬТРАЦІЇ В СИСТЕМАХ ПЕРЕГЛЯДУ ФІЛЬМІВ

Касумов Б.Р.
e-mail: bohdan.kasumov@nure.ua
Харківський національний університет радіоелектроніки, каф. СТ
м. Харків, Україна

This work is devoted to the study of clusterization methods for implementing a recommendation function based on combined filtering in movie viewing systems. The article investigates five clustering methods k-means, DBSCAN and three modifications of DBSCAN: OPTICS, HDBSCAN and ST-DBSCAN. The test data was taken from the MovieLens platform and the software was developed in the Python programming language using the Anaconda platform.

У доповіді розглядається один з методів сумісної фільтрації (Collaborative Filtering) – це метод порівняння користувачів (User-Based), який використовується в системі перегляду фільмів. Для використання цих методів використовується інтерфейс системи, що надає можливість реалізувати модель оцінювання фільмів, переглянутих користувачами, за визначеною шкалою. Модель оцінювання подається у вигляді таблиці рейтингових оцінок, виставлених користувачами переглянутим фільмам. Ця таблиця є неповною. Це зумовлено тим, що є фільми, яким користувач системи ще не виставив оцінку. З використанням методу порівняння користувачів (User-Based), визначають прогнозні оцінки для неоцінених фільмів і надаються рекомендації до їх перегляду за розрахованим рейтингом прогнозних оцінок.

Для прогнозу оцінок за методом User-Based потрібно визначити групи користувачів зі схожими уподобаннями (зі схожими оцінками). Для цього використовується метод кластеризації K-Means (k-середніх).

Головною метою досліджень є порівняльна оцінка методів кластеризації K-Means, DBSCAN (Density-Based Spatial Clustering of Applications With Noise) та трьох модифікацій методу DBSCAN: OPTICS, HDBSCAN і ST-DBSCAN. Для проведення досліджень використовувався відкритий датасет системи перегляду фільмів MovieLens [1]. Програмна система для досліджень розроблена мовою Python з використанням платформи Anaconda.

Для порівняльної оцінки методів кластеризації у якості параметрів якості обрані продуктивність (час кластеризації) та три метрики:

– індекс Калінські-Харабаша (Calinski-Harabasz Index), який використовується як внутрішня метрика для оцінки якості кластеризації. Цей індекс дозволяє отримати відносну оцінку якості кластеризації між

кластерами та всередині кластерів (розраховується як відношення міжкластерної та внутрішньокластерної дисперсій). Високе значення індексу вказує про добре відокремлені і компактні кластери. Низьке значення вказує на те, що кластери дуже розмиті (точки всередині кластерів розташовані далеко від центрів кластерів), або кластери погано відокремлені один від одного [2];

– індекс Дейвіса-Болдіна (Davies-Bouldin Index) є внутрішньою метрикою оцінки якості кластеризації і дозволяє оцінити співвідношення між внутрішньою компактністю кластерів і віддаленістю між кластерами. Низьке значення індексу вказує на те, що кластери добре відокремлені один від одного і мають низьку внутрішню дисперсію. Високе значення індексу фіксує факт, що кластери мають більшу схожість між собою або більшу варіативність всередині кластерів, що може свідчити про низьку якість кластеризації [3];

– силуетний коефіцієнт (Silhouette Score) – це метрика що дозволяє поєднати показники внутрішньокластерної та міжкластерної відстані, щоб надати загальну оцінку якості кластеризації. Індекс може приймати значення в діапазоні [-1,1]. Високе значення індексу вказує на те, що об'єкт добре збігається з власним кластером і погано збігається з сусідніми кластерами [4].

За результатом досліджень визначено, що найкращим методом кластеризації за обраними параметрами оцінки якості є метод DBSCAN. Його продуктивність вище за інших (майже втричі менше за часом ніж метод HDBSCAN). Також цей метод отримав найвищі оцінки якості за Індексом Калінські-Харабаша (значення 2462547). Найгіршим методом у результаті дослідження виявився метод ST-DBSCAN, який за силуетним коефіцієнтом отримав значення в чотири рази гірше, ніж будь який інший метод. Для реалізації рекомендаційної функції системи за порівнянням користувачів (User-Based), запропоновано використовувати алгоритм DBSCAN.

Список використаних джерел:

1. MovieLens. GroupLens : website. URL: https://grouplens.org/datasets/movielens/ (date of access 13.02.2025).

2. Перова І.Г., Мірошниченко Н.С. Огляд існуючих методів зменшення розмірності та класифікації великих вибірок даних. АСУ та прилади автоматики. 2023. № 1(179). С. 42–50.

3. GeeksforGeeks. Davies-Bouldin Index : website. URL: https://www.geeksforgeeks.org/davies-bouldin-index/ (date of access: 13.02.2025).

4. Koli S. How to Evaluate the Performance of Clustering Algorithms Using Silhouette Coefficient. Medium. URL: https://medium.com/@MrBam44/how-to-evaluate-the-performance-of-clustering-algorithms-3ba29cad8c03 (date of access 13.02.2025).