УДК 004.9:791.65

ДОСЛІДЖЕННЯ МЕТОДІВ КЛАСТЕРИЗАЦІЇ ДАНИХ ДЛЯ РЕАЛІЗАЦІЇ РЕКОМЕНДАЦІЙНОЇ ФУНКЦІЇ CRM-СИСТЕМИ МЕРЕЖІ КІНОТЕАТРІВ

Одинцова В. О.
Науковий керівник – к.т.н., с.н.с. Коваленко А. І.
Харківський національний університет радіоелектроніки, каф. СТ
м. Харків, Україна
email: viktoriia.odyntsova@nure.ua

This work is showing the investigation of the clustering methods for the web cinema CRM-system «Cinema Art» that helps to implement the recommendation function for the films. The article covers the research of three clustering methods: k-means, hierarchical clustering and DBSCAN clustering. Based on the research one of the clustering methods is developed on Python using an open-source platform Anaconda. The data for the clustering research was used from an open dataset MovieLens.

The developed clustering method allows business to increase the interest of users in their application by providing them a personal-oriented experience with the recommended films and by that increase the profit and customer base.

Необхідність координації діяльності мережі кінотеатрів, розподілених за районами міста, з продажів квитків на фільми за різними сеансами, визначає важливість використання систем управління взаємовідносинами з клієнтами (Customer Relationship Management System). Застосування CRM-систем забезпечує ефективну функцію видачі рекомендаційної інформації у вигляді рекламних пропозицій, що визначаються персональними уподобаннями клієнтів.

За проведеним аналізом для реалізації рекомендаційної функції CRM-системи мережі кінотеатрів обрані два методи сумісної (колаборативної) фільтрації (Collaborative Filtering) – метод порівняння користувачів (User-Based) та метод порівняння елементів (Item-Based).

За методом порівняння користувачів визначається прогнозна оцінка рекомендованих фільмів, яка розраховується за мірою схожості уподобань інших клієнтів. За методом порівняння елементів також визначається прогнозна оцінка, яка розраховується на основі сумісної схожості оцінок фільмів. Для отримання даних за мірою схожості оцінок клієнтів або фільмів використовується методи кластеризації.

У докладі розглядається порівняльний аналіз трьох методів кластеризації: метод k-середніх, ієрархічна кластеризація та метод DBSCAN (Density-Based Spatial Clustering of Applications With Noise) з метою вибору одного з них.

Для дослідження методів кластеризації використовувалися дані з відкритого датасет ресурсу MovieLens [1]. Програмне забезпечення для

дослідження методів розроблювалось мовою Python за допомогою відкритої платформи Anaconda.

Метод кластеризації k-середніх є найбільш поширеним. За цим методом множина елементів поділяється на задану кількість кластерів k, розташованих якнайдалі один від одного [2]. Перевага цього методу у тому, що він швидкий та ефективний завдяки тому, що не потребує обчислення всіх попарних відстаней між елементами, на відміну більшості інших методів кластеризації, включаючи ті, що використовуються в процедурах ієрархічного кластерного аналізу. Головними визначеними недоліками методу k-середніх є потрібність заздалегідь задавати кількість кластерів та параметри початкового центру їхнього визначення, що за результатами може надавати різні кінцеві кластери.

Ієрархічна кластеризація поділяється на два види: «згори-вниз» або «знизу-вгору». За обраним для аналізу методом «знизу-вгору» на початку елементи розглядаються як окремі кластери, які далі послідовно об'єднуються за парами до моменту коли всі вони будуть об'єднані в єдиний кластер. Ця ієрархія кластерів подається у вигляді дерева або дендрограми. Корінь дерева – це єдиний кластер, який містить у собі всю множину елементів. Листя – це кластери, що складаються лише з одного елемента. На відміну від методу k-середніх, за ієрархічною кластеризацією отримується однозначний результат, який не залежить від завдання початкового центу і кількості кластерів. Основним визначеним недоліком є отримання надлишкової ієрархії кластерів, яка може бути зайвою в контексті поставленої задачі

Кластеризація за щільністю DBSCAN [3] має значні переваги перед іншими алгоритмами кластеризації тому що безпосереднє сканує базу даних. За цим методом визначаються області концентрації елементів. Вони відокремлюються від розріджених (порожніх) областей і визначаються як кластери. Метод DBSCAN не потребує апріорного завдання кількості кластерів, на відміну від методу k-середніх, і визначається автоматично в ході сканування.

За проведеним порівняльним аналізом для реалізації обраний метод кластеризації за щільністю DBSCAN.

Список використаних джерел:

1. MovieLens : вебсайт. URL: https://grouplens.org/datasets/movielens/ (дата звернення 26.02.2024).

2. K-Means Clustering in Machine Learning: A Deep Dive : вебсайт. URL: https://datascientest.com/en/k-means-clustering-in-machine-learning-a-deep-dive (дата звернення 27.02.2024).

3. DBSCAN Clustering in ML, Density based clustering : вебсайт. URL: https://www.geeksforgeeks.org/dbscan-clustering-in-ml-density-based-clustering/ (дата звернення 27.02.2024).