ВИКОРИСТАННЯ МЕТОДІВ КЛАСТЕРИЗАЦІЇ ТЕКСТІВ В РЕКОМЕНДАЦІЙНИХ СИСТЕМАХ

Мінець А.А., Хайрова Н.Ф.
Національний технічний університет «Харківський політехнічний інститут», м. Харків

В наш час стрімкого зростання обсягів інформації виникає потреба в її структуризації та певному механізмі категоризації, або фільтрації. Одним із напрямків використання методів кластеризації є розділення інформації за певними критеріями запитів користувача рекомендаційних систем. За допомогою кластеризації користувач може отримувати релевантну та актуальну для нього інформацію не звертаючись до додаткових інструментів пошуку або фільтрації.

Серед алгоритмів кластеризації можна виділити: алгоритм ієрархічної кластеризації, квадратичної помилки, нечіткі алгоритми, алгоритм виділення зв’язаних компонентів, алгоритм мінімального покриваючого дерева, пошарова кластеризація та інші. Отже, основне завдання нашого дослідження полягає у визначенні найбільш відповідного методу кластеризації для рекомендаційних систем.

В загальному вигляді застосування кластерного аналізу зводиться до наступних етапів: формування вибірки об’єктів для кластеризації; визначення множини змінних, за якими будуть оцінюватися об'єкти у вибірці; обчислення значень міри схожості між об'єктами; застосування методу кластерного аналізу для створення груп схожих об'єктів.

У нашому дослідженні вибірка об’єктів для кластеризації являє собою описи фільмів різних жанрів інтернет ресурсу Вікіпедія, на базі яких створено текстовий корпус для дослідження.

На основі створеного корпусу та бібліотек кластеризації виділяються ознаки, які властиві певним групам жанрів та проводиться нормалізація для рівномірного розподілу внеску певних ознак в розрахунок «відстані» між об’єктами. Наступним етапом з використанням таких метрик, як: евклідова відстань, квадрат евклідової відстані, манхеттенська відстань, відстань Чебешева ті інших, визначається ступінь схожості. На основі отриманих даних рекомендаційною системою буде запропонована вибірка із найбільш близьких фільмів за вибраним вектором ознак.

Література:

1. Jain A., Murty M., Flynn P. Data Clustering: A Review. // ACM Computing Surveys. 1999. Vol. 31, no. 3.

2. Вікіпедія – вільна енциклопедія. [Электронний ресурс]. – Режим доступу: https://ru.wikipedia.org/wiki/ (дата звернення 14.03.2020)

3. Content-Based Recommendation System- URL: https://medium.com/towards-artificial-intelligence/content-based-recommender-system-4db1b3de03e7 (дата звернення 11.03.2020)