

МАТЕМАТИЧНЕ ОБҐРУНТУВАННЯ АЛГОРИТМУ КЛАСТЕРИЗАЦІЇ КОРИСТУВАЧІВ У РЕКОМЕНДАЦІЙНИХ СИСТЕМАХ

Вінницький національний технічний університет

Анотація

У роботі розглянуто математичні основи алгоритмів кластеризації для рекомендаційних систем. Проведено порівняльний аналіз алгоритмів k-середніх, спектральної кластеризації та алгоритму Louvain з точки зору їх ефективності для групування користувачів з подібними вподобаннями. Досліджено цільові функції, що оптимізуються кожним з алгоритмів, та обґрунтовано їх застосування для різних типів даних про користувачів. Запропоновано модифікацію спектральної кластеризації, що покращує точність рекомендацій для соціальних мереж за рахунок оптимізації нормалізованого розрізу графа. Експериментальні результати демонструють підвищення якості рекомендацій на 12% при застосуванні модифікованого алгоритму порівняно з класичним підходом.

Ключові слова: рекомендаційні системи, кластеризація користувачів, алгоритм k-середніх, спектральна кластеризація, алгоритм Louvain, графи соціальних мереж, модулярність.

Abstract

This paper examines the mathematical foundations of clustering algorithms for recommender systems. A comparative analysis of k-means, spectral clustering, and the Louvain algorithm is conducted in terms of their effectiveness for grouping users with similar preferences. The objective functions optimized by each algorithm are investigated, and their application for different types of user data is substantiated. A modification of spectral clustering is proposed that improves the accuracy of recommendations for social networks by optimizing the normalized cut of the graph. Experimental results demonstrate a 12% improvement in recommendation quality when applying the modified algorithm compared to the classical approach.

Keywords: recommender systems, user clustering, k-means algorithm, spectral clustering, Louvain algorithm, social network graphs, modularity.

ВСТУП

Рекомендаційні системи відіграють важливу роль у вирішенні проблеми інформаційного перевантаження користувачів, пропонуючи їм персоналізований контент на основі їхніх уподобань та поведінки. Одним із ключових підходів до покращення якості рекомендацій є кластеризація користувачів — групування людей зі схожими інтересами для формування колаборативних рекомендацій.

Відомо, що ефективність рекомендаційних систем суттєво залежить від обраного алгоритму кластеризації та його математичного обґрунтування. Існуючі підходи, такі як k-середніх, часто не враховують складну структуру взаємозв'язків між користувачами, особливо в середовищі соціальних мереж.

Метою даного дослідження є математичне обґрунтування алгоритмів кластеризації користувачів для покращення якості рекомендацій та розробка модифікації спектральної кластеризації, що оптимально враховує структуру соціальних зв'язків користувачів.

Результати дослідження

У рекомендаційних системах користувачі можуть бути представлені різними способами: як вектори оцінок товарів, як набори характеристик профілю або як вершини в графі соціальних зв'язків. Для кожного з цих представлень оптимальними є різні алгоритми кластеризації.

Алгоритм k-середніх оптимізує суму квадратів відстаней між точками та центроїдами кластерів, що визначається цільовою функцією:

$$J = \sum_{i=1}^n \sum_{j=1}^k r_{ij} \|x_i - \mu_j\|^2,$$

де r_{ij} — бінарний індикатор приналежності точки x_i до кластера j , μ_j — центроїд кластера j . Цей алгоритм ефективний для даних векторного представлення користувачів, наприклад, коли x_i містить оцінки користувача для набору товарів. Проте k -середніх обмежений припущенням про сферичну форму кластерів і чутливий до вибору початкових центроїдів.

Для складніших структур даних, особливо коли користувачі представлені як вузли графа соціальної мережі, більш доречною є спектральна кластеризація. Цей алгоритм базується на матриці Лапласа L графа $G=(V,E)$ і оптимізує нормалізований розріз:

$$NCut(C^1, C^2, \dots, C^k) = \sum_{l=1}^k (cut(C_l, \bar{C}_l) / vol(C_l)),$$

де $cut(C_l, \bar{C}_l)$ — сума ваг ребер між кластером C_l та його доповненням, $vol(C_l)$ — сума степенів вершин у кластері. Матриця Лапласа обчислюється як:

$$L = I - D^{-\frac{1}{2}} A D^{-\frac{1}{2}},$$

де A — матриця суміжності графа, D — діагональна матриця степенів вершин. Кластеризація виконується на k власних векторах, що відповідають k найменшим ненульовим власним значенням L .

Для великих соціальних мереж ефективним є алгоритм Louvain, який оптимізує модулярність графа:

$$Q = (1/2m) \sum_{(i,j)} [A_{ij} - ((k_i k_j) / 2m)] \delta(c_i, c_j),$$

де k_i — степінь вершини i , m — загальна сума ваг ребер, $\delta(c_i, c_j)$ — індикатор приналежності вершин i та j до одного кластера. Модулярність вимірює різницю між фактичною кількістю ребер усередині кластерів та очікуваною кількістю в випадковому графі.

На основі проведеного аналізу запропоновано модифікацію спектральної кластеризації, яка враховує не лише структуру графа соціальних зв'язків, але й схожість профілів користувачів. Модифікована матриця суміжності визначається як:

$$A'_{ij} = A_{ij} + \alpha \cdot sim(x_i, x_j),$$

де $sim(x_i, x_j)$ — міра схожості профілів користувачів (наприклад, косинусна схожість), α — коефіцієнт балансування. Це дозволяє об'єднати інформацію про соціальні зв'язки та вподобання користувачів.

Таблиця 1 - Порівняння алгоритмів кластеризації на наборі даних MovieLens

Алгоритм	Точність (Precision)	Повнота (Recall)	F1- міра
k -середніх	0.68	0.64	0.66
Спектральна кластеризація	0.72	0.71	0.71
Алгоритм Louvain	0.76	0.73	0.74
Модифікована спектральна кластеризація	0.81	0.79	0.80

Експериментальні результати на наборі даних MovieLens, що містить інформацію про оцінки фільмів та соціальні зв'язки користувачів, показують, що запропонована модифікація спектральної кластеризації дає найвищу якість рекомендацій. Зокрема, F1-міра збільшилась на 12% порівняно з алгоритмом k -середніх та на 8% порівняно з класичною спектральною кластеризацією.

Висновки

У роботі проведено математичне обґрунтування алгоритмів кластеризації користувачів для рекомендаційних систем. Показано, що ефективність алгоритму залежить від способу представлення даних про користувачів. Розглянуто цільові функції трьох основних алгоритмів: k -середніх (оптимізація суми квадратів відстаней), спектральної кластеризації (мінімізація нормалізованого розрізу) та алгоритму Louvain (максимізація модулярності).

Запропоновано модифікацію спектральної кластеризації, що поєднує інформацію про соціальні зв'язки та схожість профілів користувачів. Експериментальні результати підтверджують підвищення якості рекомендацій на 12% порівняно з базовим алгоритмом. Отримані результати можуть бути застосовані для покращення рекомендаційних систем у соціальних мережах, онлайн-магазинах та сервісах потокового відтворення контенту.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Мелешко, Ю. В. Методи кластеризації графів соціальних мереж для побудови рекомендаційних систем / Ю. В. Мелешко // Журнал Національного університету "Острозька академія". Серія: Технічні науки. — 2019. — № 2. — С. 45–52.
2. Ricci, F. Recommender Systems Handbook / F. Ricci, L. Rokach, B. Shapira. — 2nd ed. — New York : Springer, 2015. — 1003 p. — ISBN 978-1-4899-7637-6.
3. Aggarwal, C. C. Data Clustering: Algorithms and Applications / C. C. Aggarwal, C. K. Reddy. — Boca Raton : CRC Press, 2013. — 652 p. — ISBN 978-1-4665-5821-2.

Пакула Антон Артурович — аспірант, Вінницький національний технічний університет, м. Вінниця, e-mail: anton.pakula.2000@gmail.com, <https://orcid.org/0009-0002-5388-5386>

Гармаш Володимир Володимирович — кандидат технічних наук, доцент, Вінницький національний технічний університет, м. Вінниця, <https://orcid.org/0009-0007-1861-8772>

Pakula Anton — PhD Student, Vinnytsia National Technical University, Vinnytsia, e-mail: anton.pakula.2000@gmail.com, <https://orcid.org/0009-0002-5388-5386>

Garmash Volodymyr — Candidate of Technical Sciences, Associate Professor, Vinnytsia National Technical University, Vinnytsia