А. А. Пакула
В. В. Гармаш

Методи кластеризації для сегментації користувачів у рекомендаційних системах

Вінницький національний технічний університет

Анотація

З розвитком інформаційних технологій та постійним зростанням кількості цифрового контенту питання надання якісних персоналізованих рекомендацій стає важливим аспектом сучасних інформаційних систем. Розробка та впровадження рекомендаційних систем, здатних ефективно адаптувати пропозиції під індивідуальні вподобання користувачів, є ключовим напрямом досліджень у галузі аналізу даних. Останнім часом усе більше уваги приділяється алгоритмам, що забезпечують точнішу сегментацію даних для поліпшення якості рекомендацій. Зокрема, методи кластеризації дозволяють групувати користувачів на основі схожих поведінкових характеристик, що позитивно впливає на точність наданих рекомендацій і сприяє підвищенню їхньої релевантності.

У процесі цього дослідження розглянуто сучасні підходи до використання методів кластеризації у контексті створення рекомендаційних систем. Значну увагу приділено принципам адаптації рекомендацій на основі аналізу користувацької поведінки та вподобань, а також можливостям підвищення точності прогнозів через використання групування даних. Розглянуто підходи до сегментації, що дозволяють створювати узгоджені й точні прогнози для різних категорій користувачів, враховуючи їхні унікальні риси та уподобання.

Проаналізовано перспективи застосування різних стратегій обробки даних і технік для підвищення якості рекомендацій. Результати дослідження підкреслюють важливість вибору відповідного методу аналізу даних для забезпечення релевантності контенту, що отримують користувачі. Визначено, що ефективна кластеризація може значно покращити не лише точність рекомендацій, але й загальний користувацький досвід, адже допомагає виявити приховані патерни у поведінці та уподобаннях, що, в свою чергу, сприяє більш глибокому розумінню потреб користувачів та підвищує ймовірність їхньої задоволеності сервісом. Таким чином, дане дослідження має значний потенціал для розвитку і вдосконалення рекомендаційних систем у різних сферах.

Ключові слова: *кластеризація, рекомендаційні системи, сегментація користувачів, машинне навчання, K-Means, DBSCAN, Android.*

Вступ

У сучасному цифровому світі користувачі щоденно стикаються з величезною кількістю інформації та контенту. Платформи, що надають доступ до фільмів, музики, книг і товарів, пропонують сотні тисяч варіантів для вибору. В умовах такої інформаційної перенасиченості користувачі потребують інструментів, які допоможуть їм знайти релевантний контент швидко й ефективно. Одним із найпоширеніших рішень для цієї проблеми є рекомендаційні системи, що здатні надавати персоналізовані пропозиції на основі історії взаємодій користувача, його вподобань та поведінкових патернів.

Рекомендаційні системи застосовують різні алгоритми для передбачення, який контент може зацікавити користувача. Однак, для забезпечення максимальної точності рекомендацій важливо правильно сегментувати користувачів за групами з подібними інтересами. Це дозволяє системі надавати не лише індивідуальні рекомендації, але й використовувати колективний досвід груп схожих користувачів. Одним із ключових підходів для такої сегментації є кластеризація — процес автоматичного групування користувачів на основі схожих характеристик або поведінкових патернів.

Кластеризація є одним з методів машинного навчання без нагляду (unsupervised learning), що не вимагає попередньо розмічених даних. Метою кластеризації є поділ набору користувачів або об'єктів на кілька груп, або кластерів, де кожен елемент у межах кластера максимально схожий до інших елементів цього ж кластера, і водночас значно відрізняється від елементів інших кластерів. У контексті рекомендаційних систем кластеризація дозволяє групувати користувачів із подібними інтересами, що дозволяє значно підвищити точність персоналізованих рекомендацій.

Основні методи кластеризації

Для ефективної сегментації користувачів у рекомендаційних системах використовуються різні алгоритми кластеризації, кожен з яких має свої унікальні властивості та підходить для певних типів даних або завдань. У цій роботі розглядаються три основні методи кластеризації: K-Means, DBSCAN та агломеративна кластеризація. Кожен з цих методів має свої переваги та недоліки, що впливають на вибір алгоритму для конкретного випадку.

K-Means (алгоритм K-середніх). Алгоритм K-Means є одним із найбільш популярних і широко використовуваних методів кластеризації. Він підходить для обробки великих наборів даних і забезпечує швидку та ефективну кластеризацію. Основний принцип роботи цього алгоритму полягає у поділі даних на **K** кластерів, де **K** — це наперед визначена кількість. Алгоритм працює за такою схемою:

1. На початку випадковим чином вибираються **K** центроїдів (центрів кластерів).
2. Кожен об'єкт або точка даних призначається до найближчого центроїда на основі відстані (зазвичай використовується Евклідова відстань).
3. Після призначення всіх точок до кластерів обчислюються нові центроїди як середнє значення точок у кожному кластері.
4. Процес повторюється, доки центроїди не перестануть змінюватися або не досягнуто максимальної кількості ітерацій.

Візуалізацію алгоритму **К-середніх** зображено на рисунку 1.

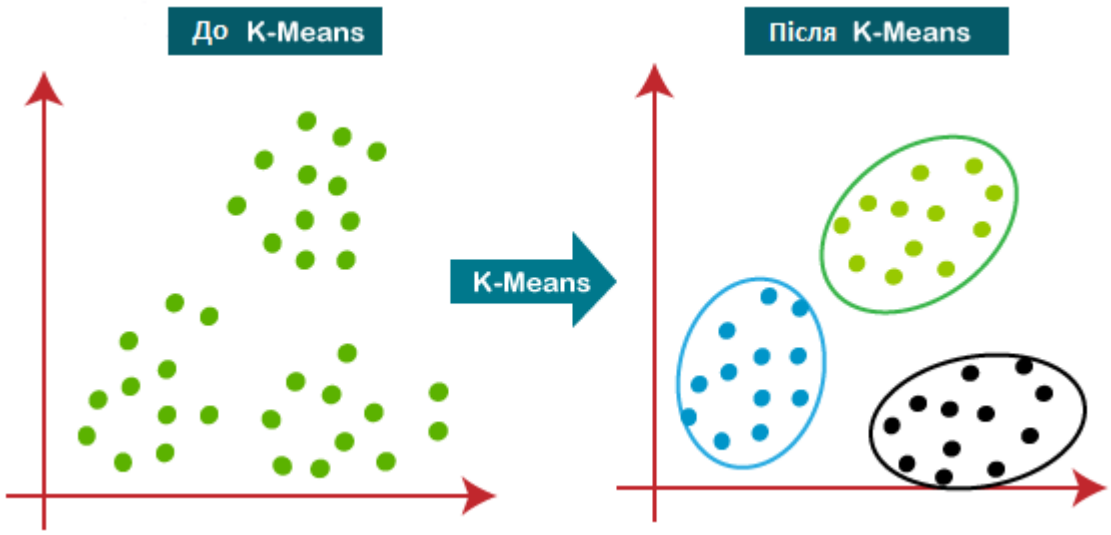

Рисунок 1 – візуальна інтерпретація алгоритму K-means.

Однією з головних переваг **K-Means** є його висока швидкість і простота реалізації, що робить його ідеальним для великих обсягів даних. Однак, цей алгоритм має кілька суттєвих недоліків:

- Потрібно наперед знати кількість кластерів (**K**), що не завжди можливо.
- Алгоритм чутливий до початкового вибору центроїдів, що може призвести до різних результатів при різних ініціалізаціях.
- **K-Means** неефективний для кластерів складної форми або даних з нерівномірною щільністю.

DBSCAN (Density-Based Spatial Clustering of Applications with Noise). DBSCAN є методом кластеризації на основі щільності, який не потребує наперед визначати кількість кластерів і здатен виявляти кластери складної форми. Алгоритм працює за таким принципом:

1. Для кожної точки визначається її щільність шляхом підрахунку кількості сусідів у межах заданого радіусу (**eps**).
2. Якщо точка має достатньо сусідів (більше, ніж заданий поріг **minPts**), вона стає "ядром" кластера.
3. Точки, що знаходяться в межах радіуса ядра, приєднуються до кластера.
4. Точки, які не мають достатньої кількості сусідів, вважаються шумом і не відносяться до жодного кластера.

На рисунку 2 зображено принцип роботи алгоритму.

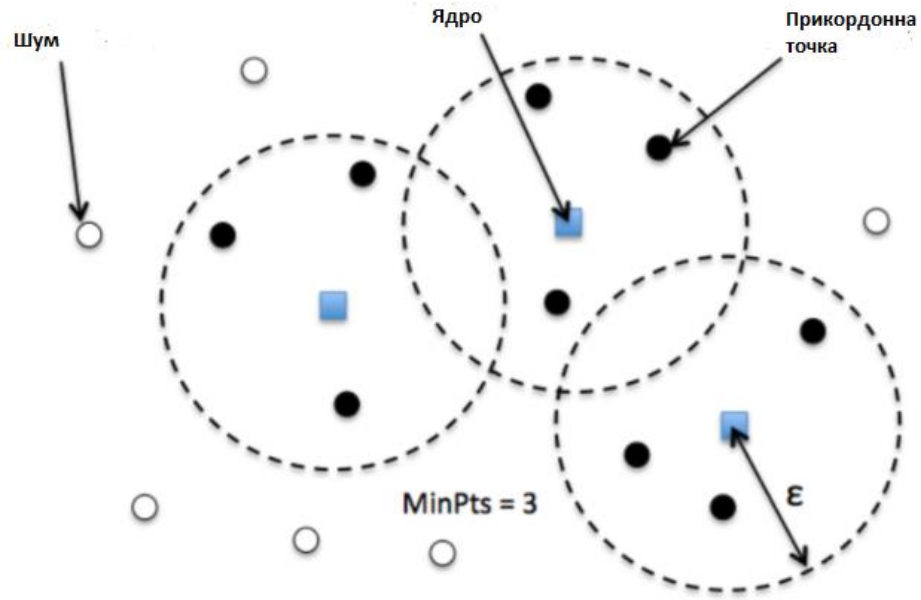

Рисунок 2 – візуальна інтерпретація алгоритму DBSCAN.

Переваги **DBSCAN**:

- Він автоматично визначає кількість кластерів на основі щільності даних.
- Здатен виявляти кластери довільної форми, що робить його більш універсальним у порівнянні з **K-Means**.
- **DBSCAN** добре працює в умовах наявності шумових точок, які можуть бути виключені з кластерів як "шум".

Проте, **DBSCAN** має певні обмеження:

- Параметри **eps** і **minPts** значно впливають на результат кластеризації, і їх правильний вибір може бути складним.
- Алгоритм погано працює на даних із нерівномірною щільністю, де одна частина даних може мати дуже щільні кластери, а інша — розріджені.

Agglomerative Clustering (агломеративна кластеризація). Агломеративна кластеризація є одним із видів ієрархічної кластеризації, яка починається з того, що кожен об'єкт є окремим кластером. Далі, найближчі кластери поступово об'єднуються доти, доки не залишиться один великий кластер або не досягнуто бажаної кількості кластерів. Алгоритм працює за наступною схемою:

1. На початковому етапі кожна точка даних вважається окремим кластером.
2. Алгоритм обчислює відстані між усіма кластерами та об'єднує ті, що мають найменшу відстань.

3. Процес повторюється до тих пір, поки всі об'єкти не будуть об'єднані в один кластер або поки не досягнуто бажаної кількості кластерів.

Особливість агломеративної кластеризації полягає в тому, що вона будує **дендрограму** — ієрархічне дерево кластерів. Це дозволяє візуалізувати процес кластеризації і вибирати потрібний рівень деталізації (кількість кластерів) на основі побудованого дерева. Приклад дендрограми наведено на рисунку 3.

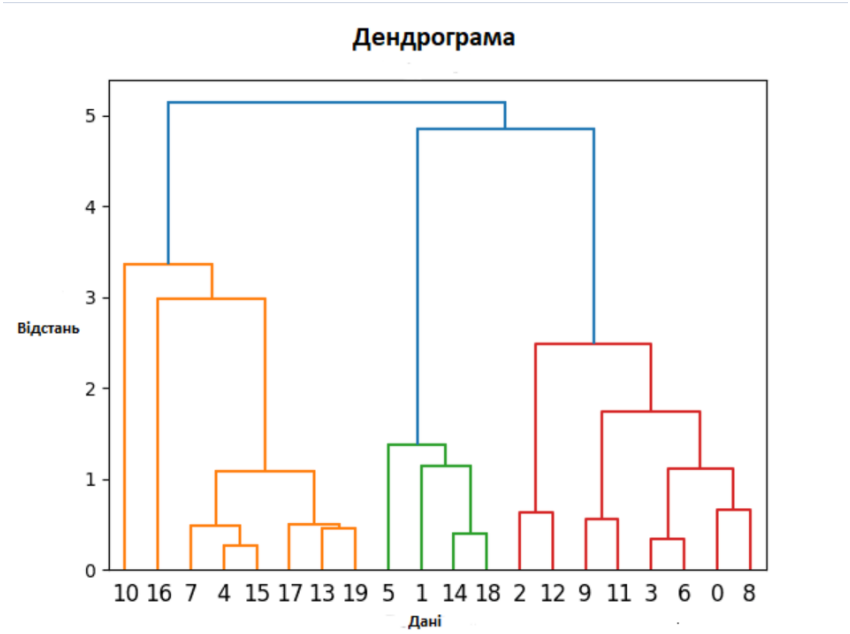

Рисунок 3 – приклад дендрограми.

Переваги **агломеративної кластеризації**:

- Алгоритм не потребує наперед визначати кількість кластерів, і це можна зробити після побудови дендрограми.
- Він підходить для роботи з невеликими або середніми наборами даних.
- Візуалізація у вигляді дендрограми дозволяє легко зрозуміти структуру даних та взаємозв'язки між кластерами.

Недоліки:

- Висока обчислювальна складність при роботі з великими наборами даних, оскільки необхідно обчислювати відстані між усіма кластерами на кожній ітерації.
- Нестабільність результатів через те, що малі зміни у даних можуть вплинути на формування кластерів.

Висновок

Було розглянуто три методи кластеризації: K-Means, DBSCAN та агломеративну кластеризацію, які є важливими інструментами для розуміння різних підходів до аналізу даних та формування кластерів. У процесі роботи було детально розглянуто специфіку кожного з методів, їхні переваги та недоліки, а також можливості їх застосування в різних контекстах.

K-Means виявився простим і швидким методом, що ефективно працює з великими наборами даних. Однак його основні недоліки, такі як необхідність заздалегідь визначити кількість кластерів та чутливість до початкових значень, обмежують його використання у випадках зі складною структурою даних або великою кількістю шумових точок.

На противагу цьому, DBSCAN показав свою ефективність у виявленні кластерів довільної форми та ігноруванні викидів. Ці характеристики роблять його особливо корисним у ситуаціях, коли дані мають нерівномірну щільність. Проте, успіх цього алгоритму значною мірою залежить від правильного вибору параметрів, що може вимагати додаткових налаштувань і експериментів.

Агломеративна кластеризація забезпечує більшу гнучкість у визначенні кількості кластерів і є надійним методом для малих і середніх наборів даних. Однак вона може бути менш ефективною для великих обсягів даних через свою обчислювальну складність.

Вибір методу кластеризації має бути ретельно обґрунтований, виходячи з характеристик даних та специфіки завдання. Правильний підбір алгоритмів кластеризації може значно підвищити ефективність аналізу даних і надання персоналізованих рішень у сучасному інформаційному середовищі. Важливо враховувати як переваги, так і обмеження кожного методу, щоб забезпечити успішне впровадження кластеризації в практичних застосуваннях.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Bishop, C. M. Pattern Recognition and Machine Learning. Springer, 2006.
2. Jain, A. K., Murty, M. N., & Flynn, P. J. Data Clustering: A Review. ACM Computing Surveys, 2008.
4. Rokach, L., & Maimon, O. Clustering Methods. In Data Mining and Knowledge Discovery Handbook. Springer, 2005.
5. Ghosh, A., & Mondal, M. A Survey on Clustering Techniques. International Journal of Computer Applications, 2018.

Пакула Антон Артурович – студент групи 174-23а, факультет інтелектуальних інформаційних технологій та автоматизації, Вінницький національний технічний університет, Вінниця, e-mail: anton.pakula.2000@gmail.com

Гармаш Володимир Володимирович – канд. техн. наук, доцент кафедри AIBT, Вінницький національний технічний університет, Вінниця, e-mail: garmash.v.v@vntu.edu.ua

A. A. Pakula
V. V. Garmash

Clustering methods for user segmentation in recommender systems

Vinnytsia National Technical University

Abstract.

With the development of information technologies and the constant growth of digital content, the issue of providing high-quality personalized recommendations is becoming an important aspect of modern information systems. The development and implementation of recommender systems that can effectively adapt suggestions to individual user preferences is a key area of research in data mining. Recently, more and more attention has been paid to algorithms that provide more accurate data segmentation to improve the quality of recommendations. In particular, clustering methods allow grouping users based on similar behavioral characteristics, which positively affects the accuracy of the recommendations provided and contributes to their relevance.

In the course of this study, modern approaches to the use of clustering methods in the context of creating recommender systems are considered. Considerable attention is paid to the principles of adapting recommendations based on the analysis of user behavior and preferences, as well as the possibilities of improving the accuracy of forecasts through the use of data grouping. Segmentation approaches that allow creating consistent and accurate forecasts for different categories of users, taking into account their unique features and preferences, are considered.

The prospects of using different data processing strategies and techniques to improve the quality of recommendations are analyzed. The results of the study emphasize the importance of choosing an appropriate data analysis method to ensure the relevance of the content received by users. It has been determined that effective clustering can significantly improve not only the accuracy of recommendations but also the overall user experience, as it helps to identify hidden patterns in behavior and preferences, which in turn contributes to a deeper understanding of user needs and increases the likelihood of their satisfaction with the service. Thus, this study has significant potential for the development and improvement of recommender systems in various fields.

Keywords: *clustering, recommendation systems, user segmentation, machine learning, K-Means, DBSCAN, Android.*

Pakula Anton A. – student of group 174-23a, faculty of intellectual information technologies and automation, Vinnytsia National Technical University, Vinnytsia, e-mail: anton.pakula.2000@gmail.com

Garmash Volodymyr V. – candidate technical of Sciences, associate professor of AIVT department, Vinnytsia National Technical University, Vinnytsia, e-mail: garmash.v.v@vntu.edu.ua