https://orcid.org/0000-0001-8361-374X

УДК 519.254

А.А. МУХА*

КЛАСТЕРИЗАЦІЯ АСОРТИМЕНТУ ОБ'ЄКТІВ ТОРГОВЕЛЬНОЇ МЕРЕЖІ НА ОСНОВІ ВІДСТАНІ ЖАККАРА ТА ДЕНДРОГРАМ

*Institute of Mathematical Machines and Systems Problems of the National Academy of Sciences of Ukraine, Kyiv, Ukraine

Анотація. *Робота присвячена питанням автоматизованої кластеризації асортименту роздрібних магазинів торговельних мереж. Дослідження актуальне для торговельних мереж з великою кількістю магазинів та великою кількістю асортименту понад 2000–3000 товарів. За таких умов порівняльний аналіз ефективного асортименту потребує значних часових і фінансових витрат, оскільки виконується аналітиками вручну або з обмеженим рівнем автоматизації. Оптимізація процесів управління асортиментом зводиться до поділу магазинів на однорідні групи, що дозволяє спростити прийняття управлінських рішень, стандартизувати асортиментну політику та підвищити ефективність брендування. Традиційно класифікація магазинів здійснюється за площею торгового залу та загальним обсягом асортименту, що не завжди відображає реальну подібність товарних пропозицій. У статті запропоновано алгоритм автоматизованої кластеризації магазинів на основі аналізу перетину асортиментів. Описаний метод базується на побудові бінарної матриці наявності товарів у магазинах та обчисленні попарних відстаней Жаккара між асортиментами магазинів. На основі отриманої матриці відстаней виконується ієрархічна кластеризація з побудовою дендрограм, що дозволяє визначити оптимальну кількість кластерів відповідно до заданих параметрів. Додатково виконана візуалізація результатів кластеризації з використанням методу Т-розподіленого вкладення стохастичної близькості (t-distributed Stochastic Neighbor Embedding, t-SNE), що забезпечує наочну інтерпретацію структури даних у просторі зниженої розмірності. Запропонований підхід дозволяє не лише сформувати групи магазинів зі спільним базовим асортиментом, але й визначити товарні позиції, які відрізняють окремі торгові точки всередині мережі. Отримані результати можуть бути використані для оптимізації асортиментної матриці, формування ефективних торгових форматів та зниження витрат на управління асортиментом у великих роздрібних мережах.*

Ключові слова: *відстань Жаккара, ієрархічна кластеризація, дендрограма, класифікація магазинів, асортиментна матриця.*

Abstract. *This paper addresses the issues of automated clustering of product assortments in retail networks. The study is particularly relevant for retail chains with a large number of stores and a broad assortment exceeding 2,000–3,000 items. Under such conditions, comparative analysis of effective assortments requires substantial time and financial resources, as it is typically performed manually by analysts or with a limited level of automation. Optimization of assortment management processes can be achieved by grouping stores into homogeneous clusters, which simplifies managerial decision-making, enables standardization of assortment policies, and improves branding efficiency. Traditionally, store classification is based on sales area size and total assortment volume. However, these criteria do not always reflect the actual similarity of product offerings. This paper proposes an algorithm for automated store clustering based on the analysis of assortment overlap. The method is based on constructing a binary matrix representing product availability across stores and computing pairwise Jaccard distances between store assortments. Using the resulting distance matrix, hierarchical clustering is performed with dendrogram construction, allowing the de-*

ISSN 1028-9763. Математичні машини і системи. 2026. № 1

termination of an optimal number of clusters according to predefined parameters. Additionally, the clustering results are visualized using the t-distributed Stochastic Neighbor Embedding (t-SNE) method, which provides an intuitive interpretation of the data structure in a reduced-dimensional space. The proposed approach enables not only the formation of store groups with a common core assortment but also the identification of product items that distinguish individual stores within the retail network. The obtained results can be applied to optimize assortment matrices, design effective store formats, and reduce assortment management costs in large retail chains.
Keywords: *Jaccard distance, hierarchical clustering, dendrogram, store classification, assortment matrix.*

DOI: 10.34121/1028-9763-2026-1-92-99

1. Вступ

Сучасні роздрібні торговельні мережі представляють собою розгалужену систему, до складу якої входять тисячі торгових точок, при цьому ринкові умови вимагають подальшого зростання.

Для великих мереж кількість товарних позицій у межах однієї категорії може сягати кількох тисяч, а в цілому асортимент мережі може налічувати сотні тисяч одиниць. Цей фактор суттєво ускладнює процеси аналізу, планування та оптимізації асортименту. За таких умов ефективне управління асортиментною матрицею стає критично важливим чинником забезпечення прибутковості та конкурентоспроможності роздрібного бізнесу.

Традиційні підходи до управління асортиментом ґрунтуються на індивідуальному аналізі кожної торгової точки або на використанні спрощених критеріїв класифікації магазинів, зокрема площі торгового залу, формату магазину чи загального обсягу товарної пропозиції. Подібні критерії не завжди адекватно відображають реальну структуру асортименту та ступінь подібності товарних пропозицій між магазинами, що може призводити до неефективних управлінських рішень та зайвих витрат ресурсів.

Одним із шляхів оптимізації процесів управління асортиментом є поділ магазинів торговельної мережі на однорідні групи, для яких можуть застосовуватися уніфіковані правила формування асортименту, ціноутворення та мерчандайзингу. Такий підхід дозволяє зменшити складність управління, стандартизувати асортиментну політику та забезпечити формування чітких торгових форматів, зрозумілих для кінцевого споживача.

У контексті розвитку методів аналізу даних та машинного навчання особливої уваги набувають алгоритми автоматизованої кластеризації, здатні обробляти великі обсяги інформації з мінімальним втручанням аналітика. Зокрема, перспективним є використання метрик подібності, що ґрунтуються на аналізі перетину множин товарних позицій, оскільки асортимент магазину за своєю природою є бінарною структурою типу «товар присутній / відсутній».

Метою даної статті є розробка та дослідження алгоритму автоматизованої кластеризації асортименту роздрібних магазинів на основі аналізу перетину асортиментів із використанням відстані Жаккара та методів ієрархічної кластеризації.

Основною задачею є виділення спільних кластерів асортименту в кожній товарній категорії та знаходження товарних позицій, що не входять до кластерів асортименту.

Практична цінність отриманих результатів полягає у можливості їх застосування в системах підтримки прийняття рішень для категорійних менеджерів та аналітиків роздрібних мереж із метою зниження витрат на управління асортиментом, підвищення операційної ефективності та формування більш точних торгових форматів.

Визначення 1. Під кластером розуміється набор товарів, який має бути присутній у певній товарній категорії кожного магазину одного формату.

2. Огляд існуючих робіт

У більшості ранніх і прикладних робіт класифікація магазинів здійснюється на основі агрегованих кількісних показників, як-от обсяг продажів, товарообіг, середній чек, площа торгового залу або географічне розташування [1, 2]. Подібні підходи дозволяють швидко отримати загальну сегментацію торговельної мережі, проте вони не враховують структурні відмінності у складі асортименту та можуть приховувати суттєві розбіжності у товарній пропозиції між магазинами з подібними фінансовими показниками. Обслуговування різного асортименту має свої недоліки, наприклад, складну складську та транспортну логістику.

У сучасних дослідженнях для сегментації магазинів використовуються такі алгоритми, як k -середніх [3] , ієрархічна кластеризація, самоорганізовані карти та методи на основі різних розподілів. Як правило, вхідними даними для таких моделей є вектори продажів за категоріями, частотні характеристики покупок або поведінкові ознаки споживачів. Хоча ці методи дозволяють виявляти приховані закономірності у великих масивах даних, їх результати значною мірою залежать від масштабу продажів та сезонних коливань.

Окремий напрям досліджень присвячений аналізу асортиментної подібності магазинів. У таких роботах розглядається гіпотеза про те, як магазини зі схожою структурою асортименту можуть бути об’єднані в єдині формати для спрощення управління [4]. Проте у більшості випадків асортимент подається у вигляді зважених ознак або агрегованих показників, що ускладнює інтерпретацію результатів та не дозволяє безпосередньо виділити спільні й унікальні товарні позиції.

Для аналізу подібності множин у науковій літературі [5] широко застосовуються метрики, призначені для бінарних даних, зокрема коефіцієнт та відстань Жаккара. Ця метрика є особливо придатною для задач, у яких важливо оцінити ступінь перетину між наборами об’єктів, і не залежить від абсолютної кількості елементів у множинах. Вона успішно використовується у задачах аналізу рекомендаційних систем та кластеризації об’єктів з ознаками типу «наявність/відсутність».

Ієрархічна кластеризація є одним із найбільш інтерпретованих методів групування об’єктів і широко застосовується у маркетингових дослідженнях. Побудова дендрограм дозволяє візуально аналізувати структуру подібності між об’єктами та обирати кількість кластерів відповідно до аналітичних або бізнес-вимог. На відміну від методів з фіксованою кількістю кластерів, ієрархічні алгоритми забезпечують гнучкість у прийнятті рішень та прозорість результатів.

Отже, аналіз існуючих робіт показує, що, незважаючи на широкий спектр методів класифікації магазинів, питання автоматизованого групування торговельних точок безпосередньо на основі перетину асортиментів залишається недостатньо висвітленим. Більшість підходів орієнтована на продажі або агреговані показники, тоді як структура асортименту як первинне джерело подібності використовується обмежено. Це визначає доцільність розробки методів, що дозволяють кластеризувати магазини за асортиментною подібністю з можливістю подальшої практичної інтерпретації результатів.

3. Виклад основного матеріалу

3.1. Формалізація задачі

Нехай торговельна мережа складається з множини торгових точок

$$S = \{s_1, s_2, \dots s_n\},$$

кожна з яких має асортимент товарів із множини

$$P = \{p_1, p_2, \dots p_m\}.$$

3.1.1. Побудова бінарної асортиментної матриці

Для формалізації асортименту кожного магазину будується бінарна матриця

$$A = [a_{ij}], a_{ij} = \begin{cases} 1, & \text{якщо товар } p_j \text{ присутній у магазині } s_i, \\ 0 & \text{— в іншому випадку.} \end{cases}$$

3.1.2. Оцінка подібності асортиментів

Для оцінки ступеня подібності між асортиментами двох магазинів використовується метрика Жаккара [6]. Відстань Жаккара між магазинами s_i та s_j визначається як

$$d_J(s_i, s_j) = 1 - \frac{| A_i \cap A_j |}{| A_i \cup A_j |},$$

де A_i та A_j — множини товарів, представлених у відповідних магазинах.

У результаті для всієї множини магазинів формується симетрична матриця попарних відстаней, що відображає рівень асортиментної подібності між усіма торговими точками мережі.

3.1.3. Обчислення матриці попарних відстаней

Для кожної пари магазинів (s_i, s_j) обчислюється відстань Жаккара:

$$d_{ij} = 1 - \frac{\sum_{k=1}^{m} a_{ik} \cdot a_{jk}}{\sum_{k=1}^{m} \max(a_{ik}, a_{jk})}.$$

У результаті формується симетрична матриця відстаней

$$D = [d_{ij}], i, j = 1..n.$$

3.1.4. Ієрархічна кластеризація

На основі матриці відстаней D виконується агломеративна ієрархічна кластеризація [7] з використанням методу середнього зв'язування (average linkage).

На початковому етапі кожен магазин утворює окремий кластер. Далі на кожній ітерації об'єднуються два кластери C_p та C_q, для яких середня відстань є мінімальною і виражається як

$$d(C_p, C_q) = \frac{1}{|C_p||C_q|} \sum_{s_i \in C_p} \sum_{s_j \in C_q} d_{ij}.$$

Обчислення виконується за циклом і триває до об'єднання всіх магазинів у єдину ієрархічну структуру.

3.1.5. Визначення кількості кластерів

Результатом ієрархічної кластеризації є дендрограма (рис. 1), де кількість кластерів визначається шляхом вибору порогу відсікання τ за значенням відстані від одного об'єднання кластерів до іншого. Отже, кількість об'єднань буде дорівнювати кількості кластерів.

Поріг τ може визначатися автоматично на основі аналізу приростів відстаней між послідовними об'єднаннями кластерів. Оптимальний поріг відповідає максимальному стрибку відстані, що інтерпретується як точка переходу між групами магазинів.

3.1.6. Формування кластерів магазинів

Після визначення порогу τ дендрограма розсікається, і кожному магазину s_i присвоюється мітка кластера:

$$c_i \in \{1, 2, ..., K\},$$

де K — кількість отриманих кластерів.

3.1.7. Визначення базового асортименту кластерів

Для кожного кластера C_k визначається базовий асортимент:

$$B_k = \bigcap_{s_i \in C_k} P_i,$$

який містить товарні позиції, присутні у всіх магазинах даного кластера B_k.

Крім того, для кожного магазину визначаються товарні позиції O_i, що не входять до базового асортименту кластера:

$$O_i = P_i \setminus B_{c_i},$$

що дозволяє ідентифікувати індивідуальні відхилення магазинів.

4. Практичне застосування

4.1. Приклад

Наочний розрахунок виконаної кластеризації для малої кількості магазинів. Аналізуємо дані про наявність товарів (SKU) у торговій точці за кожною категорією товару.

Для кожної категорії формуємо бінарну матрицю рядка — магазини, стовпці — товари, де значення 1 — присутній, 0 — відсутній (табл. 1).

Таблиця 1 — Бінарна матриця наявності асортименту у магазинах

	SKU 1	SKU 2	SKU 3	SKU 4	SKU 5
Магазин А	1	0	1	1	0
Магазин Б	1	1	1	0	0
Магазин В	0	1	1	1	1
Магазин Г	1	0	1	1	0

Далі порівнюємо магазини один з одним і формуємо симетричну матрицю відстаней.
Магазин А: {1,3, 4}.
Магазин В: {1,2, 3}.
Перетин: {1,3}=2 спільних SKU.
Об’єднання: {1,2, 3, 4}=4 унікальних SKU.
Відстань Жаккара=1–(2 / 4)=0,5.

Аналогічні обчислення виконуються для всіх пар магазинів, у результаті чого формується симетрична матриця попарних відстаней. Для наведеного прикладу така матриця має вигляд (табл. 2).

Таблиця 2 — Матриця відстаней асортименту магазинів відносно один одного

	Магазин А	Магазин Б	Магазин В	Магазин Г
Магазин А	0	0,5	0,6	0
Магазин Б	0,5	0	0,5	0,5
Магазин В	0,6	0,5	0	0,6
Магазин Г	0	0,5	0,6	0

Отримана матриця відстаней використовується як вхідні дані для ієрархічної агломеративної кластеризації. На її основі будується дендрограма (рис. 1), що відображає структуру подібності між магазинами. У даному прикладі магазини А та Г, які мають ідентичні асортиментні множини, об’єднуються на мінімальному рівні відстані, тоді як інші магазини приєднуються до кластерів на вищих рівнях.

Рисунок 1 — Дендрограма подібності

4.2. Практичне застосування

Аналізуємо дані про наявність товарів (SKU) у торгових точках реальної торгової мережі з трьома різними форматами магазинів: формат супермаркету (Market), формат магазину біля дому (Smart), формат магазину екотоварів (Eko). Для зручності подальшого аналізу, а також з огляду на те, що асортимент має спорідненість або відмінності в розрізі товарних категорій, розрахунки проводилися окремо за кожною категорією товарів. Для наочності наведено один рядок тестового набору даних з табл. 3. Загалом набір містив близько 9000 SKU в 105 товарних категоріях. До вибірки було додано 14 торгових точок.

Таблиця 3 — Приклад набору даних

Shop_id	shop_name	Category_id	Category_name	Sku	Name
100	222 Лозова (Маркет)	02	Солодка вода	2436	Напій Кока-Кола 0,5л

Алгоритм розрахунку був реалізований у програмі на мові Python із використанням бібліотек: Pandas, Sklearn, Scipy, Numpy. Результатом роботи програми став відсортований асортимент. Дані виводяться у виді файлів. Рис. 2 відповідно до тек кожного магазину: список кластерів для кожної категорії та товари, що в них входять, а також список асортименту,

який не увійшов у кластер, але доступний у конкретному магазині. А також графіки дендрограм по кожній категорії кожного магазину (рис. 3) та t-SNE-діаграми (рис. 4). При необхідності ця інформація може бути завантажена до корпоративного сховища для подальшого аналізу та прийняття управлінських рішень, а також відображення в ERP-системі.

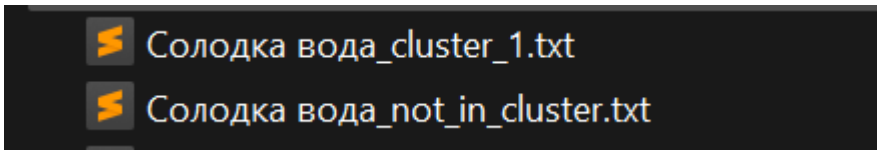

Рисунок 2 — Файли з даними (солодка вода_clusster_1.txt — товари в кластерах, солодка вода_clusster_not_in_cluster.txt — товари, які не увійшли в кластер)

Також для кожної категорій отримані дендрограми спорідненості магазинів (рис. 3).

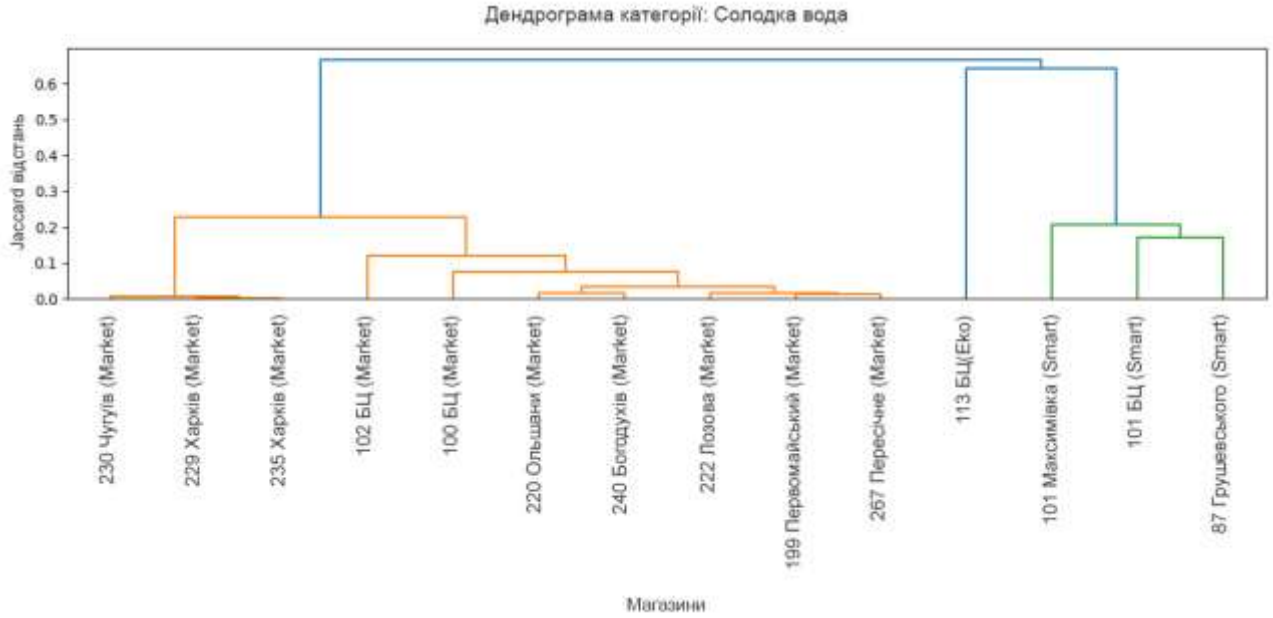

Рисунок 3 — Дендрограма подібності (кластер 1 — помаранчевий, кластер 2 — зелений, кластер 3 — синій)

Як видно з дендограми, алгоритм вірно визначив кількість форматів магазинів, що зображена кольором, а динамічний розрахунок ступеня зміни відмінності вірно визначив кількість мінімально необхідних кластерів асортименту, що дорівнює кількості форматів магазинів. Наочно видно, значний стрибок у відстані між кластером 3 і 1, що означає значне розходження в асортименті.

Для більшої наочності було реалізовано відображення результатів розрахунків за категоріями у вигляді розподіленого вкладення стохастичної близькості t-distributed Stochastic Neighbor Embedding, (t-SNE) [8] (рис. 3).

На рис. 4 що ближче точки одна до одної, то більше спорідненість. Отже, видно, що магазини помаранчевого кольору не є однорідними за своїм асортиментом і потребують вирівнювання асортименту або додання ще одного кластера для зручності управління. Як виявилося, сім із них мають значно більшу площу та ширший асортимент, а не лише його представленість. У будь-якому випадку на таку ситуацію в категорії відповідний категорійний менеджер може звернути додаткову увагу.

Отримані дані дозволяють оперативно скоригувати асортимент, щоб досягти однорідності представленості, а також визначити, якою мірою магазини відрізняються за рівнем представленості.

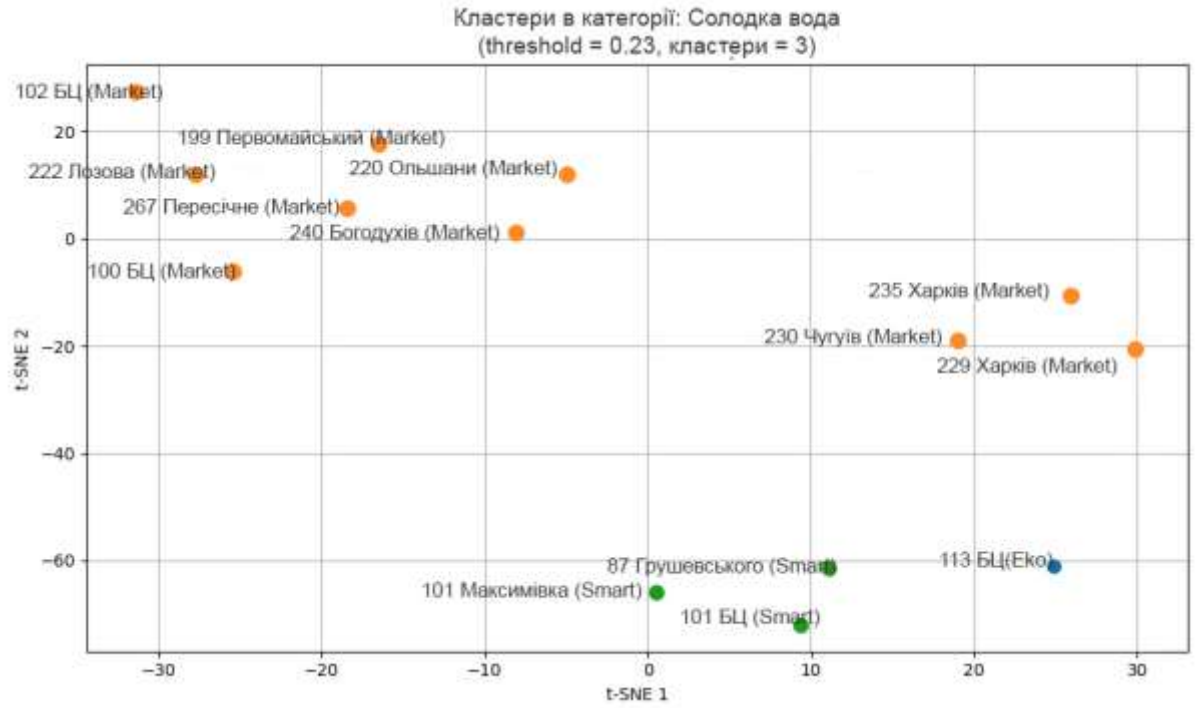

Рисунок 4 — T-sne діаграма кластерів у категорії солодка вода

Без автоматизації такий аналіз зайняв би значний час аналітиків. У випадках, коли кількість магазинів наближається до 1000, автоматизований підхід не має альтернатив.

5. Висновки

Запропонований у роботі метод є ефективним засобом автоматизації, який здатний вирішити проблему класифікації великих масивів даних, спрощуючи у такий спосіб значну кількість ручної праці аналітиків та категорійних менеджерів. Засоби візуалізації надають значні переваги для сприйняття інформації і можуть бути використані для контролю стану асортименту в категоріях товарів.

Метод може знайти свій розвиток у випадку зміни бінарної моделі на модель із показниками, наприклад, обсягами продажів. Але в такому випадку відстань Жаккара слід змінити на більш складні оціночні критерії, наприклад, відстань Махаланобіса.

СПИСОК ДЖЕРЕЛ

1. Reutterer T., Teller C. Store format choice and shopping trip types. *International Journal of Retail & Distribution Management*. 2009. Vol. 37 (8). P. 695–710.
2. Fox E.J., Montgomery A.L., Lodish L.M. Consumer shopping and spending across retail formats. *Journal of Business*. 2014. Vol. 77 (2). P. S25–S60.
3. Dolnicar S. A review of unquestioned standards in using cluster analysis for data-driven market segmentation. *Conference Proc. of the Australian and New Zealand Marketing Academy Conference*. Deakin University, Melbourne, 2002. P. 31–37.
4. Mantrala M.K. Why is assortment planning so difficult for retailers? *Journal of Retailing*. 2009. Vol. 85 (1). P. 71–83.
5. Levandowsky M., Winter D. Distance between sets. *Nature*. 1971. Vol. 234. P. 34–35.
6. Kaufman L., Rousseeuw P.J. Finding Groups in Data: An Introduction to Cluster Analysis, Wiley, 1990. pp. 47–51. Analysis of Agglomerative Clustering. *Algorithmica*. 2012. Vol. 69 (1). P. 184–215.
8. Linderman G.C., Steinerberger S. Clustering with t-SNE, provably. *SIAM journal on mathematics of data science 1.2.* 2019. Vol. 1, 2. P. 313–332.

Стаття надійшла до редакції 11.11.2025 / прийнята до друку 12.02.2026