УДК 004.9

М. В. Лобур, М. Є. Шварц, Ю. В. Стех
Національний університет “Львівськав політехніка”,
кафедра систем автоматизованого проектування

МОДЕЛІ І МЕТОДИ ПРОГНОЗУВАННЯ РЕКОМЕНДАЦІЙ ДЛЯ КОЛАБОРАТИВНИХ РЕКОМЕНДАЦІЙНИХ СИСТЕМ

У цій статті проаналізовано сучасний стан моделей і методів побудови рекомендаційних систем. Виділено основні класи задач, які вирішують рекомендаційні системи. Показано особливості застосування методу колаборативної (спільної) фільтрації. Розроблено метод мішаної категоріально-чисельної кластеризації для пошуку груп користувачів, який використовує числові рейтингові і демографічні характеристики користувачів, розроблено гібридний метод пошуку груп користувачів, який використовує коефіцієнт розрідженості матриці користувач-предмет.

Ключові слова: колаборативна фільтрація, прогнозування рекомендацій, категоріальна кластеризація, групові рекомендації.

This article analyzes the current state of models and methods for constructing recommender systems. The main classes of tasks that solve recommender systems are highlighted. The features of the application of the method of collaborative (joint) filtering are shown. A mixed numerical-categorical clustering method for searching for user groups that uses numerical rating and demographic characteristics of users has been developed, a hybrid method for searching for user groups has been developed that uses the coefficient of user-subject matrix sparseness.

Key words: collaborative filtration, forecasting of recommendations, categorical clustering, group recommendations.

Вступ

Останні десятиліття спостерігається значне зростання інформації у світовому інформаційному просторі. Широке впровадження Інтернет-технологій у всі сфери суспільного життя, доступність інформації, вимагає розроблення нових методів пошуку інформації. Пошукові системи Google, Yahoo, Altavista не враховують персоналізацію інформації. Рекомендаційні системи – це системи, які працюють із певним типом інформації, системою фільтрів, що рекомендують інформаційні елементи, які можуть викликати інтерес користувача [1]. Типова рекомендаційна система приймає рекомендації користувачів як вхідних даних, агрегує і відправляє їх до відповідних одержувачів у вигляді рекомендацій. Ця технологія дозволяє користувачам витрачати мінімум часу для знаходження потрібної інформації в мережі Інтернет. Рекомендаційні системи порівнюють зібрані дані від користувачів і створюють список елементів, які рекомендовані користувачеві. Вони є альтернативою алгоритму пошуку, оскільки допомагають користувачам швидко знаходити предмети та інформацію, які б ті не змогли знайти самостійно [2]. Особливо широко рекомендаційні системи використовуються в електронній комерції [3–5]. Застосування рекомендаційних систем поширюється останнім часом на стаціонарну роздрібну торгівлю, довідкові центри, пошук з програмного забезпечення, наукових статтях і т. п. Це застосування характеризується наданням рекомендацій користувачам автоматично, на підставі вже вчинених дій (покупок, виставлених рейтингів, відвідувань тощо) і прийомом від них зворотного зв’язку (замовлення в магазинах, перехід за посиланнями). До найвідоміших інтернет-порталів, які використовують рекомендаційні системи

належать: Amazon.com, Inc. – американська компанія, найбільша в світі за оборотом серед інтернет-компаній, які продають товари та послуги, і один з перших інтернет-сервісів, орієнтованих на продажу реальних товарів масового попиту; eBay Inc. – американська компанія, що надає послуги у сфері інтернет-аукціонів (основне поле діяльності), інтернет-магазинів, миттєвих платежів, управляє веб-сайтом eBay.com і його місцевими версіями в декількох країнах, володіє компанією PayPal і Ebay Enterprise; MovieLens – рекомендаційна система і віртуальний веб-сайт спільноти, що рекомендує фільми для своїх користувачів, рекомендації надаються з врахуванням профілів (рейтингів) користувачів та використовують алгоритм спільної фільтрації (collaborative filtering); Rozetka.ua™ – сьогодні є найпопулярнішим інтернет-магазином електроніки і побутової техніки в Україні, представництва компанії є у всіх областях України. Рекомендаційні системи є одним із важливих розділів інтелектуальних систем підтримки прийняття рішення.

Аналіз останніх досліджень та публікацій

Рекомендаційні системи виникли і почали розвиватися з середини 90-х років минулого століття [10]. Основне завдання рекомендаційної системи – це надання персоналізованих рекомендацій користувачу, які враховують його уподобання при виборі предметів (товарів, об'єктів або послуг). Найширше застосування сьогодні отримали колаборативні рекомендаційні системи [6, 7]. Основна ідея алгоритмів колаборативної фільтрації полягає в пропозиції нових елементів для конкретного користувача на основі попередніх переваг користувача або думки інших однодумців користувача. Сьогодні дослідники розробили низку алгоритмів колаборативної фільтрації [8–11], які можна розділити на три основні категорії:

1. Методи, засновані на аналізі наявних оцінок (Memory-based). Цей підхід ще називають методом найближчих сусідів: використання попередніх оцінок, зроблених клієнтом, і аналіз оцінок інших користувачів, які мають подібні переваги. Тоді рекомендації (прогноз) для цільового користувача формуються на підставі обчислення певної міри схожості по всіх накопичених даних.

2. Методи, засновані на побудові моделі даних (Model-based). У цьому випадку спочатку за сукупністю оцінок формується описова модель переваг користувачів, товарів і взаємозв'язку між ними, а потім формуються рекомендації на підставі отриманої моделі. Процес формування рекомендацій розбитий на два етапи: ресурсомістке навчання моделі в відкладеному режимі і досить просте обчислення рекомендацій на основі існуючої моделі в реальному часі. Ці алгоритми можуть базуватися на імовірнісному підході, кластерному аналізі, аналізі прихованих чинників [10, 12].

3. Методи, засновані на об'єднанні попередніх алгоритмів, – гібридні методи [10–12].

До важливих задач розвитку рекомендаційних систем на даний час належать: підвищення точності прогнозування рекомендацій; вирішення проблеми впливу розрідженості і розмірності матриці користувач-предмет на точність прогнозування рекомендацій; вирішення проблеми нового користувача і нового предмету.

Останнім часом значна увага дослідників присвячена розробленню і дослідженню методів групової рекоменндації. В багатьох випадках надання рекомендацій для груп користувачів є більш доцільним, аніж надання рекомендацій для окремих користувачів [13–16].

Мета статті

Мета досліджень – розробити методи і алгоритми прогнозування рекомендацій для груп користувачів. Для вирішення поставленої задачі розробити метод змішаної категоріально-чисельної кластеризації. Розробити гібридний метод пошуку груп користувачів, який охоплює чисельну кластеризацію, змішану категоріально-чисельну і категоріальну кластеризацію.

Основні результати дослідження

Формальна постановка задачі прогнозування рекомендацій для груп користувачів полягає в такому. Нехай $\mathbf{U} = \{\mathbf{U_1}, \mathbf{U_2}, \mathbf{K}, \mathbf{U_n}\}$ – множина векторів профілів користувачів, $\mathbf{G} = \{\mathbf{G_1}, \mathbf{G_2}, \mathbf{K}, \mathbf{G_M}\}$ –

множина груп користувачів, $\mathbf{G_i} = \{\mathbf{U_{1Gi}}, \mathbf{U_{2Gi}}, \mathbf{K}, \mathbf{U_{kGi}}\}$ – множина профілів користувачів для групи $\mathbf{G_i}$. Необхідно здійснити прогноз рекомендацій для груп користувачів $\hat{r}_{Gi} = \text{Predict}(\mathbf{G_i})$. Узагальнену схему роботи методу прогнозування рекомендацій для спільнот користувачів наведено на рис. 1.

Для пошуку груп подібних користувачів використовуються методи кластеризації [10, 12].

Особливістю матриці користувач-предмет є те, що вона містить значну кількість нульових елементів. Кількість ненульвих елементів не перевищує 10 % від загальної кількості елементів матриці предмет-користувач [10, 12]. Тому для кластеризації користувачів у групи доцільно використовувати демографічні характеристики користувачів. Основними демографічними атрибутами користувачів є такі: вік, стать, освіта, рід занять. Вік – це числовий атрибут, стать, освіта, рід занять – категоріальні атрибути.

Нехай рейтинговий вектор профілю *i*-того користувача задається таким вектором (1)

$$\mathbf{U_i} = (u_{1i}, u_{2i}, \mathbf{K}, u_{mi}) \ , \tag{1}$$

де u_{ji} – рейтингова оцінка j-го предмета і-тим користувачем.

Розширимо цей вектор за допомогою демографічних атрибутів користувача (2)

$$\mathbf{U_i^{ext}} = (u_{1i}, u_{2i}, \mathbf{K}, u_{mi}, d_{1i}, d_{2i}, d_{3i}, d_{4i}) \ . \tag{2}$$

де $d_{1i}, d_{2i}, d_{3i}, d_{4i}$ – категоріальні атрибути користувача.

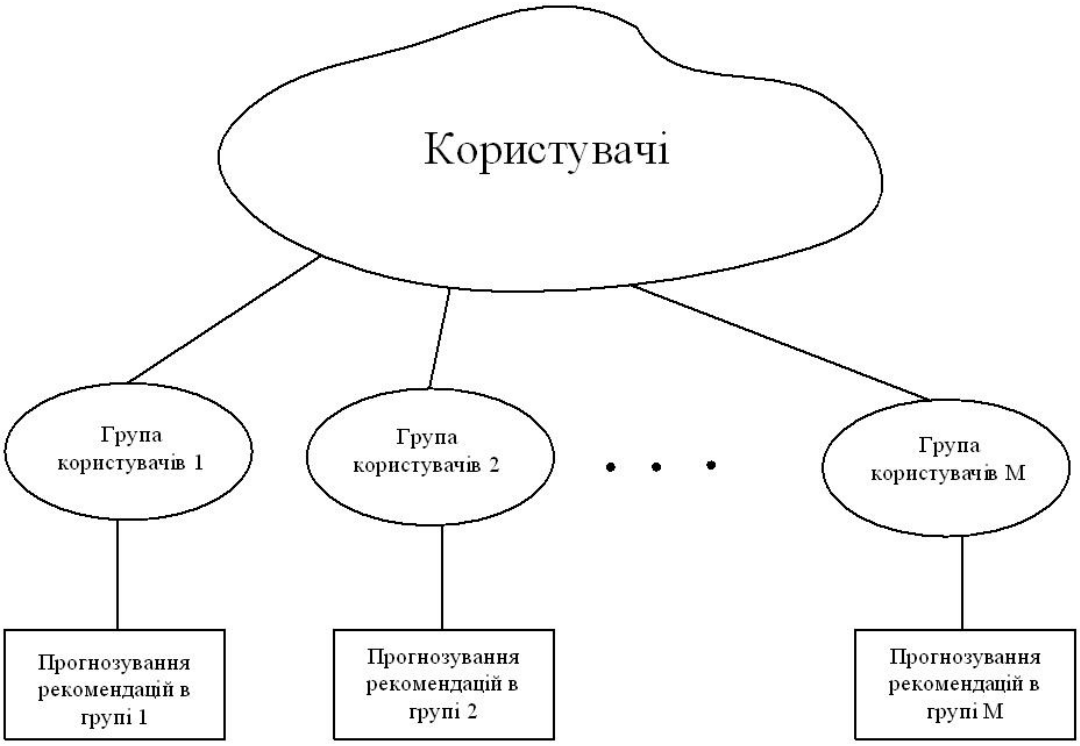

Рис. 1. Узагальнена схема роботи методу прогнозування рекомендацій для груп користувачів

Для спрощення опису методу позначатимемо вектор $\mathbf{U_i^{ext}}$ за допомогою вектора $\mathbf{X_i} = (x_{1i}, x_{2i}, \mathbf{K}, x_{ni})$.

Категоризація числового атрибута “вік” наведена в таблиці 1.

Так отримуємо мішаний вектор профілю користувача, який містить числові і категоріальні значення (2).

Кластеризація мішаних векторів профілів користувачів здійснюється за допомогою методу мішаної кластеризації. Метод мішаної кластеризації оснований на розрахунку щільності

розміщення мішаних векторів профілів користувачів і визначає кількість і положення центрів кластерів. Щільність визначається як кількість векторів профілів користувачів, які перебувають в околі радіусом d_c біля кожного користувача (3), (4)

$$r_i = \sum_{j=1}^{N} f(d_{ij} - d_c), \tag{3}$$

де d_{ij} – відстань між *i*-м та *j*-м векторами профілів користувачів; d_c – порогове значення; N – кількість користувачів.

$$f(x) = \begin{cases} 1, x = d_{ij} - d_c \le 0 \\ 0, x = d_{ij} - d_c > 0 \end{cases} \tag{4}$$

Таблиця 1

Категоризація атрибуту користувача – вік

Вік			
вік<18	18<вік<30	30<вік<50	50<вік
1	2	3	4

Відстань між векторами профілів користувачів обчислюється як середня зважена сума відстаней між атрибутами векторів профілів користувачів (5)

$$d_{ij} = D(\mathbf{X_i}, \mathbf{X_j}) = \frac{\sum_{k=1}^{N} w_{ij}^k l_{ij}^k}{\sum_{k=1}^{N} w_{ij}^k}, \tag{5}$$

де $w_{ij}^k = 0$, якщо атрибут d_{ij}^k відсутній і $w_{ij}^k = 1$ в протилежному випадку, l_{ij}^k – відстань між *i*-м і *j*-м атрибутамим об'єктів $\mathbf{X_i}$ і $\mathbf{X_j}$.

Відстань l_{ij} обчислюється окремо для числових і категоріальних атрибутів векторів $\mathbf{X_i}$ і $\mathbf{X_j}$.

Для числових атрибутів (6)

$$l_{ij}^k = \frac{\left| x_i^k - x_j^k \right|}{\max z_{ij}^k - \min z_{ij}^k}, \tag{6}$$

де $\max z_{ij}^k$ – максимальне значення в множині атрибутів векторів $\mathbf{X_i}$ і $\mathbf{X_j}$; $\min z_{ij}^k$ – мінімальне значення в множині атрибутів векторів $\mathbf{X_i}$ і $\mathbf{X_j}$.

Для категоріальних атрибутів (7)

$$l_{ij}^k = \begin{cases} 0, x_i^k = x_j^k \\ 1, x_i^k \ne x_j^k \end{cases} \tag{7}$$

Для знаходження положення і кількості центрів кластерів визначаються мінімальні відстані між множинами об'єктів з різними щільностями (8)

$$d_i = \min_j (d_{ij}),\ r_j > r_i \tag{8}$$

Будуються два вектори в порядку спадання значень $r_a > r_b$, $d_a > d_b$ (рис. 2).

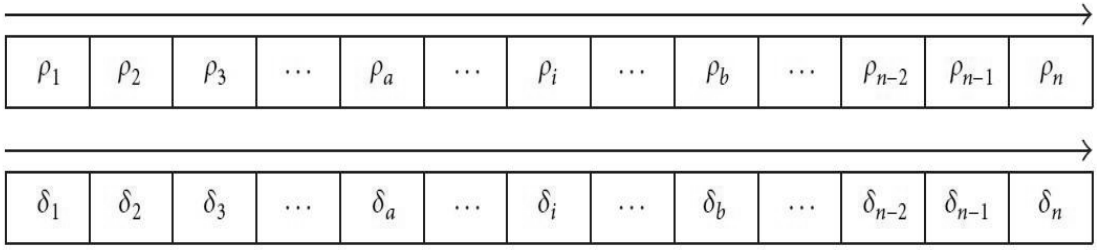

Рис. 2. Вектори щільностей і відстаней між множинами

Вважатимемо, що $r_1 - r_a$ – це "великі" значення щільностей, $r_b - r_n$ – це "малі" значення щільностей. Подібні умови накладемо на $d_1 \div d_a$ і $d_b \div d_n$.

Якщо виконується умова $r_i \in (r_1, r_a)$ і $d_i \in (d_1, d_a)$, тоді $\mathbf{X_i}$ центр наступного кластера.

Якщо виконується умова $r_i \in (r_b, r_n)$ і $d_i \in (d_1, d_a)$, тоді $\mathbf{X_i}$ об'єкт "шуму" і в подальших розрахунках не враховується. Графічна інтерпретація методу в 2D координатах наведена на рис. 3.

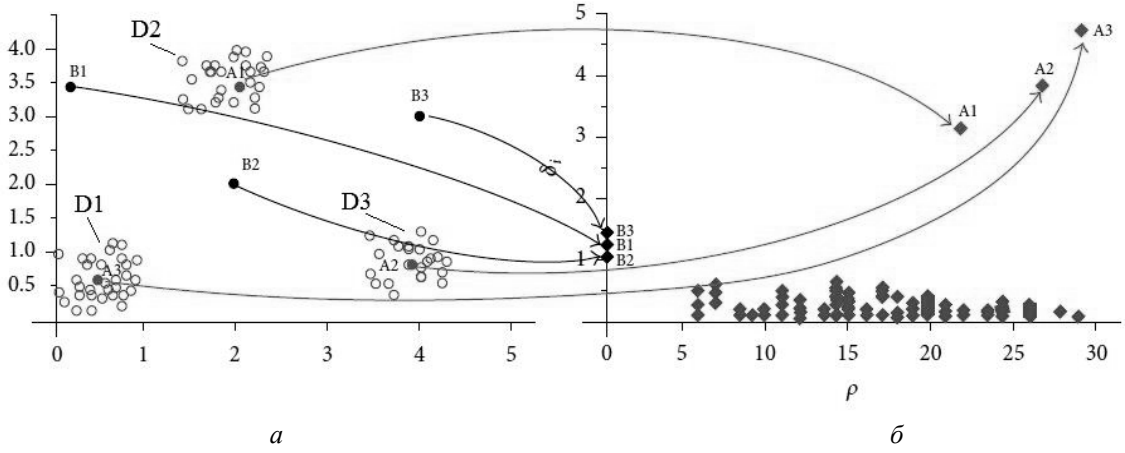

Рис. 3. Графічна інтерпретація мішаного методу кластеризації в 2D координатах

На рис. 3, а представлений приклад розподілу точок в 2D просторі. На рис. 3, б представлений приклад розподілу параметрів r і d для попереднього розподілу. D1, D2, D3 – області згущення точок з великою щільністю. A1, A2, A3 – центри кластерів, B1, B2, B3 – точки "шуму". Після визначення центрів кластерів здійснюється поділ об'єктів на кластери. Для пошук кластерів в системі прогнозування рекомендацій передбачено модифікований метод сканування щільності і модифікований метод k – середніх.

Результати пошуку груп користувачів великою мірою залежать від розрідженості матриці користувач-предмет. Розрідженість матриці користувач-предмет може бути розрахована за допомогою такого виразу (9)

$$SP = \frac{nR}{nUSER * nITEM}, \tag{9}$$

де nR – кількість відмінних від нуля елементів матриці користувач-предмет; $nUSER$ – кількість користувачів системи; $nITEM$ – кількість предметів в системі.

Розрідженість матриці користувач-предмет використовується в гібридному методі пошуку груп користувачів. Структурна схема методу представлена на рис. 4.

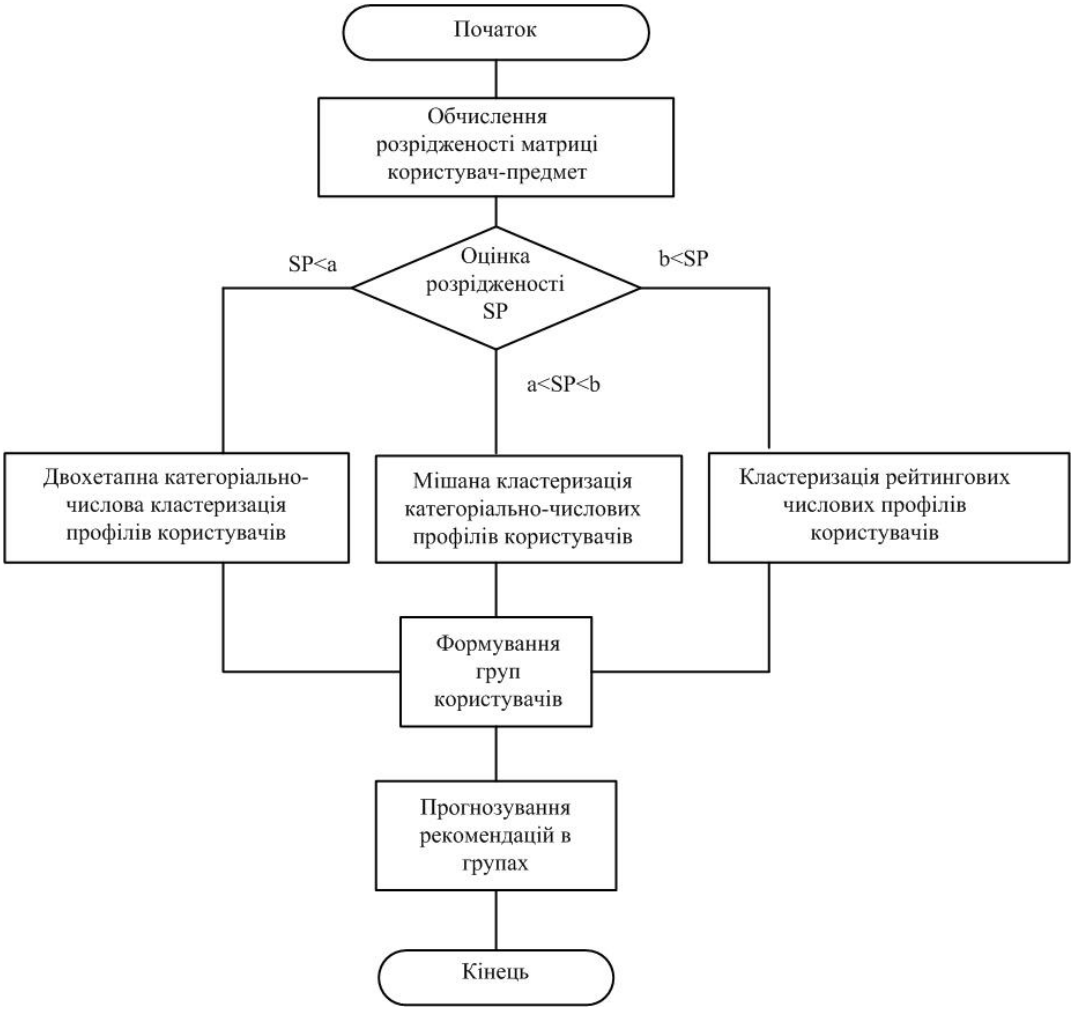

Рис. 4. Структурна схема гібридного методу кластеризації для пошуку груп подібних користувачів

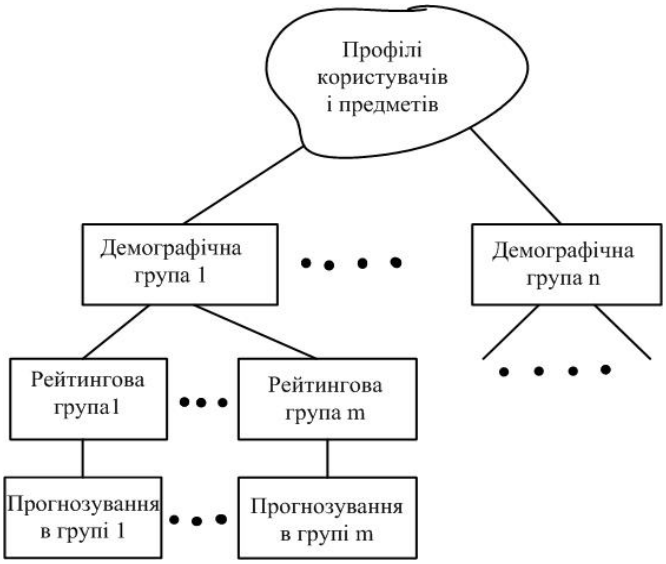

Рис. 5. Узагальнена схема двоетапного методу категоріально-числової кластеризації

Гібридний метод кластеризації для пошуку груп користувачів передбачає такі методи: модифікований метод неієрархічної кластеризації числових векторів профілів користувачів, який

базується на методі k-середніх; метод мішаної кластеризації категоріально-числових векторів профілів користувачів; двоетапний метод категоріально-числової кластеризації. Вибір методу здійснюється за допомогою оцінки розрідженості матриці користувач-предмет і двох параметрів *a* і *b*. За малої розрідженості використовують модифікований метод неієрархічної кластеризації числових векторів профілів користувачів, який оснований на методі k-середніх. При середньому значенні розрідженості використовується метод мішаної кластеризації категоріально-числових векторів профілів користувачів. За великої розрідженості використовують двоетапний метод категоріально-числової кластеризації.

Узагальнена схема двоетапного методу категоріально-числової кластеризації представлена на рис. 5.

На першому етапі здійснюється категоріальна кластеризація векторів демографічних профілів користувачів і будуються групи користувачів, які близькі за своїми демографічними характеристиками. На другому етапі здійснюється числова кластеризація векторів профілів користувачів, які містять числові рейтингові оцінки предметів. Прогнозування рекомендацій в отриманих групах може бути виконано як класичним методом колаборативного прогнозування користувач-користувач або предмет-предмет, так і методом прогнозування в групі (методом адитивної корисності, або методом мультиплікативної корисності [13–16]). Для категоріальної кластеризації векторів демографічних профілів користувачів використовується модифікований метод ROCK (A Robust Clustering Algorithm for Categorical Attributes) [17].

ROCK являє собою агломеративний ієрархічний алгоритм кластеризації категоріальних атрибутів. Замість звиклої міри близькості для методів числової кластеризації (норма Евкліда) в методі ROCK вводиться міра зв'язку між атрибутами і множинами атрибутів, яка не задовольняє аксіом метричного простору, однак ефективна при кластеризації категоріальних векторів у n-мірному просторі. Однак метод ROCK може будувати хибні кластери на кінцевій стадії кластеризації. Модифікований метод ROCK вимагає менше часу для розрахунку і вирішує проблему кластеризації на кінцевій стадії.

Висновки

Розроблено метод мішаної категоріально-числової для пошуку груп користувачів у рекомендаційних системах з урахуванням демографічних характеристик користувачів. Розроблено гібридний метод пошуку груп користувачів, який включає метод числової неієрархічної кластеризації, метод мішаної категоріально-числової кластеризації, двохетапний метод кластеризації. Двохетапний метод здійснює на першому етапі категоріальну кластеризацію, на другому етапі числову кластеризацію. Показано застосування розроблених методів в колаборативній рекомендаційній системі.

1. J. A. Konstan Recommender systems: from algorithms to user experience / J. A. Konstan J. A. // User Modeling and User-Adapted Interaction. – 2012 – Vol. 22. – No. 1–2. – P. 101–123. 2. Schafer J. B. E-Commerce Recommendation Applications / J. B. Schafer J. B., J. A. Konstan, J. Riedl // Data Mining and Knowledge Discovery. – 2001. – Vol. 5 – No. 1–2. – P. 115–123. 3. Sarwar B. Analysis of recommendation algorithms for e-commerce / B. Sarwar, G. Karypis, J. Konstan, J. Riedl // In Proceedings of the 2nd ACM conference on Electronic. – Minnesota, USA – October 17–20, 2000. – P. 158–167. 4. Pu P, Chen L, Hu R. A user-centric evaluation framework for recommender systems / P. Pu, L. Chen, R. Hu // In: Proceedings of the fifth ACM conference on Recommender Systems (RecSys'11), ACM. – New York, NY, USA. – 2011. – P. 57–164. 5. я к 2. 6. Candillier L. Comparing State-of-the-Art Collaborative Filtering Systems / L. Candillier, F. Meyer, M. Boullé // In Proceedings of the 5th International Conference on Machine Learning and Data Mining in Pattern Recognition, LNCS. – Vol. 4571. – 2007. – P. 548–562. 7. Su X., Khoshgoftaar T. M. A survey of collaborative filtering techniques / X. Su, T. M. Khoshgoftaar // Adv. Artif. Intell. – Vol. 4571. – 2007 – P. 1–19. 8. Isinkaye F. O. Recommendation systems: Principles, methods and evaluation / F. O. Isinkaye, Y. O. Folajimi,

B. A. Ojokoh // Egyptian Informatics Journal. – Vol. 16. – 2015. – P.261–273. 9. Das D. A Survey on Recommendation System / D. Das, L. Sahoo, S. Datta // International Journal of Computer Applications. – Vol. 160. – No. 7. – 2017. – P.6–10. 10. Bobadilla J. Recommender systems survey / J. Bobadilla, F. Ortega, A. Hernando, A. Gutiérrez // Knowledge-Based Systems. – Vol. 46. – 2013. – P. 109–132. 11. Resnick P., Varian H. R. Recommender systems / P. Resnick, H. R. Varian // Communications of the ACM. – Vol. 40. – 1997. – P. 56–58. 12. G. Adomavicius, A. Tuzhilin Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions / Adomavicius G., Tuzhilin A. // IEEE Transactions on Knowledge and Data Engineerin. – Vol. 17. – 2005. – P. 734–749. 13. Jameson A., Smyth B. Recommendation to groups / Jameson A., Smyth B. // In The adaptive web: methods and strategies of web personalization. – 2007. – P. 596–627. 14. Konstan J. GroupLens: applying collaborative filtering to usenet news / J. Konstan, B. Miller, D. Maltz, J. Herlocker, L. Gordon, J. Riedl // Commun. ACM – Vol. 40. – No. 3. – 1997. – P.77–87. 15. J. Masthoff Group modeling: selecting a sequence of television items to suit a group of viewers / J. Masthoff // User Model. User-Adap. Inter. – Vol. 14. – No. 1. – 2004 – P. 37–85. 16. L. Boratto, S. Carta, "State-of-the-art in group recommendation and new approaches for automatic identification of groups" / L. Boratto, S. Carta // In Information Retrieval and Mining in Distributed Environments. – Vol. 324. – Springer Berlin Heidelberg – 2011. – P. 1–20. 17. Guha S. Rock: A robust clustering algorithm for categorical attributes / S. Guha, R. Rastogi, K. Shim // Information Systems. – Vol. 25. – No. 5. – 2000. – P. 345–366.