УДК 004.67 DOI: 10.20837/0135-1710.2023.179.032

П.Е. СИТНІКОВА, М.О. ГРЕБЕНЮК

РЕКОМЕНДАЦІЙНА СИСТЕМА НА ОСНОВІ КОМПАКТНОЇ ГІБРИДНОЇ МОДЕЛІ КОРИСТУВАЧА

Запропоновано компактну гібридну модель користувача для методу спільної фільтрації інформації. Ця модель долає обмеження, які часто виникають при використанні традиційних підходів, і дозволяє ефективніше генерувати персоналізовані рекомендації. Для наочного прикладу цієї моделі було взято рекомендаційну систему кінофільмів. Запропонована модель пов'язує оцінки користувачів з деяким описом вмісту об'єктів. Для цього введено поняття міри цікавості жанру, що являє собою гібридну ознаку, яка поєднує оцінки користувачів і жанри фільмів, та представляє уподобання користувача на рівні моделі. Таким чином, запропонована модель зберігає точність спільної фільтрації на основі пам'яті та масштабованість на основі моделі. Це сприяє значному зменшенню складності системи, її розрідженості та забезпечує виконання взаємовідношення транзитивності між сусідами.

1 Вступ

У сучасному цифровому світі, коли обсяг інформації та доступного матеріалу надзвичайно великий, рекомендаційні системи грають важливу роль у полегшенні процесу вибору продуктів, послуг та контенту для користувачів. Вони допомагають надавати індивідуальні рекомендації, враховуючи уподобання та інтереси користувачів, що робить процес прийняття рішень зручнішим та ефективнішим. Рекомендаційні системи використовують різні алгоритми та методи для аналізу даних і надання персоналізованих рекомендацій. Такі системи діють в різних галузях, включаючи електронну комерцію, музику, фільми, книги та інші області. Вони можуть базуватися на аналізі історії оцінок користувача, на основі схожості переглянутих об'єктів, які знаходяться у системі, та на інших даних.

У цьому контексті однією з важливих складових вподобань користувачів є така ознака класифікації об'єктів, як жанри фільмів, музичних композицій, літератури тощо. Знання жанру допомагає користувачам знаходити той контент, який їх цікавить, і зробити вибір серед маси доступних альтернатив. Проте, для ефективної роботи рекомендаційних систем необхідно враховувати не лише жанр, але й інші фактори, такі як історія взаємодій користувачів, рейтинги, відгуки та інші параметри.

Тут на допомогу приходять гібридні рекомендаційні системи, які поєднують різні підходи та методи, включаючи інформацію про жанри, для надання більш точних та персоналізованих рекомендацій. Гібридні системи дозволяють враховувати індивідуальні смаки та вподобання користувачів, а також забезпечують більш масштабовані та ефективні рекомендації, навіть у випадках, коли об'єм доступного контенту дуже великий.

Таким чином, гібридні рекомендаційні системи важливі для рекомендацій об'єктів з різними жанрами, оскільки вони допомагають підвищити точність, рівень персоналізації та релевантність рекомендацій для користувачів, що робить процес вибору контенту більш зручним і задовольняє індивідуальні потреби кожного користувача. Тому розробка саме таких систем є актуальною і важливою задачею та має практичну значущість.

2. Аналіз існуючих методів і постановка проблеми

Для керування обчисленням подібності на основі доступних вхідних даних існує багато методів фільтрації інформації. В загальному випадку кожен метод фільтрації можна віднести до одного з таких класів: демографічний, на основі вмісту, спільний або гібридний [2]. Далі розглянемо різні методи фільтрації інформації з відповідними вхідними даними (табл. 1).

Таблиця 1

Методи фільтрації інформації

Метод фільтрації інформації		Демографічна інформація	Опис вмісту	Рейтинг
Демографічна фільтрація (DF)		✓		
Фільтрація на основі вмісту (CBF)			✓	
Спільна фільтрація (CF)				✓
Гібридна фільтрація	DF/CBF	✓	✓	
	DF/CF	✓		✓
	CBF/CF		✓	✓
	DF/CBF/CF	✓	✓	✓

Демографічна фільтрація (DF) [1] використовує описи користувачів для того, щоб визначити взаємозв'язок між конкретним об'єктом і типом користувачів, яким він подобається. Відповідно, користувачеві будуть рекомендовані товари, подібні до тих, які вподобають інші люди з тієї ж демографічної групи. У [3] розглядається саме цей підхід для генерації рекомендацій. Перевага демографічного підходу полягає в тому, що він може не вимагати історії рейтингів, необхідної для спільної фільтрації та фільтрації на основі вмісту. Основні недоліки демографічної фільтрації такі:

– труднощі отримання даних. Важко отримати особисту інформацію від користувачів. Вони зазвичай залишаються анонімними або не розголошують свої особисті дані [4]. Здебільшого це пов'язано з недовірою до політики конфіденційності сучасних веб-сайтів. Таким чином, користувачі або приховують персональні дані, або надають неправдиві дані;

– проблема сірої вівці. Це стосується користувачів з унікальними вподобаннями та смаками, які ускладнюють розробку точних профілів. Вирішення саме цієї проблеми запропонували у [5].

Фільтрація на основі вмісту (CBF) використовує описи вмісту об'єктів [1], які сподобалися користувачеві в минулому, щоб зробити для нього нові рекомендації. Тому йому будуть рекомендовані об'єкти, схожі на ті, яким він віддавав перевагу в минулому. Цей метод визначає об'єкти за пов'язаними з ними ознаками. Наприклад, у [6] для рекомендацій використовуються теги та жанр музики. Рекомендаційні системи на основі вмісту не підходять для динамічних і дуже великих середовищ, де об'єктів мільйони і вони часто додаються в систему. Основними недоліками фільтрації на основі вмісту є:

– обмеження вмісту. Можна надати лише дуже поверхневий аналіз певних видів контенту;

– надмірна спеціалізація [2]. Користувач може бачити лише предмети, подібні до вже оцінених ним. Це обмеження не дозволяє користувачеві отримувати випадкові рекомендації;

– проблема нового користувача. Система повинна дізнатися уподобання користувача з описів об’єктів, які йому подобалися раніше. Тому користувач повинен оцінити деякі об’єкти, щоб система могла дійсно зрозуміти вподобання користувача і дати надійні та точні рекомендації.

Спільна фільтрація (CF) [1] стимулює соціальний процес пошуку у сусідів рекомендацій щодо об'єктів, які раніше не були розглянуті користувачем. Таким чином, будуть рекомендовані предмети, які подобалися людям зі схожими смаками та уподобаннями в минулому. Велика сила спільної фільтрації по відношенню до фільтрації на основі вмісту полягає в тому, що вона не потребує апріорної інформації про властивості об'єктів (наприклад, жанр фільму чи тип продукту) і здатна пропонувати рекомендації, які можуть бути неочевидними чи "нестандартними". Більше того, спільна фільтрація є повністю незалежною від будь-якого машиночитаного представлення рекомендованих об’єктів. На відміну від фільтрації на основі вмісту, спільна та демографічна фільтрація можуть надавати випадкові рекомендації. Однак спільна фільтрація окрім вже згаданих проблеми нового користувача та проблеми сірої вівці має ще такі слабкі сторони:

– проблема нового об’єкта. Новий об’єкт, який має дуже малу кількість оцінок, не можна легко рекомендувати. Ця проблема також відома як проблема «холодного старту», оскільки перший користувач, який оцінить об’єкт, отримує від цього незначну користь. Такі ранні оцінки не дозволяють користувачам ефективно порівнювати свої вподобання з вподобаннями інших користувачами;

– розрідженість. Зазвичай кожен користувач оцінює лише дуже обмежений відсоток об’єктів у порівнянні з наявною загальною кількістю. Це призводить до розріджених матриць "користувач-об’єкт", і тому можуть бути створені слабкі рекомендації, оскільки успішні сусіди не можуть бути знайдені. Тому було запропоновано багато методів зменшення розмірів для зменшення ефекту розрідженості (наприклад, у [7]);

– втрата транзитивності сусідів. Пряма опора на рейтинги об’єктів призводить до втрати транзитивності сусідів. Припустимо, що ми маємо трьох користувачів u_i, u_j і u_k. Користувачі u_i і u_j дуже корелюють, також u_j і u_k дуже корелюють. Через транзитивність існує ймовірність того, що користувачі u_i та u_k також сильно корелюють. Такий перехідний зв’язок не відображається в традиційних методах спільної фільтрації, якщо користувачі u_i та u_k не оцінили багато спільних об’єктів.

Гібридні системи [1] поєднують два або більше методів фільтрації, які доповнюють один одного для досягнення кращої продуктивності.

У таблицях 2, 3, 4 наведено зразки профілів для користувача [8] відповідно до демографічної фільтрації, фільтрації на основі вмісту та спільної фільтрації.

Таблиця 2

Профіль користувача для демографічної фільтрації

№	Пол	Вік	Професія	Освіта	…
1	Чоловік	23	Програміст	Магістр	…
…	…	…	…	…	…

Таблиця 3

Профіль користувача для фільтрації на основі вмісту

№	Режисер	Актор	Рік	Країна		…
1	Крістофер Нолан	Меттью Макконехі	2014	США		…
…	…	…	…	…		…

Таблиця 4

Профіль користувача для спільної фільтрації

№	Тітанік	Зоряні війни	Месники	Форсаж	…
1	5	4	2	3	…
…	…	…	…	…	…

Різниця між профілем користувача та моделлю користувача полягає в різному рівні складності. Профіль користувача — це просто сукупність інформації, зібраної від користувача, яку можна описати простою моделлю. Залежно від змісту та кількості інформації про користувача, яка зберігається в профілі, можна побудувати модель. Точніше, модель користувача — це уявлення про знання та особистісні характеристики, якими, на думку системи, володіє користувач. Відповідно, рекомендаційні системи можна класифікувати на системи на основі пам'яті або на основі моделі [2]. Для того, щоб побудувати модель користувача, необхідно вирішити, які конкретні ознаки або параметри будуть включені до моделі користувача для того, щоб рекомендаційна система могла краще розуміти і передбачати інтереси та потреби користувача. Від цього залежить якість та ефективність системи рекомендацій. Рекомендаційні системи на основі пам'яті використовують просту модель та весь набір даних для процесу рекомендацій. Хоча система, що базується на пам'яті, є простою і надає рекомендації високої точності, вона є дорогою з точки зору обчислень, оскільки розмір вхідного набору даних збільшується. З іншого боку, рекомендаційна система на основі моделі отримує модель з офлайн-даних, які будуть використовуватися для онлайн-рекомендацій. Після формування моделі користувача матриця об'єктів користувача більше не потрібна для генерації рекомендацій. Це призведе до систем із нижчими вимогами до пам'яті та швидкого генерування рекомендацій. Хоча такі системи дозволяють зменшити витрати на обробку в режимі онлайн, часто це супроводжується втратою точності в рекомендаціях.

Аналіз існуючих методів підкреслює актуальність підвищення ефективності та точності систем рекомендацій в умовах зростаючого обсягу інформації та складності задачі забезпечення користувачів персоналізованими рекомендаціями. З ростом обсягу даних та різноманітності інтересів користувачів традиційні методи рекомендаційних систем стикаються з численними обмеженнями, такими як розрідженість даних, недостатній обсяг оцінок користувачів та обмеженість інформації про вміст об'єктів тощо. Ці обмеження можуть призводити до низької точності та обмеженої масштабованості систем рекомендацій, ускладнюючи надання користувачам рекомендацій, які відповідають їхнім індивідуальним потребам та інтересам.

3. Мета і задачі дослідження

Метою даного дослідження є розробка покращеної моделі користувача для рекомендаційних систем для поліпшення точності та ефективності у ситуаціях, коли

інформація стає дедалі об'ємнішою, і завдання забезпечення персоналізованих рекомендацій стають складнішими.

Для досягнення цієї мети пропонується вирішити такі задачі:

1. Розробити компактну гібридну модель користувача, яка буде долати проблеми існуючих методів, такі як розрідженість даних, недостатній обсяг оцінок користувачів та обмеженість інформації про вміст об'єктів. Для цього ввести гібридну ознаку, яка буде точно показувати інтерес користувача до різних жанрів і поєднає оцінки користувачів та інформацію про фільми, а саме про їх жанр;

2. Провести обчислення, спрямовані на демонстрацію та підтвердження ефективності розробленої моделі користувача та її ключової характеристики, а саме гібридної ознаки.

4. Компактна гібридна модель користувача

Спільна фільтрація на основі пам’яті є точнішою, але її масштабованість, порівняно з рекомендаційною системою на основі моделі, гірша. Крім того, фактичні уподобання користувача не завжди можуть бути охоплені лише оцінками, і тому потрібні деякі описи вмісту об’єктів. Цього можна досягти, якщо ми побудуємо гібридну модель користувача. Ідея об'єднання або інтеграції багатьох джерел інформації може значною мірою полегшити проблеми розрідженості та масштабованості. Під час пошуку інформації ознаки з відповідних джерел часто додаються до базового представлення. Це зазвичай називають розширенням запиту. Цей спосіб не допоможе в нашому випадку, тому що якщо ми подолаємо проблему розрідженості, ефект проблеми масштабованості збільшиться. Однак ми хочемо одночасно зменшити вплив цих проблем. Для цього запропоновано набір функцій для обчислення гібридних ознак, що поєднують деякі властивості користувачів і фільмів.

Для наочного прикладу моделі, що пропонується, було взято рекомендаційну систему, яка пропонує користувачам кінофільми. Система рекомендацій повинна отримувати ідентифікатор користувача як вхідні дані і повертати список фільмів та інформацію про те, наскільки цікавими вони будуть для користувача. Ця інформація, фактично, є прогнозом системи на основі її алгоритму і може бути інтерпретована, наприклад, як повідомлення користувачу про топ-10 фільмів для нього. Система рекомендацій також повинна відповідно обчислювати рейтинги для всіх відомих фільмів, які користувач не дивився, і для яких отримано унікальний номер як вхідні дані. Для створення такої системи рекомендацій потрібна база даних, яка повинна складатися з двох частин. Перша частина містить список усіх користувачів системи та інформацію про їх взаємодію з фільмами (рейтинги та перегляди). Друга частина - це детальна та обширна база даних про самі фільми. Отже, перед створенням системи рекомендацій повинні бути доступні такі вхідні дані:

– вся доступна інформація про фільми: назва, короткий сюжет, жанр, ключові слова, акторський склад, режисер та ін.;

– інформація про користувачів: дані про користувача, для кожного користувача – конкретний набір деякої кількості рейтингів, кожен з яких пов'язаний з певним фільмом.

Формально маємо множину користувачів $U = \{u_1 \dots, u_M\}$, де M – кількість користувачів, які мають явні або неявні оцінки множини об’єктів (фільмів) $S = \{s_1 \dots, s_K\}$, де K – кількість об’єктів.

Формально маємо множину користувачів $U = \{u_1 \dots, u_M\}$, де M – кількість користувачів, які мають явні або неявні оцінки множини об’єктів (фільмів) $S = \{s_1 \dots, s_K\}$, де K – кількість об’єктів.

Множини S та U є великими, а у деяких випадках можуть бути величезними. Кожен

користувач $u_i, i = 1,\dots,M$, оцінив підмножину об'єктів S_i. Оцінку користувача u_c для об'єкту s_k, $k = 1,\dots,K$, позначено як $r_{c,k}$. Усі доступні оцінки збираються в $(M \times K)$ матрицю "користувач-об'єкт", позначену як R. Архітектура різних систем рекомендацій може бути централізованою або розподіленою. У цій роботі представлена централізована архітектура, де система рекомендацій розташована в одному конкретному місці. Під час розробки системи рекомендацій можна виділити такі п'ять етапів [1]:

– збір даних;
– формування моделі користувача;
– обчислення схожості;
– вибір сусідів;
– прогнози та рекомендації.

Взаємодію цих етапів можна описати схемою, наведеною на рис. 1.

Рис. 1. Схема етапів процесу рекомендації

На основі рейтингів користувача для набору високо оцінених фільмів та описів жанрів, які відповідають цьому набору фільмів, за допомогою формул (1)-(8) виведено «Міру цікавості жанру» (Genre Interestingness Indicator - *GII*) (9) - гібридну ознаку, яку детально описано у [9] і яка поєднує оцінки користувачів та жанри фільмів. Нижче описано процес виведення «Міри цікавості жанру».

Для певного користувача u_i і певного жанру, G_i набір фільмів з високою оцінкою H_i є:

$$H_i = \{r_{i,k} : s_k \in S_i, r_{i,k} \geq 3\}, \quad (1)$$

Тут S_i — набір фільмів, оцінених u_i.

Загальна оцінка $TR(u_i)$ користувача u_i:

$$TR(u_i) = \sum_{s_k \in S_i} r_{i,k}. \quad (2)$$

Жанровий рейтинг $GR_{u_i}(G_j)$ (відповідно частота жанру, $GF_{u_i}(G_j)$) для високо оцінених фільмів жанру G_j, що відповідають користувачу u_i:

$$GR_{u_i}(G_j) = \sum_{s_k \in G_j \subset H_i} r_{i,k}, \tag{3}$$

$$GF_{u_i}(G_j) = \sum_{s_k \in G_j \subset H_i} \delta_z(r_{i,k}), \quad z \in \{3,4,5\}, \tag{4}$$

де

$$\delta_z(r_{i,k}) = \begin{cases} 1 & z = r_{i,k} \\ 0 & z \neq r_{i,k} \end{cases}. \tag{5}$$

Слід зазначити, що ми розглядаємо лише фільми, які мають рейтинг $r_{i,k} \geq 3$, тобто фільми, оцінені як "добре" – 3, "дуже добре"– 4 або "відмінно" – 5. Нарешті, відносний жанровий рейтинг (RGR) (6) є відношенням жанрового рейтингу (3) користувача u_i до загальної оцінки користувача (2). Відповідно, відносна частота жанру (RGF) (7) є відношенням частоти жанру (4) користувача u_i до загальної частоти оцінок користувача.

$$RGR_{u_i}(G_j) = \frac{GR_{u_i}(G_j)}{TR(u_i)}, \tag{6}$$

$$RGF_{u_i}(G_j) = \frac{GF_{u_i}(G_j)}{TF(u_i)}, \tag{7}$$

де загальна частота $TF(u_i) = |S_i|$ є потужністю S_i. Схема для запропонованої моделі користувача наведена на рис. 2. Перш за все, для створення гібридних ознак використовуються жанри та явні оцінки.

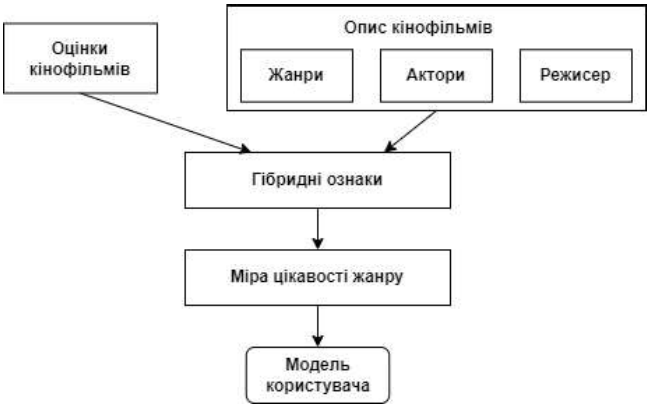

Рис. 2. Схема компактної гібридної моделі користувача

Як простий приклад розглянемо табл. 5 лише для трьох користувачів, які оцінили фільми, що належать до чотирьох жанрів. У стовпцях $G_j, j = 1,2,3,4$ значення "1" вказує на належність даного фільму до G_j, а "0" – в іншому випадку. Також ненульове значення в стовпцях оцінок користувачів вказує на те, що фільм був оцінений, а нуль вказує, що фільм ще не оцінений.

У табл. 6 наведено рейтинги, пов’язані з різними жанрами. Слід зазначити, що оцінка конкретного фільму пов’язується з усіма жанрами, до яких належить фільм. Наприклад, оцінки User-3 для G_4 — це 4, 5 і 3, де вони відповідають фільму-5, фільму-8 та фільму-11 (табл. 5) відповідно. Як обговорювалося раніше, гібридні ознаки використовуються як основа для вимірювання цікавості жанру, тому користувача більше цікавить G_i, якщо він має вищі оцінки, тобто "добре", "дуже добре" або "відмінно". Кількість гібридних ознак залежить від кількості жанрів. На основі цього можна зробити висновок, що User-1, User-2 і User-3 більше цікавляться G_3, G_1 і G_4 відповідно (табл. 6). Хоча деякі жанри мають низькі рейтинги

Таблиця 5

Оцінки користувачів

Фільм	Жанри фільмів				Рейтинги користувачів		
	G_1	G_2	G_3	G_4	User-1	User-2	User-3
1	1	0	1	0	1	5	0
2	1	0	1	1	3	0	0
3	1	1	0	0	1	5	3
4	0	1	1	0	5	1	0
5	0	1	1	1	4	0	4
6	0	1	0	0	0	2	0
7	1	1	0	1	0	0	0
8	0	0	1	1	0	0	5
9	0	1	1	0	0	2	0
10	1	1	1	0	0	4	1
11	1	1	0	1	0	3	3
12	1	0	0	0	0	0	3

Таблиця 6

Оцінки користувачів по жанрах

User	TF	TR	G_1	G_2	G_3	G_4
1	5	14	1,3,1	1,5,4	1,3,5,4	3,4
2	7	22	5,5,4,3	5,1,2,2,4,3	5,1,2,4	3
3	6	19	3,1,3,3	3,4,1,3	4,5,1	4,5,3

($r_{i,k} < 3$), їх можна відфільтрувати на етапі спільної фільтрації. Необхідно знайти компактну модель користувача, щоб виявити найближчих сусідів для цього користувача. Після цього отриманий набір сусідів використовується для спільної фільтрації, щоб рекомендувати певні фільми, які можуть сподобатися користувачеві. Тому не можна рекомендувати активному користувачеві будь-який фільм із прогнозованим рейтингом, меншим за "добре", тобто 3.

Для того, щоб дослідити можливість вибору формули для GII, ми розглянули формули (6) і (7). Формула для RGR (6) добре працює для User-1 та User-2, але для User-3 вона дає G_1 і G_3 однакові значення (табл. 7), тоді як оцінки користувача для фільмів цих двох жанрів досить різні. G_1 має три оцінки "добре" -3, а G_3 має оцінки "дуже добре" -4 і "відмінно" -5. Це тому, що кількість фільмів для кожного жанру ігнорується в цій формулі. Формула для RGF (7) добре працює для User-2, але для User-3 вона дає G_1, G_2 і G_4 однакові значення. Крім того, значення G_3 нижче, ніж у G_1 і G_2, тоді як оцінки для G_3 – "дуже добре" та "відмінно". Це пов’язано з тим, що враховується лише кількість фільмів із високим рейтингом, і всі високі оцінки вносяться в цю формулу однаково. Очевидно, що RGR і RGF є формулами GII з деякими недоліками. Насправді, користувач представляє своє вподобання до фільму за допомогою надання оцінки цьому фільму, наприклад "добре", "дуже добре" або "відмінно". Цю детальність уподобань неможливо відобразити, якщо ми візьмемо просте усереднення частот, оскільки воно надає всім високим рейтингам однакову вагу. Тому введено нову версію формули RGF, яка намагається відобразити точне уподобання до фільмів з високими

рейтингами (схожий підхід наведено у [10]). Тобто модифікована відносна частота ($MRGF$) жанру G_j для користувача u_i визначається як:

$$MRGF_{u_i}(G_j) = \frac{\sum_{s_k \in G_j \subset H_i} \delta_3(r_{i,k}) + 2\times\delta_4(r_{i,k}) + 3\times\delta_5(r_{i,k})}{3\times TF(u_i)} \quad (8)$$

Результати обчислень, наведені у табл. 7, показують, що, як і очікувалося, $MRGF$ долає недоліки RGF.

Таблиця 7

Результати обчислень

GII формула	User	G_1	G_2	G_3	G_4
RGR	1	0.214	0.643	0.857	0.500
	2	0.773	0.545	0.409	0.136
	3	0.474	0.526	0.474	0.632
RGF	1	0.200	0.400	0.600	0.400
	2	0.571	0.429	0.286	0.143
	3	0.500	0.500	0.333	0.500
MRGF	1	0.067	0.333	0.400	0.200
	2	0.429	0.286	0.238	0.048
	3	0.167	0.222	0.278	0.333

Для того, щоб розробити точніший GII, необхідно також враховувати відносний жанровий рейтинг. Поєднання формул RGR і $MRGF$ допоможе взаємно усунути недоліки, які притаманні окремим формулам. Для випадку жанру G_j для користувача u_i міра цікавості жанру (GII) визначається як:

$$GII_{u_i}(G_j) = \frac{2\times nf\times RGR_{u_i}(G_j)\times MRGF_{u_i}(G_j)}{RGR_{u_i}(G_j)\times MRGF_{u_i}(G_j)}, \quad (9)$$

де nf - коефіцієнт нормалізації для даної системи.

Формула (9) дає гармонічне середнє значення $RGR_{u_i}(G_j)$ і $MRGF_{u_i}(G_j)$, помножене на nf. Діапазон $GII_{u_i}(G_j)$ становить $[0, \text{MAX}]$ Це узгоджується зі структурою рейтингу системи, тобто $1 -$ "дуже погано",...,MAX $-$ "відмінно". Тут MAX - найвищий можливий рейтинг для даної системи. Коефіцієнт нормалізації nf приймає значення MAX або глобального середнього рейтингу користувача ($TR(u_i)/TF(u_i)$). Результати обчислень на основі формули (9) з $nf = MAX$ показані в табл. 8. Ці результати відображають саме ступінь цікавості для кожного

Таблиця 8

Результати обчислень GII

Користувач	$GII_{u_i}(G_j)$			
	G_1	G_2	G_3	G_4
1	0.510	2.194	2.727	1.429
2	2.759	1.876	1.505	0.355
3	1.235	1.561	1.752	2.181

жанру: User-1 більше зацікавлений у G_3, User-2 – G_1, User-3 – в G_4. За допомогою формули (9) отримуються три основні переваги. По-перше, уподобання користувачів у жанрах легко обчислити. По-друге, отримується компактна модель користувача. По-третє, GII гарантує транзитивність сусіда, тому користувачі, які мають близький GII, будуть корелювати між собою (мати схожі смаки).

Дана компактна модель користувача вводить гібридизацію на двох різних рівнях, а саме, на рівні пам'яті і на рівні моделі. Гібридні функції використовують як оцінки користувачів для фільмів з високим рейтингом, так і деякі описи вмісту фільмів (жанрів). На рівні моделі компактна модель користувача використовується для пошуку набору однодумців, серед яких здійснюється пошук на основі пам'яті. Цей набір набагато менший за розміром, ніж оригінальний набір, що робить систему масштабованою. Очевидно, що $GII_{u_i}(G_j)$ представляє інтерес u_i користувача до жанру G_j на основі оцінок фільмів у жанрі G_j, що були надані іншими користувачами.

5. Висновки та перспективи подальших досліджень

Розглянуто та проаналізовано різні методи фільтрації інформації для генерації персоналізованих рекомендацій. Розроблено компактну гібридну модель користувача для методу спільної фільтрації. Така модель пов'язує оцінки користувачів з деяким описом вмісту об'єктів. Для цього введено поняття міри цікавості жанру, що являє собою гібридну ознаку, яка поєднує оцінки користувачів і жанри фільмів, та представляє уподобання користувача на рівні моделі. Такий підхід дозволяє зберігати точність спільної фільтрації на основі пам'яті та водночас забезпечує масштабованість завдяки моделі користувача. Це робить процес рекомендації компактнішим та швидшим у порівнянні із стандартними методами. Важливо відзначити, що ця модель допомагає вирішити проблеми розрідженості даних та обмеженої інформації про об'єкт, які пов'язані з розглянутими у статті підходами, і водночас зберігає транзитивність відносин між сусідніми користувачами. Для можливого поліпшення результатів у майбутніх дослідженнях можна розглянути впровадження модифікованого методу фільтрації з використанням алгоритмів нечіткої логіки, що допоможе уникнути недоліків при побудові моделі з ознаками, яким притаманна нечіткість та невизначеність.

Перелік посилань:

1. *Charu C. Aggarwal.* Recommender Systems: The Textbook, 2016, http://pzs.dstu.dp.ua/DataMining/recom/bibl/1aggarwal_c_c_recommender_systems_the_textbook.pdf (Last accessed: 10.10.2023).

2. *Francesco Ricci, Lior Rokach, Bracha Shapira.* Recommender Systems Handbook, 2015, https://www.cse.iitk.ac.in/users/nsrivast/HCC/Recommender_systems_handbook.pdf (Last accessed: 10.10.2023).

3. *M.Sridevi, Dr .R.Rajeswara Rao*, DECORS: A Simple and Efficient Demographic Collaborative Recommender System for Movie Recommendation, 2017, ISSN 0973-6107 Volume 10, Number 7, pp. 1969-1979. https://www.ripublication.com/acst17/acstv10n7_01.pdf (Last accessed: 20.10.2023).

4. *Ruogu Kang, Stephanie BrownSara, KieslerSara Kiesler*, Why do people seek anonymity on the Internet? Informing policy and design, 2013 https://www.researchgate.net/publication/262273589_Why_do_people_seek_anonymity_on_the_Internet_Informing_policy_and_design (Last accessed: 20.10.2023).

5. *Abdellah El Fazziki, Ouafae El Aissaoui, Yasser EL Madani El Alami, Youssouf El Allioui*, Mohammed Benbrahim, A new collaborative approach to solve the gray-sheep users problem in recommender systems, 2019, https://www.researchgate.net/publication/338361794_A_new_collaborative_approach_to_solve_the_gray-sheep_users_problem_in_recommender_systems (Last accessed: 20.10.2023).

6. *Juuso Kaitila*, A content-based music recommender system, 2017, https://www.cs.rit.edu/usr/local/pub/GraduateProjects/2161/kxd8041/Report.pdf (Last accessed: 20.10.2023).

7. *Guan, Xin*, On reducing the data sparsity in collaborative filtering recommender systems., 2017, https://wrap.warwick.ac.uk/97978/1/WRAP_Theses_Guan_2017.pdf (Last accessed: 20.10.2023).

8. *Wanvimol Nadee*, Modelling user profiles for recommender systems, 2016, https://eprints.qut.edu.au/93723/1/Wanvimol_Nadee_Thesis.pdf (Last accessed: 20.10.2023).

9. *Andreu Vall, Matthias Dorfer, Hamid Eghbal-zadeh, Markus Schedl, Keki Burjorjee & Gerhard Widmer.* Feature-combination hybrid recommender systems for automated music playlist continuation, 2019, https://link.springer.com/article/10.1007/s11257-018-9215-8 (Last accessed: 10.10.2023).

10. *Sebastien Fremal, Fabian Lecron*. Weighting Strategies for a Recommender System Using Item Clustering Based on Genres, 2017, p. 6-11, https://www.sciencedirect.com/science/article/abs/pii/S0957417417300404 (Last accessed: 10.10.2023).

Надійшла до редколегії 13.10.2023 р

Ситнікова Поліна Едуардівна, канд. техн. наук, доцент, доцент кафедри системотехніки ХНУРЕ, м. Харків, Україна, e-mail: polina.sytnikova@nure.ua, ORCID: https://orcid.org/0000-0002-6688-4641.
Гребенюк Микита Олександрович, аспірант кафедри системотехніки ХНУРЕ, м Харків, Україна, e-mail: mykyta.hrebeniuk@nure.ua. ORCID: https://orcid.org/0009-0008-0989-7957.

УДК 004.627 DOI: 10.20837/0135-1710.2023.179.042

І.Г. ПЕРОВА, Н.С. МІРОШНИЧЕНКО

ОГЛЯД ІСНУЮЧИХ МЕТОДІВ ЗМЕНШЕННЯ РОЗМІРНОСТІ ТА КЛАСИФІКАЦІЇ ВЕЛИКИХ ВИБІРОК ДАНИХ

Аналіз великих вибірок даних, який проводиться з метою виявлення прихованих закономірностей і тенденцій, за останні роки стає все важливішим і кориснішим. Такі великі вибірки на поточний час характеризуються загальнодоступністю, складністю структур і великими розмірами.

Для вирішення проблеми великої розмірності даних пропонується ознайомлення з існуючими методами зменшення розмірності великих вибірок даних та порівняння ефективності цих методів на репозиторних вибірках. Розглядаються такі методи, як аналіз головних компонент (Principal Component Analysis), лінійний дискримінантний аналіз (Linear Discriminant Analysis), аналіз головних компонент ядра (Kernel Principal Component Analysis), багатовимірне масштабування (MDS), метод t-розподільного стохастичного вбудовування сусідів (t-SNE) та аналіз незалежних компонент (Independent Component Analysis). Як приклади великих вибірок даних використовуються набір даних ініціативи з нейровізуалізації хвороби Альцгеймера (ADNI) та набір даних про щитоподібну залозу, який є одним з декількох баз даних про щитоподібну залозу, доступних в репозиторії UCI.

1. Вступ

Для роботи з великими вибірками даних попередньо необхідно зменшити кількість параметрів вибірки. Цей процес називається зменшенням розмірності вибірки.

Зменшення розмірності як етап попередньої обробки машинного навчання є ефективним для видалення нерелевантних і надлишкових даних, підвищення точності навчання та покращення зрозумілості результату за допомогою візуалізації розмірності [1]. Дуже важливо зменшити розмірність набору даних без втрати будь-якої інформації, що міститься в них.