http://www.moderntechno.de/index.php/meit/article/view/meit34-00-013
DOI: 10.30890/2567-5273.2024-34-00-013

УДК 519.816

COLLABORATIVE FILTERING AS ONE OF THE MAIN METHODS OF CONTENT RECOMMENDATIONS: MAIN PROBLEMS AND WAYS TO SOLVE

КОЛАБОРАТИВНА ФІЛЬТРАЦІЯ ЯК ОДИН ІЗ ОСНОВНИХ МЕТОДІВ РЕКОМЕНДАЦІЙ КОНТЕНТУ: ОСНОВНІ ПРОБЛЕМИ ТА СПОСОБИ ЇХ ВИРІШЕННЯ

Yevhen Ivokhin/Євген Івохін
DSc (Phys. & Math.), Prof/д-р фіз.-мат. наук, проф.
ORCID ID: 0000-0002-5826-7408
Glib Shelyakin/Гліб Шелякін
post-graduate/ аспірант
ORCID ID: 0009-0002-7171-6535
Taras Shevchenko National University of Kyiv, 60 Volodymyrska Street, Kyiv, 01033
Київський національний університет імені Тараса Шевченка,
Київ, вул. Володимирська, 60,01033

Анотація. *У статті розглядається колаборативна фільтрація як один із основних методів формування рекомендацій контенту. Проаналізовано основні проблеми методу, такі як синонімія об'єктів, «холодний старт», розрідженість даних та різноманітність. Розглянуто способи вирішення зазначених проблем шляхом застосування часового фактору, семантичної подібності, кластеризації та статистичних методів. Запропоновано власні шляхи удосконалення методу колаборативної фільтрації, методу кластерного аналізу HDBSCAN, сіамських нейронних мереж та семантичної подібності.*

Ключові слова: *колаборативна фільтрація, «синонімія» об'єктів, «холодний старт», розрідженість даних, кластерний аналіз.*

Вступ.

Колаборативна фільтрація – це підхід до створення рекомендацій контенту, що ґрунтується на реакціях користувачів на цей контент. Метод колаборативної фільтрації вперше запропонував Голдберг у 1992 році як систему для фільтрації електронної пошти. Головна мета цього методу полягає в тому, щоб на основі попередніх реакцій користувача передбачити, як він оцінить контент, з яким ще не взаємодіяв. Чим точніший розрахунок, тим якісніше рекомендація.

Поняття *колаборативна фільтрація* має два основні значення: вузьке та широке. У більш широкому сенсі – це процес відбору інформації або зразків через співпрацю кількох джерел, поглядів чи агентів. Цей метод часто застосовується до великих наборів даних, таких як результати зондування та моніторингу у геології, фінансові дані від різних фінансових установ, або інформація користувачів в електронній торгівлі та веб-додатках.

У вузькому значенні, колаборативна фільтрація – це метод побудови прогнозів у рекомендаційних системах, який використовує відомі оцінки групи користувачів для прогнозування невідомих уподобань іншого користувача. Основне припущення цього методу – користувачі, які однаково оцінювали певні об'єкти в минулому, ймовірно, надаватимуть схожі оцінки й у майбутньому. Наприклад, музичний додаток, використовуючи колаборативну

фільтрацію, може передбачити, яка музика сподобається користувачеві, навіть, за умови неповного переліку його уподобань. Прогнози створюються індивідуально для кожного користувача, використовуючи зібрану інформацію від багатьох інших. Це відрізняється від підходу, який дає середню оцінку об'єктів на основі кількості голосів. У цій галузі наразі проводяться активні дослідження, оскільки метод має певні нерозв'язані проблеми.

Серед існуючих методів рекомендацій найяскравішими вважають такі: метод фільтрації на основі вмісту, метод колаборативної фільтрації на основі інформації про користувача (user-based) та на основі порівнянь об’єктів рекомендацій (item-based).

Метою статті є проаналізувати основні проблеми методу колаборативної фільтрації та способи їх вирішення.

Виклад основного матеріалу. Серед проблем методу колаборативної фільтрації слід відзначити «холодний старт», розрідженість даних, забезпечення синонімії об'єктів та різноманітність.

Найбільшою проблемою в застосуванні методу колаборативної фільтрації є проблема «холодного старту». Проблема «холодного старту» для user-based методу виникає тоді, коли у користувача, який щойно зареєструвався на сервісі, відсутня історія переглядів. В результаті система діє недетерміновано замість того, щоб рекомендувати те, що може потенційно зацікавити користувача. Проблема «холодного старту» для методу колаборативної фільтрації, заснованому на item-based підході, тобто на обчисленні оцінок на основі порівнянь об’єктів, виникає у тому випадку, коли контенту не було надано оцінку жодним користувачем, через те, що система просто «не звертає на нього увагу» до того моменту, поки хоча б один користувач не оцінить цей контент.

Проблема «холодного старту» залишається однією з найактуальніших проблем методів колаборативної фільтрації, оскільки при створенні нових сервісів із рекомендаціями на основі колаборативної фільтрації дані про користувачів та їх взаємодії із сервісом відсутні. Вирішенню даного питання присвячено низьку спеціальних досліджень. Зокрема, Чарльз Емануєль Діаз, Вінсент Гуі та Патрік Галінарі [3] вирішували проблему «холодного старту» завдяки впровадженню оцінок на основі змісту текстової інформації. Для цього вони використовують метод *word2vec*, щоб перетворити текстову інформацію на набір векторів і завдяки ним навчити систему «передбачати» найближчі по контексту вектори, та на їх основі проводити рекомендації, застосовуючи стандартний метод колаборативної фільтрації. В якості метрики подібності автори використали подібність косинусів такого виду:

$$sim(u,v) = \alpha_{uv} = \frac{\frac{\langle u,v \rangle}{\|u\| \times \|v\|} + 1}{2}$$

В результаті були виведені формули для розрахунку рейтингів як для користувача, так і для об’єктів:

$$\hat{r_{ui}} = \mu_i + \frac{\sum_{j \in N_i} \alpha_{ij}(r_{uj} - \mu_j)}{\sum_{j \in N_i} \alpha_{ij}}, \hat{r_{vi}} = \mu_u + \frac{\sum_{v \in N_u} \alpha_{uv}(r_{vi} - \mu_v)}{\sum_{v \in N_u} \alpha_{uv}}$$

Не зважаючи на використання набору даних, для якого матриця оцінок була сильно розрядженою, вдалося зменшити похибки при апроксимації оцінки користувачів на 10% - 15%, перевершивши такі існуючі моделі як наївний метод колаборативної фільтрації на основі порівнянь користувачів та матричну факторизацію [3].

Кін Лу [11] запропонував новий спосіб підбору користувачів для user-based методу шляхом розділення ознак (атрибутів) користувача на лінійні та ієрархічні. На думку автора прикладом лінійних ознак можуть бути заробітна плата або вік, оскільки їх можна умовно поділити на відповідні відрізки – до 18 років, від 18 до 30, від 30 років і так далі. Тоді такий параметр як адреса проживання можна відобразити у вигляді ієрархічної структури – дерева. Автор також запропонував формули для розрахунку подібності лінійних атрибутів:

$$sim(u,v) = \frac{1}{1+|u-v|}$$

та ієрархічних атрибутів:

$$sim(u,v) = 1 - \frac{L(u,v)}{H}$$

H – висота ієрархічного дерева, а $L(u,v)$ - найдовший шлях від атрибуту до загального вузла.

Крім цього, Кін Ліу запропонував нову кількісну оцінку, яку назвав *User Attention*, для розрахунку схожості користувачів за їх увагою та формулу до неї:

$$sim(u,v) = \frac{|Iu,v|}{|Iu|+|Iv|+|Iu,v|}$$

$|Iu,v|$ — кількість спільно оцінених об’єктів користувачами u і v; а $|Iu|,|Iv|$ — кількість оцінених об’єктів користувачами u і v відповідно.

Був розроблений індекс *Project Popularity*, який має на меті «прирівняти» товари, що є занадто популярними до тих, які не є такими:

$$norIP_i = \frac{IP_i - \text{minIP}}{maxIP - minIP}$$

$$w_i = w_0 * \frac{norIP_i}{\sum_{i=1}^{N} norIP_i},$$

де $IP_i -$ популярність пункту i, яка може бути виражена частотою елемента i у сукупності даних.

В результаті практичного випробування методу показано, що алгоритм з авторськими модифікаціями працює краще, ніж звичайний метод колаборативної фільтрації, але алгоритм є чутливим до розрідженості матриці оцінок, що виражається у дуже незначному зменшенні похибки у порівнянні зі звичайним алгоритмом.

Етієн Брангбур, Пієрік Бруню, Томас Тамісієр та Стефан Мачанд [1] пропонують використовувати алгоритми кластеризації для розв’язання

проблеми «холодного старту» у випадку незбалансованого набору даних (тобто такого, де даних одного класу значно більше, ніж іншого). Запропоновані ними алгоритми використовують імовірнісний підхід з використанням *Індексу якості кластеру*:

$$\Phi_k = 1 - \frac{\sum_{i \in I_k, j \notin I_k} w_{ij}}{\min(a_k, a_{\bar{k}})},$$

$$a_k = \sum_{i \in I_k, j \in |1,N|} w_{ij},\ a_{\bar{k}} = \sum_{i \notin I_k, j \in |1,N|} w_{ij},$$

I_k – позначає множину членів (або індексів), які належать до кластеру k

Це дає змогу оцінити належність кожного об'єкту до відповідних кластерів, а також дозволяє більш точно оцінити вплив того чи іншого об'єкту на результуючу оцінку. Запропонований підхід дозволяє також зменшити обчислювальні ресурси, оскільки замість того, щоб штучно збільшувати або зменшувати кількість об'єктів у вибірці, вони можуть використовувати розраховані коефіцієнти належності до кластерів.

Для вирішення проблеми «холодного старту» нами була використана додаткова матриця, яка зберігає інформацію про подібність об'єктів завдяки використанню семантичного фактору [16].

Нехай M – множина текстової інформації, що відображає суть деякого об'єкту. Кожен елемент множини є вектор $(a_1, a_2, \dots, a_k)$ – де a_i – це деяке текстове представлення i-го атрибуту поточного об'єкту. Це можуть бути описи, рецензії, набір технічних характеристик, відгуки.

Тоді величину подібності (близькості) двох об'єктів можна обчислити за такою формулою:

$$sim_{meta}(i, \hat{i}) = \frac{\sum_{j=1}^{k} sim(i_j, \hat{i_j})}{k}$$

де $sim(i_j, \hat{i_j})$ - коефіцієнт подібності j-х характеристик об'єктів i та $\hat{i}$ із застосуванням сіамських нейронних мереж у якості оцінювачів.

Введемо матрицю $OS\ (Object\ Similarity) \in R^{n,n}$, де n – кількість об'єктів на сервісі. Матриця є квадратною та симетричною, $OS_{i,j}\ i > j = \overline{1, n-1}$ – оцінки подібності i-го та j-го об'єкту.

Цей підхід дозволяє здійснювати рекомендації навіть при мінімальній кількості даних про користувача або об'єкт. Застосування матриці OS для вирішення проблеми «холодного старту» показало значне покращення якості рекомендацій для нових користувачів та об'єктів. Навіть при мінімальній кількості доступних даних система змогла надавати релевантні рекомендації. Таким чином, запропонований підхід довів свою ефективність і може бути використаний як важливий інструмент для покращення систем колаборативної фільтрації.

Наступною проблемою є розрідженність даних. Розрідженість даних зустрічається у ситуаціях, коли значна частина даних у датасеті відсутня або містить неточні значення. Це часто трапляється у великих наборах даних, де багато можливих взаємодій між об'єктами не реєструються або не відомі. Наприклад, у контексті рекомендаційних систем, більшість користувачів взаємодіє лише з дуже малою частиною доступного контенту, що залишає більшість користувацьких оцінок невідомими. В результаті матриці, що визначають відношення «користувач-предмет», є сильно розрідженими.

Дослідники намагалися вирішити проблему розрідженості завдяки застосуванню методів кластеризації. Наприклад, Лілі Юнгар та Дін Фостер [12] пропонують використати метод кластеризації K-means. Ідея авторів полягає у наступному: згрупувати людей на основі переліку фільмів, які вони дивилися, а потім згрупувати фільми на основі людей, які їх дивилися. Потім людей можна повторно кластеризувати на основі кількості фільмів у кожному кластері фільмів, які вони переглянули. Сукупність фільмів можна так само повторно кластеризувати на основі кількості людей у кожному кластері осіб, які їх дивилися. У своєму дослідженні автори провели зазначену процедуру декілька разів, для того щоб коректно розбити відповідні множини на кластери. В той же час автори зазначають, що таке розбиття потрібно повторювати кожен раз при додаванні нового користувача чи фільму у базу даних. В результаті додавання такої модифікації вдалося зменшити розрідженість матриці оцінок та прискорити швидкість обрахунків.

Ївеі Сяо та Ральф Кламма [13] пропонують використовувати метод агломеративної кластеризації для розбиття користувачів на відповідні групи, використовуючи оцінки, які вони поставили фільмам. Після розбиття користувачів на кластери автори застосовують метод рекомендації типу «сусідство» (neighbourhood method). Він полягає у тому, щоб знайти користувачів зі схожими оцінками. Оскільки в кластерах знаходяться користувачі, які є «подібними» один до одного, це дозволило більш точно рекомендувати фільми користувачам, зменшивши похибку в середньому на 15%.

Марк О'Коннор та Джон Херлокер [10] порівнювали 3 метода кластеризації – метод агломеративної кластеризації, метод ROCK та kMetis для розбиття групи фільмів за оцінками, які їм поставили користувачі. Їх гіпотеза полягала у тому, щоб підвищити точність передбачення, об'єднавши фільми з однаковими рейтингами. Однак у експериментах вони не виявили, що це відбувається постійно, а результати показали зворотнє – похибка не зменшилась, а навпаки, збільшилась на 5%. Лише в одному чи двох випадках точність для елементів у розділі була більшою, ніж у базовому нерозділеному випадку. Автори пояснюють це тим, що кореляція рейтингів між двома елементами вимірює, наскільки подібно два елементи оцінені, а не обов'язково, наскільки ці два елементи схожі за змістом. Це також може бути пов'язано з тим, що вони обмежують елементи виключно в одному кластері. Певні елементи можуть мати значну прогностичну цінність для кількох кластерів, і їх видалення може зменшити загальний показник точності. Хоча автори

зазначають позитивні моменти свого дослідження, наприклад, швидкість обрахунку збільшилась, оскільки не потрібно обраховувати великі набори даних.

Для вирішення проблеми розрідженості даних нами було запропоновано використання методу кластерного аналізу HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) [16]. HDBSCAN є покращеним варіантом алгоритму DBSCAN і дозволяє виявляти кластери різної щільності та ідентифікувати шум у даних. Цей метод добре підходить для роботи з семантичними відстанями, обчисленими з матриці OS, яка будується на основі семантичного аналізу об'єктів, що включає обробку текстових описів, категорій та інших метаданих. Це дозволяє виявляти подібність між об'єктами на основі їх змісту, навіть якщо для них немає користувацьких оцінок. Таким чином, при перегляді хоча б одного об'єкту користувачем, система може запропонувати релевантні рекомендації, використовуючи інформацію про семантичну подібність інших об'єктів. Застосування методу HDBSCAN для кластерного аналізу об'єктів на основі семантичних відстаней показало високу ефективність у виявленні природних груп об'єктів. Система змогла більш точно визначати схожі об'єкти та надавати користувачам більш релевантний контент. Додатково, метод HDBSCAN дозволив ідентифікувати шум у даних, що також сприяло підвищенню загальної точності рекомендацій.

Ще одним змістовним недоліком методу колаборативної фільтрації є так звана «синонімія» об’єктів, за якою однакові за своєю суттю об’єкти мають, наприклад, різні теги, назви, описи, тощо. Більшість рекомендаційних систем не здатні виявити ці приховані зв'язки і тому розпізнають ці предмети як різні. Наприклад, «фільми для дорослих» та «дорослі фільми» належать до одного жанру, але система сприймає їх як різні.

Автори Скотт Дивестер, Томас Ландауер, Річард Хашман [2] у своєму дослідженні намагаються подолати недоліки пошуку за збігом термінів, розглядаючи ненадійність асоціацій *термін-документ* як статистичну проблему. Припускається, що існує прихована семантична структура в даних, частково затемнена випадковістю вибору слів. Вони використовують статистичні методи для оцінки цієї структури та видалення «шуму». Опис термінів і «документів» на основі цієї структури використовується для індексації та пошуку. Алгоритм використовує сингулярний розклад, що створює «семантичний» простір, де асоційовані терміни та документи розміщуються близько один до одного. Це відображає основні асоціативні патерни в даних, ігноруючи менш важливі впливи, призводячи до того, що навіть якщо певний термін не використовується у конкретному документі, він все одно може виявитися близьким за значенням до цього документа у семантичному просторі, створеному сингулярним розкладом. Позиція в цьому просторі визначає новий тип семантичної індексації, а пошук позиції здійснюється через ідентифікацію точки в просторі за термінами запиту, після чого користувачу повертаються документи з найближчого сусідства. Цей семантичний простір і використовується для пошуку подібних «документів».

Айман С. Габаєн та Шахрул Азман [4] пропонують рахувати семантичну

подібність між тегами за допомогою відомої лексичної бази даних WordNet для англійської мови. WordNet є великою концептуальною базою даних, що містить іменники, дієслова, прикметники та прислівники, згруповані в набори когнітивних синонімів (синсети). У WordNet існують концептуально-семантичні зв'язки та лексичні відносини між синсетами. Терміни з однаковим значенням називаються синонімами і належать до одного синсету. Ці ієрархічні концепти можуть кількісно визначити, наскільки концепт «A» подібний до концепту «B». Наприклад, автомобіль більш подібний до човна, ніж до дерева, оскільки «човен» і «автомобіль» мають спільного предка «транспортний засіб» у структурі WordNet. Даний підхід базується на формалізації у вигляді трійок <користувач, елемент, тег>, що широко використовується у спільнотах спільного тегування. Кожна трійка представляє анотацію користувача до елемента з тегом. Для семантичного обґрунтування вони використовують WordNet як зовнішній семантичний простір для вимірювання семантичної подібності між тегами. Вимірювання семантичної подібності в WordNet можна здійснити, вимірюючи відстань між вузлами, пов'язаними з відповідними концептами. Зв'язки між цими вузлами оцінюються з точки зору відстані, яка вказує на те, наскільки подібні концепти. Вони вимірюють подібність між тегами, використовуючи семантичну подібність Ліна, яка визначається на основі величини показника *Інформаційного спільного предка* для розрахунків:

$$sim_{Lin}(c_1, c_2) = \frac{2 \times IC(C_{MICA})}{IC(c_1) + IC(c_2)},$$

$IC(c_1), IC(c_2)$ – значення інформаційного вмісту концепцій (тегів) c_1, c_2, а C_{MICA} – найбільш інформативний спільний предок. В результаті для розрахунку подібності автори визначили формулу:

$$SUsim(u, v) = \sum_{m \ni m_{u,v}} \sum_{t_u, t_v} sim_{Lin}(t_u, t_v),$$

$m_{u,v}$ – набір об'єктів, які спільно позначені тегами обома користувачами u та v, t_u – це теги, які користувач u поставив об'єкту m, t_v – це теги, які користувач v поставив об'єкту m.

Лян Хуйчжі, Сюй Юе, Лі Юефен і Наяк Річі [6] пропонують вирішення проблеми таким чином: нехай у кожного користувача є деякий набір тегів, які йому найбільш до вподоби та нехай є деякий набір популярних тегів на сервісі. Вони розраховують вагу:

$$w(c_x, t, u) = \frac{1}{|P(t,u)|} \sum_{p_j \in P(t,u)} f(p_j, c_x),$$

$P(t, u)$ – набір елементів, які віднесені до тегу t користувачем u; c_x – деякий тег з множини популярних тегів C; $f(p_j, c_x) \in [0, 1]$ – міра приналежності тегу c_x до об'єкта p_j, тобто наскільки часто c_x зустрічається у об'єкті p_j.

Оскільки кількість елементів у різних тегах може бути різною, вони

нормалізують вагу за кількістю елементів у тегі t користувача u. Таким чином, зберігається точка зору кожного користувача на класифікацію його/її елементів, водночас отримуючи набір популярних тегів для представлення семантичного значення кожного тега. Для різних користувачів представлення того самого тега можуть бути різними, а також представлення різних тегів можуть бути однаковими або схожими. Незважаючи на те, що терміни тегів вільно обираються окремими користувачами, представлення кожного тега за допомогою набору популярних тегів робить усі теги порівнянними, оскільки всі вони представлені одним і тим самим набором термінів. Використовуючи популярні теги для представлення, ті непопулярні теги, які часто викликають плутанину та шум, стають зрозумілими для інших користувачів через їх відповідне представлення популярними тегами. Подання популярних часто виявляє інші пов'язані популярні теги, які самі по собі мають велику вагу у своїх представленнях. Оскільки кожен тег представлений набором популярних тегів, що забезпечує основу для порівняння, цей підхід допомагає вирішити проблеми, викликані вільним стилем написання тегів, такі як синоніми тегів (різні теги з однаковим значенням), семантична неоднозначність тегів (один тег має різні значення для різних користувачів) і варіації написання тощо.

Для вирішення проблеми «синонімії» об'єктів рекомендацій нами було запропоновано використовувати сіамські нейронні мережі [16]. Сіамська нейронна мережа складається з двох або більше підмереж, що мають однакову архітектуру та спільні ваги. Вона дозволяє порівнювати об'єкти та виявляти їх подібність на основі векторних представлень. Застосування сіамських нейронних мереж для вирішення проблеми «синонімії» об'єктів рекомендацій показало значне покращення якості рекомендацій. Попередньо навчена модель змогла адекватно ідентифікувати схожі об'єкти, що дозволило уникнути дублювання рекомендацій та забезпечити користувачам більш різноманітний та релевантний контент.

І, нарешті, останній суттєвий недолік, на якому треба зупинитись – це різноманітність. Колаборативна фільтрація спочатку збільшує різноманітність, щоб дозволяти відкривати користувачам нові продукти з незліченної множини. Однак деякі алгоритми створюють дуже складні умови для просування нових і маловідомих продуктів, оскільки перевага надається популярним продуктам, які давно перебувають на ринку.

Це питання також знайшло відображення у дослідженнях фахівців. Зокрема, Ї Дін, Сюе Лі [14] використовували монотонно спадну функцію $f(t) = e^{-\alpha t}$ для впливу в методі *Nearest neighbourhood* на етапі створення рекомендації, надаючи більшої ваги елементам, що оцінювалися користувачами недавно. Тут t — це час, коли було надано оцінку, а α — параметр, який контролює швидкість спадання функції за часом.

Ї Дін, Сюе Лі та Марія Орловська [15] використовували зважені оцінки за останніми рейтингами, наданими подібним елементам, для прогнозування рейтингу елемента.

Також Луй Натаннан, Чжао Мін, Сян Еванвей, Ян Цян [8] представили функції старіння даних у часі як при обчисленні подібності, так і в кроках прогнозування рейтингу алгоритму на основі елементів. Для кроку обчислення подібності використовується спадна експоненціальна функція $f(t) = e^{-\alpha t}$ з параметром α. На практиці це призводить до того, що пари елементів стають все менш і менш схожими, оскільки їхні оцінки стають менш побідними зчасом. На кроці прогнозування рейтингу подібна функція старіння $g(t) = e^{-\beta t}$ використовується для прогнозування рейтингу, використовуючи той самий метод, що й у Дінг Ю. та Лі Х. Єдина відмінність між функціями f і g — параметри старіння α і β, які насправді можуть бути однаковими. Автори стверджували, що використання різних коефіцієнтів старіння для обчислення подібності та прогнозування рейтингу забезпечує точніший контроль над алгоритмом. Однією з суттєвих особливостей цього алгоритму є те, що обчислення подібності є поетапним і має невисоку складність, що дозволяє алгоритму виконувати он-лайн оновлення, оскільки користувачі постійно дають оцінки.

Олфа Насрауї, Джефф Сервінске, Карлос Рохас і Фабіо Гонсалес [9] запропонували так званий «алгоритм сусідства», який використовує ковзаюче вікно, що містить фіксовану кількість екземплярів. Алгоритм обчислює подібність між останніми сеансами користувача. Кожна сесія користувача характеризується конкретною кількістю оцінок, наданих користувачем за заданий невеликий проміжок часу – наприклад, протягом 1 години.

Інший підхід із використанням часових інтервалів дослідили Ніл Лафія, Лісія Капра, Стівен Хейлз [5]. Автори використовували набір алгоритмів на основі елементів, що відрізняються лише кількістю найближчих сусідів, які розглядаються для прогнозування рейтингів. Алгоритми перенавчаються через фіксовані проміжки часу – 7 днів, із змінною кількістю екземплярів – з даними за той самий проміжок часу. Помилка постійно відстежується для всіх алгоритмів, і для надання рекомендацій вибирається алгоритм із найменшою помилкою.

Для вирішення цієї проблеми автори статті запропонували використовувати часову функцію $f(t) = e^{-t}$, яка зменшує вагу більш старіших переглядів у процесі саме розрахунку оцінки. Це дозволяє системі надавати пріоритет новішим переглядам, що відображає актуальні інтереси користувачів. Як результат, нові перегляди мають більший вплив на рекомендації, що дозволяє системі адаптуватися до змін у перевагах користувачів. Це сприяє підвищенню задоволеності користувачів та ефективності системи рекомендацій.

Результати та виновки. Отже, проаналізувавши основні недоліки застосування методу колаборативної фільтрації, проаналізовано підходи та запропоновано спроби для подолання зазначених проблем. В результаті аналізу зосереджено увагу на принципових модифікаціях методики.

Для вирішення проблеми «холодного старту» використано додаткову матрицю OS, яка зберігає інформацію про подібність об'єктів завдяки використанню семантичного фактору. Цей підхід дозволяє здійснювати рекомендації навіть при мінімальній кількості даних про користувача або об'єкт. Застосування такої матриці дозволило значно покращити якість рекомендацій для нових користувачів та об'єктів. Навіть за умови мінімальної кількості доступних даних система змогла надавати релевантні рекомендації. Таким чином, запропонований підхід довів свою ефективність і може бути використаний як важливий інструмент для покращення систем колаборативної фільтрації.

Для вирішення проблеми розрідженності даних запропоновано використання методу кластерного аналізу HDBSCAN. Цей метод добре підходить для роботи з семантичними відстанями, обчисленими з матриці OS, яка будується на основі семантичного аналізу об'єктів, включаючи обробку текстових описів, категорій та інших метаданих. Це дозволяє виявляти подібність між об'єктами на основі їх змісту, навіть якщо для них немає користувацьких оцінок.

Застосування методу HDBSCAN для кластерного аналізу об'єктів на основі семантичних відстаней показало високу ефективність у виявленні природних груп об'єктів. Система змогла більш точно визначати подібні об'єкти та надавати користувачам більш релевантний контент. Додатково, метод HDBSCAN дозволив ідентифікувати шум у даних, що також сприяло підвищенню загальної точності рекомендацій.

Для вирішення проблеми «синонімії» об'єктів рекомендацій запропоновано використання сіамських нейронних мереж. Вони дозволяють порівнювати об'єкти та виявляти їх подібність на основі векторних представлень. Застосування сіамських нейронних мереж також дозволило покращити якість рекомендацій. Запропоновані модифікації покращили можливості методу адекватно ідентифікувати схожі об'єкти, що дозволило уникнути дублювання рекомендацій та забезпечити користувачам більш різноманітний та релевантний контент.

Для актуалізації рекомендацій запропоновано використання часову функцію старіння, яка зменшує вагу старіших переглядів. Це дозволяє системі надавати пріоритет новішим переглядам, що відображає актуальні інтереси користувачів. Як результат, нові перегляди мають більший вплив на рекомендації, що, в свою чергу, дозволяє системі адаптуватися до змін у перевагах користувачів. Це сприяє підвищенню задоволеності користувачів та ефективності системи рекомендацій.

Отримані результати, за якими спостерігалось зменшення середньо-квадратичної похибки на 20% та зменшення часу виконання в середньому на 40%, підтверджують конструктивність запропонованих модифікацій, завдяки яким можна покращити швидкість та якість функціонування рекомендаційних систем у різних проблемних областях.

Розглянута у статті тематика досліджень залишається актуальною, потрубує подальшого поглиблення за змістом та поширення за сферами

використання, впровадження результатів дослідження може бути продовжене з урахуванням особливостей відповідних галузей, серед яких інтернет речей, фінанси та медіа.

Література:

1. Brangbour Etienne, Bruneau Pierrick, Tamisier Thomas, Marchand-Maillet Stephane. Active Learning with Crowdsourcing for the Cold Start of Imbalanced Classifiers. Cooperative Design, Visualization, and Engineering. 2020. P .192–201. URL: https://www.researchgate.net/publication/346246697_Active_Learning_with_Crowdsourcing_for_the_Cold_Start_of_Imbalanced_Classifiers

2. Deerwester Scott, Dumaism Susan T., Furnas George W., Landauer Thomas K., Harshman Richard. Indexing by Latent Semantic Analysis. URL: http://wordvec.colorado.edu/papers/Deerwester_1990.pdf

3. Dias Charles-Emmanuel, Guigue Vincent, Gallinari Patrick. Text-based collaborative filtering for coldstart soothing and recommendation enrichment. AISR2017, May 2017, Paris, France. URL: https://hal.science/hal-01640268/document**.**

4. Ghabayen Ayman S., Noah Shahrul Azman. Using Tags for Measuring the Semantic Similarity of Users to Enhance Collaborative Filtering Recommender Systems. International Journal on Advanced Science Engineering and Information Technology. 2017. P. 2063-2070. URL: https://www.researchgate.net/profile/Ayman-Ghabayen-2/publication/320258585_Using_Tags_for_Measuring_the_Semantic_Similarity_of_Users_to_Enhance_Collaborative_Filtering_Recommender_Systems/links/5a423075458515f6b04dd899/Using-Tags-for-Measuring-the-Semantic-Similarity-of-Users-to-Enhance-Collaborative-Filtering-Recommender-Systems.pdf .

5. Lafia Neal, Capra Licia, Hailes Stephen. Temporal Collaborative Filtering With Adaptive Neighbourhoods. Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR 2009. Boston. USA. 2009. P.796-797. URL: https://doi.org/10.1145/1571941.1572133.

6. Liang, Huizhi and Xu, Yue and Li, Yuefeng and Nayak, Richi (2009) Collaborative Filtering Recommender Systems based on Popular Tags. Proceedings of the Fourteenth Australasian Document Computing Symposium, 4 December 2009, University of New South Wales, Sydney. P. 1-8. URL: https://eprints.qut.edu.au/29732/1/29732.pdf .

7. Liang, Huizhi, Xu, Yue, Li, Yuefeng, & Nayak, Richi. Collaborative filtering recommender systems based on popular tags. Proceedings of the 14th Australasian Document Computing Symposium. University of Sydney. Australia. 2009. P. 1-8. URL: https://eprints.qut.edu.au/29732/1/29732.pdf .

8. Lui Nathannan, Zhao Min, Xiang Evanwei, Yang Qiang. Online evolutionary collaborative filtering. Proceedings of the fourth ACM conference on Recommender systems. 2010. P. 95-102. URL: https://doi.org/10.1145/1864708.1864729.

9. Nasraoui Olfa, Cerwinske Jeff, Rojas Carlos, and Gonzalez Fabio.

Performance of Recommendation Systems in Dynamic Streaming Environments. Proceedings of the 2007 SIAM International Conference on Data Mining**.** P. 569-574. URL: https://doi.org/10.1137/1.9781611972771.63

10. O'Connor Mark, Herlocker Jon Clustering Items for Collaborative Filtering. https://redirect.cs.umbc.edu/~ian/sigir99-rec/papers/oconner_m.pdf

11. Qin Lui A New Collaborative Filtering Algorithm Integrating Time and Multisimilarity. Mathematical Problems in Engineering. 2022. URL: https://www.hindawi.com/journals/mpe/2022/2340671/

12. Ungar Lyle H. and Foster. Dean P. Clustering Methods for Collaborative Filtering Technical Report. 1998. P. 114–128. URL: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=d6b99637332f818dc94c0432db617db28873e035**.**

13. Yiwei Cao, Klamma Ralf. Clustering Technique for Collaborative Filtering and the Application to Venue Recommendation. Proceeding of the 10th International Conference on Knowledge Management and Knowledge Technologies (I-KNOW 2010), 1-3 September, 2010, Graz, Austria URL: https://www.researchgate.net/publication/228446479_Clustering_Technique_for_Collaborative_Filtering_and_the_Application_to_Venue_Recommendation

14. Yi Ding, Xue Li. Time Weight Collaborative Filtering. URL: https://cseweb.ucsd.edu/classes/fa17/cse291-b/reading/p485-ding.pdf

15. Yi Ding, Xue Li, Maria Orlowska.Recency-based collaborative filtering. Proceedings of the 17th Australasian Database Conference. 2006. Hobart. Tasmania. Australia. January 16–19. 2006. URL: https://www.researchgate.net/publication/221152610_Recency-based_collaborative_filtering.

16. Івохін Є., Шелякін Г., Махно М. Удосконалення методу колаборативної фільтрації шляхом інтегрування семантичного та часового факторів і методу кластерного аналізу. Artificial Intelligence. №1, 2024. С. 57 – 63. URL: https://jai.in.ua/archive/2024/2024-1-5.pdf

Abstract. *The article considers collaborative filtering as one of the main methods of content recommendation. The main problems of the method are analyzed, such as "synonymy of objects", "cold start", sparsity of data, the ways to solve these problems by using the time factor, semantic similarity, clustering and statistical methods are considered. Own ways of improving the collaborative filtering method, the HDBSCAN cluster analysis method, Siamese neural networks and semantic similarity are proposed.*

Keywords: *collaborative filtering, "synonymy" of objects "cold" start, data sparsity, cluster analysis.*

Науковий керівник: доктор фізико-математичних наук, професор ЄВГЕН ІВОХІН

Стаття відправлена: 04.08.2024 13:33