

ЗАСТОСУВАННЯ КЛАСТЕРНОГО АНАЛІЗУ ДЛЯ ВДОСКОНАЛЕННЯ АЛГОРИТМУ КОЛАБОРАТИВНОЇ ФІЛЬТРАЦІЇ

Анотація

В статті описано основні проблеми, які виникають при розробці рекомендаційних систем на основі колаборативної фільтрації та запропоновано підходи до їх вирішення, а саме використання кластерного аналізу та згладжування за допомогою нейронної мережі на основі радіальної базисної функції.

Annotation

In this paper is described the main problems that arise when developing a recommendation system based on collaborative filtering and approaches for their solution, namely the use of cluster analysis and smoothing using a neural network based on radial basis functions

Ключові слова: колаборативна фільтрація, кластеризація, радіальна базисна функція, функція активації, алгоритм згладжування, розрідженість, масштабованість, самоорганізаційна карта Кохонена.

ВСТУП

Світ сучасної економіки наповнений товарами, властивості яких важко оцінити завчасно, виходячи з характеристик, даних виробником. Також важливим аспектом є те, що зазвичай таких параметрів не один десяток, і просумувати їх цінність важко. До того ж існує цілий ряд товарів, важливі параметри яких неможливо емпірично визначити (це стосується здебільшого індустрії розваг – книги, фільми, ігри).

Не зважаючи на можливу подібність характеристик різних товарів, вони можуть відрізнятися. Тому, щоб зрозуміти чи об'єкт дійсно потрібен, його треба випробувати. Однак товарів надзвичайно багато і не представляється можливим випробувати всі товари, навіть тільки ті, що схожі за характеристиками.

Рішенням даної проблеми є використання рекомендацій тих, хто вже знайомий з даним об'єктом. Колаборативна фільтрація надає колективній обробці інформації принципово нову якість. Кожний апробує свої об'єкти, а з допомогою колаборативних алгоритмів індивідуальний досвід враховується для всієї системи, звільняючи користувачів від непотрібних проб і помилок [1].

Експертні фільтри не справляються з потоком інформації. Колаборативна технологія дозволяє урівноважити цей дисбаланс: який би темп не набрала генерація контенту, їй протидіє не менш продуктивна масова селекція.

Колаборативна фільтрація дозволяє встановити паритет між обізнаністю користувача і виробника про якість товару, блокуючи тенденцію погіршення відбору, яка розвивається в умовах інформаційної асиметрії.

Мета дослідження: підвищення точності рекомендацій систем на основі колаборативної фільтрації.

Об'єкт дослідження – рекомендаційні системи на основі колаборативної фільтрації.

Предмет дослідження: застосування кластерного підходу до систем колаборативної фільтрації.

Наукова новизна полягає в розробці нових алгоритмічних рішень для систем колаборативної фільтрації, де на відміну від існуючих усунуто такі проблеми як масштабованість та розрідженість оцінок.

Потужний інструментарій рекомендаційних систем, що знайшов широке застосування в різних галузях, допомагає виявити об'єкти, які пропонуються експертами користувачам з урахуванням їх загальної популярності, демографічних характеристик та аналізу поведінки, що і визначає актуальність удосконалення алгоритмів фільтрації систем означеного класу.

ФОРМАЛІЗАЦІЯ ЗАДАЧІ КОЛАБОРАТИВНОЇ ФІЛЬТРАЦІЇ

Формально проблема вироблення рекомендацій може бути представлена таким чином: нехай C – множина користувачів, S – множина пропонованих об'єктів. Потужність множини пропонованих об'єктів S та множини користувачів C може сягати сотень тисяч і навіть мільйонів одиниць. Функція корисності u_{ij} описує корисність об'єкта s_j для користувача c_i , де i – порядковий номер користувача, j – порядковий номер об'єкта [2]:

$$u_{ij} : c_i \times s_j \rightarrow R, \quad (1)$$

де R – кількість замовлених об'єктів ($R \in N$). Для кожного користувача $c_i \in C$ обирається такий об'єкт $s'_j \in S$, для якого значення корисності максимальне для даного користувача:

$$\forall c_i \in C, s'_j = \arg \max_{s_j \in S} u_{ij}. \quad (2)$$

На основі базової технології, рекомендаційні системи у широкому змісті класифікуються на контент-орієнтовані та системи спільної фільтрації (Collaborative filtering, CF). CF підхід є найбільш часто використовуваним в рекомендаційних системах. При прогнозуванні потенційного інтересу даного користувача,

такий підхід спочатку визначає безліч подібних користувачів на основі попередніх оцінок, а потім робить прогноз, заснований на поведінці цих подібних користувачів.

Принципи колаборативної фільтрації [3]:

1. Ті, хто погоджувався у минулому, схильні погодитися і в майбутньому.
2. Користувачі надають дані про свої уподобання, оцінюючи товари інтернет-магазину або вітрини.
3. Чим більше даних надає про себе користувач, тим більш якісні рекомендації він отримує.
4. Чим більше сумарний обсяг даних, тим краще рекомендації.

У колаборативних рекомендаційних системах використовуються різні підходи для обчислення міри схожості між користувачами. У більшості з них схожість між двома користувачами ґрунтується на тому, як було оцінено аналізуємі об'єкти. Набули поширення кореляційний метод і метод лінійної схожості.

При цьому кореляційний метод можна представити таким чином: нехай S_{xy} – множина об'єктів оцінених користувачами x та y , тобто

$$S_{xy} = \{s \in S \mid r_{x,s} \neq \emptyset \& r_{y,s} \neq \emptyset\}. \quad (3)$$

де s – об'єкт;

$r_{x,s}, r_{y,s}$ – оцінки об'єкта s користувачами x та y .

Алгоритм лінійного методу знаходить найближчого сусіда, обчислюючи відстані між користувачами, основані на їх інформації про вподобання та рекомендації [2]. Відстані залежать від домену, числа користувачів, числа рекомендованих елементів і ступеню співпраці між користувачами. Прогноз того, як оцінить користувач певний об'єкт обчислюється як «середнє зважене» оцінок користувачів зі спільним інтересом стосовно цього елемента. Оскільки пошук таких користувачів у потужних базах даних є тривалим процесом, практичні алгоритми використовують евристичні для розв'язання даної задачі.

АНАЛІЗ ПРОБЛЕМ КОЛАБОРАТИВНОЇ ФІЛЬТРАЦІЇ

Серед основних проблем слід відзначити:

1. Проблеми на етапі “cold start” [4], до яких відносять:

— Проблема нового користувача. Нові користувачі не мають зв'язку з існуючими і таким чином не можуть отримати рекомендацій. Система має вивчити вподобання користувача на підставі даних ним оцінок для вироблення точних рекомендацій. Для вирішення цієї проблеми було запропоновано декілька методик. Більшість з них використовують гібридний рекомендаційний підхід, що поєднує тематичні і колаборативні алгоритми. У цих методиках використовуються стратегії, основані на популярності об'єктів, їх ентропії, персоналізації користувачів і комбінації всіх цих методик.

— Проблема нового об'єкту. Нові об'єкти регулярно додаються до рекомендаційної системи. Колаборативні системи при виробленні рекомендацій керуються тільки потребами користувачів. Тому рекомендаційна система не може рекомендувати об'єкт, поки він не отримає достатню кількість оцінок.

2. Розрідженість оцінок. У будь-якій рекомендаційній системі, кількість оцінок, які необхідно передбачити, перевищує кількість даних оцінок. Система повинна прогнозувати оцінки, виходячи з мінімальної кількості об'єктів та користувачів. Усунути проблему розрідженості оцінок можна за рахунок використання профіля користувача при пошуку міри схожості.

Колаборативна фільтрація побудована виключно на кореляції між користувачами. Тобто, рекомендації для активних користувачів надаються на підставі оцінок подібних користувачів. У разі високої розріджених матриць (> 90% нульових значень), дуже важко знайти кореляцію між користувачами.

3. Масштабованість. У колаборативних рекомендаційних системах користувачі зі схожими вподобаннями визначаються на основі пам'ять-орієнтованого або модель-орієнтованого підходів. У пам'ять-орієнтованому підході, заснованому на активних користувачах порівняно з усіма існуючими користувачами, тому обчислювальна складність збільшується при збільшенні числа користувачів. У модель-орієнтованому підході існуючі користувачі згруповані на основі їх подібності. При виробленні рекомендації, подібні групи групуються разом, а тільки потім порівнюються користувачі всередині кластера. Такий підхід не надто впливає на вирішення проблеми масштабованості. Традиційні підходи не виконують попереднє оффлайн обчислення, і тому кількість обчислень безпосередньо при запиті збільшується з числом користувачів і об'єктів. Алгоритм неможливо використовувати на великих наборах даних, якщо тільки він не використовує скорочення розмірності зменшуючи рівень якості рекомендацій.

Застосування кластеризації надає такі переваги:

1. Краще розуміння структури даних.
2. Багаторазове підвищення продуктивності.
3. Відсіювання шумів у вигляді величезної кількості незначущих даних.

Застосування простої кластеризації або статистичних моделей кластеру вважається перспективним засобом для проблеми розрідженості оцінок. Замість використання розріджених матриць рейтингу користувач-об'єкт, матриця спочатку згладжується на основі середнього рейтингу користувача.

Інтелектуальний інтегрований підхід вироблення рекомендацій використанням мережі побудованої на радіальних базисних функціях (Radial Basis Functions Network, RBFN) і колаборативних фільтрів дозволяє вирішити наведені вище проблеми. Причина для згладжування основана на теоремі Ковера про відокремленість

образів [5]. Так, RBFN використовується для згладжених розріджених даних. Згладження вирішує традиційні проблеми, такі як розрідженість та «cold start».

Зниження розмірності вхідних даних при кластеризації дозволяє:

1. Зменшити обсяг обчислень при розрахунку відстані між векторами.
2. Підвищити якість кластеризації за рахунок особливих властивостей векторів головних компонент.
3. Зменшити час кластеризації.

Розглянемо вирішення проблем:

1. масштабованості – шляхом об'єднання користувачів в кластери,
2. розрідженості оцінок – згладжуванням розріджених матриць рейтингу «користувач-об'єкт» за допомогою RBFN
3. холодного старту – через рекомендацію популярних об'єктів.

Таким чином, попередня обробка(навчання), тобто ресурсозатратний розподіл множини користувачів на кластери схожих користувачів, буде проводитись завчасно. Потім онлайн обчислення (рекомендація) проводиться на основі числа кластерів. Таким чином, проблема масштабованості знижується до певної міри.

Щоб подолати труднощі викликані розрідженістю вхідних даних, розріджена матриця вподобань користувачів перетворюється в повну матрицю за допомогою функції згладжування. При згладжуванні кореляція між користувачами посилюється (число кластерів до згладжування > число кластерів після згладжування). Так, використання згладжувальної моделі забезпечує високу якість рекомендацій.

ОРГАНІЗАЦІЯ НАВЧАННЯ РЕКОМЕНДАЦІЙНОЇ СИСТЕМИ З ВИКОРИСТАННЯМ КЛАСТЕРИЗАЦІЇ

Важко знайти корельованих користувачів у дуже розріджених матрицях. Тому матриця «користувач-об'єкт» повинна бути згладжена таким чином, щоб вона стала повною. Згладжування проводиться за допомогою RBFN.

Нехай маємо:

1. Групу M користувачів $\{u_i \mid i = 1, 2, \dots, M\}$.
2. Групу N різних об'єктів $\{p_j \mid j = 1, 2, \dots, N\}$.
3. Рейтингову таблицю r_{ij} - матрицю $[M \times N]$, яка містить оцінки i-ого користувача j-ого об'єкта в області оцінок. Неоцінені об'єкти представлені нульовими значеннями.

Нейронна мережа на основі радіальних базисних функцій є сімейством штучних нейронних мереж, яке має три шари: вхідний шар, прихований шар і вихідний шар. Вхідний шар містить M число нейронів, до яких на вхід подається рейтинг вектор користувачів. Цей шар повністю пов'язаний з усіма нейронами в прихованому шарі. Кожен нейрон в прихованому шарі має функцію активації. Прихований шар також повністю підключений до вихідного шару. Вихідний шар містить M число нейронів, до якого згладжується рейтинг-вектор користувачів. Вихідний шар виконує просту функцію підсумовування користувачів.

Функція активації

Техніка радіальної базисної функції полягає у виборі функції F [5]:

$$F(X_i) = \sum_{k=1}^K w_k \phi(\|X_i - C_k\|) \quad (4)$$

де w_k позначає вектор ваги від прихованого шару до вихідного шару, X_i - дана множина точок, C_k є центром даної множини точок, $\|\cdot\|$ визначає норму і ϕ є функцією активації. Зазначимо три з них:

1) Функція Гауса:

$$\phi(r) = \exp\left(-\frac{r^2}{2\sigma^2}\right), \text{ для } \sigma > 0. \quad (5)$$

2) Багатокритеріальна функція:

$$\phi(r) = r^\beta, \quad (6)$$

де $\beta > 0$ додатне непарне число.

3) Згладжуюча функція Thin-plate:

$$\phi(r) = r^k \log(r), \text{ де } k > 0. \quad (7)$$

де $r = \|X_i - C_k\|$, σ, β, k - додатні параметри.

Згладжування і алгоритм кластеризації.

Три алгоритми для згладжування розріджених рейтингових матриць та кластеризації

1) Згладжування

Вхід: Розріджена матриця оцінок користувачів $\langle r_{ij} | 1 \leq i \leq M, 1 \leq j \leq N \rangle$, $\phi_i^{\max}, \phi_i^{\min}$ - максимум та мінімум функції активації для об'єкту i , r_{ij} - рейтинг об'єкта i від користувача j .

Вихід: Згладжена матриця оцінок користувачів $\langle r'_{ij} | 1 \leq i \leq M, 1 \leq j \leq N \rangle$

Алгоритм:

1. $\text{range} = (\max_rating - \min_rating) + 1$.

2. Знайти кількість кластерів k так, щоб $\frac{\text{range}}{k} \leq 3$.

3. Для кожного j :

3.1. Розділити користувачів на k кластерів за допомогою самоорганізаційної карти Кохонена.

3.2. Обчислити центри $C_k = \frac{\sum_{p=1}^{k1} r_{ip}}{k1}$, де $k1$ - кількість користувачів, що відносяться до даного кластеру.

3.3. Розрахувати матрицю евклідових відстаней $g_{ip} = \|r_{ip} - C_k\|$, де $1 \leq p \leq k1$. Розрахувати функцію активації $\phi_{ip}(g)$ за допомогою (5), (6) або (7).

3.4. Розрахувати ваги використовуючи функцію псевдовипадкових ваг (8):

$$\omega_i = \frac{(\phi_i^{\max} - \phi_i(r_{i1})) / (\phi_i^{\max} - \phi_i^{\min})}{\sum_{x=1}^n (\phi_{i,x}^{\max} - \phi_{i,x}(r_{xi})) / (\phi_x^{\max} - \phi_x^{\min})}, \quad 1 \leq i \leq M \quad (8)$$

3.5. Обчислити $r'_{ij} = F(r_{ij})$ за допомогою формули (4).

2) Кластеризація за допомогою самоорганізаційної карти Кохонена

Вхід: Вектор оцінок об'єкта $\langle x_i | 1 \leq i \leq M \rangle$, α_{old} - швидкість навчання, $\omega_j(t)$ - вектор ваг в момент часу t , $n1$ означає кількість ітерацій.

Вихід: Розподілені k векторів.

Алгоритм:

1. Ініціалізація випадковими значеннями ваг $w_j = \{w_{1j}, w_{2j}, \dots, w_{mj}\}$ для кожного нейрона j з вихідної карти.

2. Випадково обрати для розгляду вектор x .

3. Обчислити відстань d_j (в M -вимірному просторі),

$$d_j = \sqrt{\sum (x_i - w_{ij})^2}. \quad (9)$$

4. Обчислити нейрона-переможця – нейрона k , який найближче до вектора x .

5. Змінити всі вектори ваг

$$\alpha_{new} = \alpha_{old} \cdot 0.5 \quad (10)$$

$$w_j(t+1) = w_j(t) + \alpha_{new} \cdot (x - w_j(t)) \quad (11)$$

6. Повторити кроки 2-5 $n1$ разів.

3) Кластеризація з використанням кореляційної функції.

Вхід: Згладжена матриця оцінок користувачів $\langle r'_{ij} | 1 \leq i \leq M, 1 \leq j \leq N \rangle$, t - набір об'єктів, оцінених користувачами i та j , $\bar{r}_i, \bar{r}_j$ - середні оцінки користувачів i та j відповідно.

Вихід: Розподілені k кластерів ($Clus_k$ - набір корельованих користувачів).

Алгоритм:

Для кожного користувача i ($1 \leq i \leq M$, схожість обчислюється між M користувачами).

1. $\text{sim_count}(i) = 0$.

2. Для кожного користувача j ($1 \leq j \leq M$)

3. Знайти схожість користувачів i та j за формулою кореляції Пірсона (12):

$$\text{sim}_{i,j} = \frac{\sum_{t \in \bar{r}_i \Delta \bar{r}_j} (r'_{it} - \bar{r}_i) \cdot (r'_{jt} - \bar{r}_j)}{\sqrt{\sum_{t \in \bar{r}_i \Delta \bar{r}_j} (r'_{it} - \bar{r}_i)^2} \sqrt{\sum_{t \in \bar{r}_i \Delta \bar{r}_j} (r'_{jt} - \bar{r}_j)^2}} \quad (12)$$

4. Якщо $\text{sim}_{i,j} = 1$, то $\text{sim_count}(i) = \text{sim_count}(i) + 1$.
5. Визначити k користувачів, які мають навищу подібність, як голову кластера для кластера k .
6. Позначити користувачів до відповідних кластерів.
7. Повторити пункти 2 та 3 для всіх M користувачів.

ОСОБЛИВОСТІ ОРГАНІЗАЦІЇ ФУНКЦІОНУВАННЯ РЕКОМЕНДАЦІЙНОЇ СИСТЕМИ

Архітектура рекомендаційної системи показано на рис. 1

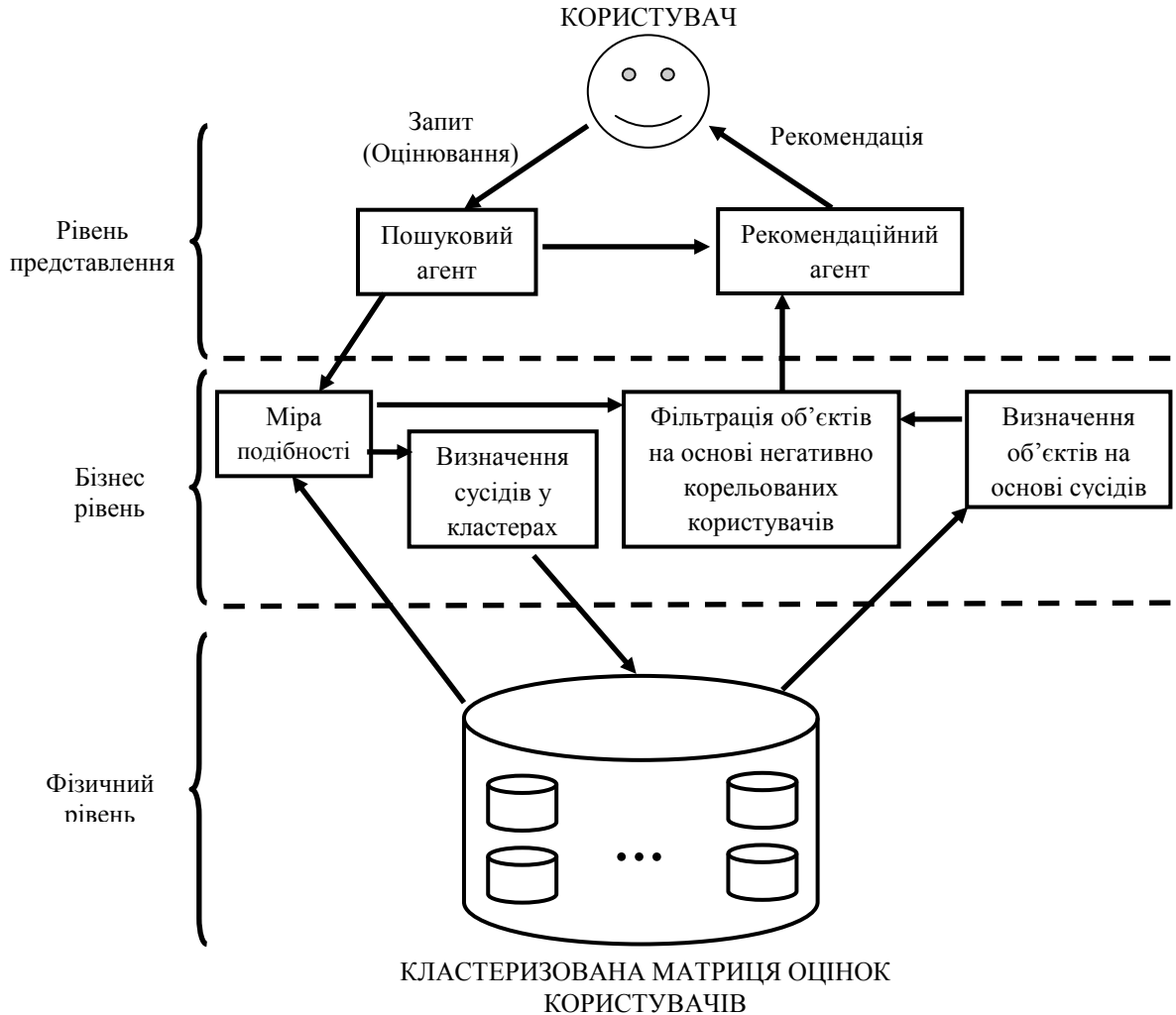


Рисунок 1 – Архітектура рекомендаційної системи

Функціональні характеристики:

1. Якщо немає оцінених об'єктів (проблема холодного старту), потім відправити T найпопулярніших об'єктів з кожного кластера рекомендаційному агенту.
2. В іншому випадку, визначаємо позитивні і негативні сусідні кластери використовуючи кореляційну функцію Пірсона [6]:

$$\text{sim}_{i,j} = \frac{\sum_{t \in r_{i \Delta r_j}} (r'_{it} - \bar{r}_i)(r'_{jt} - \bar{r}_j)}{\sqrt{\sum_{t \in r_{i \Delta r_j}} (r'_{it} - \bar{r}_i)^2} \sqrt{\sum_{t \in r_{i \Delta r_j}} (r'_{jt} - \bar{r}_j)^2}}. \quad (13)$$

3. Визначаємо найближчі позитивні і негативні сусіди з корельованих кластерів використанням Косинус-функції схожості:

$$\text{csim}_{i,j} = \frac{\sum_{k=1}^N r_{ik} \cdot r_{jk}}{\sqrt{\sum_{k=1}^N r_{ik}^2} \cdot \sqrt{\sum_{k=1}^N r_{jk}^2}}. \quad (14)$$

4. Прогнозування оцінки об'єктів з нульовою оцінкою за допомогою функції прогнозування. Обираємо підмножину K найбільш подібних користувачів на основі їх схожості з активним користувачем. Середньозважене відхилення від сусіда [7]:

$$P_{ui} = \bar{r}_i + \frac{\sum_{m=1}^c (r'_{mi} - \bar{r'_m}) \cdot c\text{sim}_{u,m}}{\sum_{m=1}^c c\text{sim}_{u,m}} \quad (15)$$

5. У крок 4 для негативну кореляцію користувачів.
6. Позначимо X, як множина рекомендованих об'єктів прогнозована на підставі результатів позитивних найближчих сусідів і Y, як множина рекомендованих об'єктів прогнозована на основі результатів негативних найближчих сусідів.
7. Розрахувати $Z = X - Y$. Множина Z направляється до рекомендаційного агента.
8. Рекомендувати множину об'єктів T активному користувачу.

Перспективи використання:

1. Впровадження в електронні магазини і вітрини для полегшення пошуку товарів, які цікавлять користувачів
2. Вивчення користувацьких вподобань для розміщення максимально затребуваних товарів
3. Можливість отримати різні вибірки користувачів по інтересам, підбір однодумців та користувачів зі спільними вподобаннями [6].

ВИСНОВКИ

Рекомендаційні системи як нові маркетингові стратегії відіграють важливу роль, особливо в електронній торгівлі. Серед різних рекомендаційних систем, колаборативні рекомендаційні системи (CRS) широко використовуються в ряді різних систем. Інтелектуальні підходи до вироблення рекомендацій пропонуються з метою подолання традиційних проблем CRS. Аналіз різних алгоритмів колаборативної фільтрації показують, що запропонований підхід є більш ефективним, ніж інші існуючі підходи з точки зору точності та актуальності вироблених рекомендацій.

Для подолання проблем колаборативної фільтрації, рекомендаційна система повинна використовувати гібридні технології (колаборативна та контент-фільтрація), щоб забезпечити точність рекомендацій.

Завдяки використанню кластерного аналізу для розбиття користувачів за схожістю вподобань вдосконалений алгоритм роботи рекомендаційної системи забезпечує зменшення часу на обробку даних та підвищує точність рекомендацій.

Таким чином, колаборативна фільтрація з попереднім згладженням набору даних та кластеризацією користувачів забезпечує високу якість рекомендацій для користувачів.

ЛІТЕРАТУРА

1. Xiaoyuan Su. A Survey of Collaborative Filtering Techniques / Xiaoyuan Su, Taghi, M. Khoshgoftaar. – Hindawi Publishing Corporation USA, 2009. – 215 p. – ISBN:1- 85233-661-7.
2. Сегаран Т. Программируем коллективный разум. / Сегаран Т.; пер. с англ. А. Слинкина – СПб: Символ-Плюс, 2008. – 368 с., ил. – ISBN 5-93286-119-3.
3. Gediminas Adomavicius and Alexander Tuzhilin, "Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions," IEEE Trans. on Knowledge and Data Engineering, vol. 17, no. 6, June 2005, pp. 734-749.
4. Bardul M. Sarwar, George Karypis, Joseph A. Konstan, and John T. Riedl, "Analysis of recommendation algorithms for e-commerce," in Electronic Commerce, 2008.
5. M. Deshpande and G. Karypis. Item-based top-n recommendation algorithms. ACM Trans. Inf. Syst., 22(1):143–177, 2004.
6. B. Mobasher, "Recommender systems," Kunstliche Intelligenz, Special Issue on Web Mining, vol. 3, pp. 41–43, 2007.
7. J. Herlocker, J. Konstan, L. Terveen, and J. Riedl. «Evaluating collaborative filtering recommender systems», ACM Transactions on Information Systems, Vol. 22(1), 2004.