Інформаційні технології

УДК 004.6

Федько Микола Романович

студент

Національного технічного університету України

«Київський політехнічний інститут імені Ігоря Сікорського»

ДОСЛІДЖЕННЯ МЕТОДІВ ПОБУДОВИ РЕКОМЕНДАЦІЙНИХ СИСТЕМ ДЛЯ ВИРІШЕННЯ ЗАДАЧ В ЕЛЕКТРОННІЙ КОМЕРЦІЇ, МЕТОДИ КОЛАБОРАТИВНОЇ ФІЛЬТРАЦІЇ

Анотація. *У роботі представлено способи побудови рекомендаційних систем, а саме метод колаборативної фільтрації. Також проведений аналіз методу, визначені переваги та недоліки.*

Ключові слова: *рекомендаційні системи, інтернет-магазин, користувачі, колаборативна фільтрація, подібність елементів, навчання моделі, рекомендації.*

Вступ. У сучасному світі на ринку товарів та послуг користувачам надається надзвичайно великий обсяг пропозицій. Таке різноманіття породжує проблему вибору найвигіднішої пропозиції, користувачеві важко оцінити та обрати оптимальний варіант, тому користувачі все частіше покидають сторінки інтернет–магазинів не виконавши покупку.

Щоб допомогти користувачам з вибором, використовуються рекомендаційні системи, а способи підбору рекомендацій постійно вдосконалюють та оновлюють задля підвищення ефективності, оскільки результат роботи цих систем значною мірою впливає на прибуток компанії–користувача.

Тож серед переваг використання рекомендаційних систем можна виділити такі можливості:

Збільшення кількості проданих товарів

Це мабуть найважливіша функція комерційних рекомендаційних систем, яка допомагає продати додаткові позиції до товарів, які продаються без будь-яких рекомендацій, тим самим збільшити середній чек покупки. Завдяки рекомендаційним системам ця мета досягається, оскільки рекомендовані товари, ймовірно, відповідають потребам користувачів.

Покращення залученості користувачів

Некомерційні програми мають подібні цілі, навіть якщо для користувача немає витрат, але є потреба у виборі контенту. Наприклад, тематична мережа націлена на збільшення кількості новин, прочитаних на її сайті. Загалом, можна сказати, що з точки зору постачальника послуг, основною метою введення рекомендаційних систем є збільшення коефіцієнта конверсії, тобто кількості користувачів, які отримують рекомендації і споживають товар, порівняно з кількістю простих відвідувачів, які просто переглядають інформацію.

Продаж широкого спектру товарів

Ще однією важливою функцією рекомендаційних систем є надання можливості користувачеві для вибору товари, які важко знайти без точної рекомендації. Наприклад такий постачальник послуг, як платформа Netflix, зацікавлені в тому, щоб користувачі переглядали якомога більше різноманітного відео контенту з каталогу, а не лише найпопулярнішого. Без рекомендаційних систем для постачальника послуг може бути досить ризиково рекламувати фільми, які, ймовірно, не відповідатимуть смакам конкретного користувача. Тож рекомендаційні системи допомагають просувати непопулярні фільми зацікавленим користувачам.

Підвищення лояльності користувачів

Користувач буде лояльним до веб-сайту, який під час відвідування впізнає старого клієнта і ставиться до нього, як до цінного відвідувача. Тож цінною особливістю рекомендаційних систем є те, що вони враховують взаємодію та попередні оцінки, виставлені рейтинги для певного контенту, це дозволяє рекомендаційним системам надавати релевантні рекомендації.

Краще розуміння користувацьких вподобань

Рекомендаційні системи допомагають прогнозувати майбутні продажі на основі історичних даних користувачів. Надалі постачальник послуг може використовувати отриману інформацію для ряду інших цілей, наприклад планування закупок певних товарів [1].

Постановка задачі. Питання підвищення ефективності підбору рекомендацій за допомогою методів колаборативної фільтрації для інтернет –магазинів, а саме: скорочення часу навчання моделі, тобто час, який необхідний для того, щоб система навчилась визначати релевантні пропозиції, що прямо впливатиме на показник конверсії інтернет–магазину залишається актуальним.

Тому наша задача полягає в розробці способу підбору рекомендацій за допомогою методів колаборативної фільтрації який дозволить скоротити час навчання моделі.

Аналіз існуючих способів побудови рекомендаційних систем

На практиці для побудови рекомендаційних систем використовуються методи контентної фільтраці, колаборативної фільтрації та гібридні методи, проте найпопулярнішим є саме метод колаборативної фільтрації. У свою чергу в колаборативній фільтрації виділяються такі підходи: Memory-based метод, який базується на аналізі оцінок елементів, Model-based метод, який базується на аналізі моделі даних та гібридний метод, що поєднує попередні два методи [2; 3; 4].

У роботі [2] запропонований Memory-based метод, який базується на аналізі оцінок елементів. Алгоритми Memory-based методу базуються на статистичних методах, основним завдання яких є виділення групи користувачів, близьких до конкретного користувача. В цьому методі використовуються попередні оцінки, які зробив користувач, і аналіз оцінок інших користувачів. Цей підхід ще називають методом найближчих сусідів: використання попередніх оцінок, які зробив клієнт, і аналіз оцінок інших користувачів, які мають подібні уподобання. Тоді рекомендації (прогноз) для цільового користувача формується на основі обрахунків деякої міри подібності по всіх накопичених даних.

Memory-based метод як метод знаходження подібності поділяється на User-based метод, який базується на аналізі подібності користувачів та Item-based метод, який базується на аналізі подібності елементів. Ціллю обох напрямків є виділення подібних об'єктів в групи на основі матриці оцінок. В першому випадку визначається подібність користувачів: знайти інших користувачів, чиї минулі оцінки поведінки схожі на ті, що і в поточного користувача, і використати їх оцінки інших елементів для прогнозування інтересів поточного користувача. Другий підхід, на основі подібності елементів, використовується в Amazon.com у наш час. В цьому випадку замість того, щоб використовувати подібність між поведінкою користувацьких оцінок для прогнозування інтересів, використовується подібність між оцінками моделей елементів. Якщо два елементи, як правило, мають однакові оцінки користувачів, то вони схожі, і користувачі повинні мати аналогічні інтереси для подібних елементів.

Також у роботі [2] запропонований Model-based метод, який базується на аналізі моделі даних. В цьому випадку спочатку по сукупності оцінок формується модель інтересів користувачів, товарів і взаємозв'язків між ними, а потім формуються рекомендації на основі отриманої моделі. Процес формування рекомендацій розбитий на два етапи: складне та

ресурсномістке навчання моделі і достатньо простий обрахунок рекомендацій на основі існуючої моделі в режимі реального часу. Ці алгоритми можуть базуватись на імовірнісному підході, кластерному аналізі, аналізі прихованих факторів.

Розглядаючи недолік розрідженості даних, ми бачимо, що як правило, більшість комерційних рекомендаційних систем основані на великій кількості даних (товарів), в той час як більшість користувачів не ставить оцінки товарам. В результаті цього матриця «предмет-користувач» виходить дуже розрідженою, що створює проблеми при формуванні рекомендацій. Ця проблема особливо гостра для нових, щойно створених систем. Також розрідженість даних посилює проблему холодного старту.

При розгляді проблеми холодного старту, новий предмет або користувач являють собою велику проблему для рекомендаційних систем. Частково цю проблему допомагає вирішити контекстний підхід, оскільки він базується не на оцінках, а на атрибутах, що дозволяє включати нові предмети в рекомендації користувачів. Однак проблему із наданням рекомендацій для нового користувача вирішити складніше.

До недоліків колаборативної фільтрації відносять синоніміку, яка являє собою тенденцію схожих та однакових предметів мати різні назви. Більшість рекомендаційних систем не здатні виявляти ці приховані зв'язки і тому відносяться до цих предметів як до різних. Наприклад, «фільми для дітей» і «дитячий фільм» відносяться до одного жанру, але система сприймає їх як різні.

Варто також врахувати, що колаборативна фільтрація була створена для надання рекомендацій та збільшення різноманітності, що дозволяє користувачам відкривати для себе нові продукти з великої кількості пропозицій. Однак, оскільки алгоритм колаборативної фільтрації базується на рейтингах та реальних продажах, то використання цього методу може

привести до зменшення різноманітності, новим продуктам у такому разі складно стати рекомендованими.

Таким чином, враховуючи вищесказане, надзвичайно важливим є створення нового способу підбору рекомендацій, який би мав модель яка швидко навчається при незначній кількості даних та зберігає властивість різноманіття в наданих користувачам пропозиціях.

Спосіб підбору рекомендацій з використанням Memory-based методів колаборативної фільтрації

Спосіб базується на аналізі оцінок елементів та формування груп подібних елементів. У якості вхідних даних розглядаються:

U – множина суб'єктів (клієнтів, користувачів: users);

R – множина об'єктів (ресурсів, товарів, предметів: items);

Y – простір описів транзакцій;

$D = (u_i, r_i, y_i)_{i=1}^{m} \in U \times R \times Y$ – транзакційні дані;

Агреговані дані:

$F = f_{ur}$ — матриця крос-табуляції розміру |U| × |R|, де $f_{ur} = aggr\{(u_i, r_i, y_i) \in \mathrm{D} \mid u_i = u, r_i = r\}$

Задачі:

- Прогнозування незаповнених комірок f_{ur}
- Оцінювання подібності: $p(u, u')$, $p(r, r')$, $p(u, r)$

Опис підходів на основі сусідства

- Тривіальна рекомендаційна система:

«Клієнти, які купляли товар r_0, також купляли $R(r_0)$»

1) $U(r_0) := \{ u \in U \mid f_{ur_0} \neq \emptyset, u \neq u_0\}$ – колаборація;

2) $R(r_0) := \{r \in R \mid B(r) = \left|\frac{U(r_0) \cap U(r)}{U(r_0) \cup U(r)}\right| > 0\}$;

де $B(r)$ - одна із можливих пар близькості r і r_0;

3) Відсортувати $R(r_0)$ у порядку спадання $B(r)$, взяти top N;

Недоліки підходу:

- Рекомендації тривіальні (пропонується все найбільш популярне)
- Не враховуються інтереси конкретного користувача u_0
- Проблема «холодного старту» (новий товар нікому не рекомендується)
- Потрібно зберігати всю матрицю F

Фільтрація по подібності користувачів (user-based)

«Клієнти, схожі на u_0, також купляли $R(u_0)$».

Даний підхід має високу точність, але недоліком є вибагливість до пам'яті (через велику кількість складних обрахунків, які потрібні для отримання рекомендацій). Також обрахунки степені подібності можуть відбуватися тільки в реальному часі, оскільки дані про поточну транзакцію стають доступні тільки в момент формування рекомендацій. Тому даний підхід може застосовуватись тільки у відносно невеликих базах даних.

1) $U(u_0) := \{ u \in U \mid corr(u, u_0) > \alpha \}$ – колаборація;

де $corr(u, u_0)$ – одна із можливих мір близькості u до u_0;

2) $R(u_0) := \{r \in R \mid B(r) = \left|\frac{U(u_0) \cap U(r)}{U(u_0) \cup U(r)}\right| > 0\}$;

де $U(r) := \{u \in U \mid f_{ur} \neq \emptyset\}$;

3) Відсортувати $r \in R(u_0)$ по спаданню $B(r)$, взяти top N;

Недоліки:

- ~~Рекомендації тривіальні~~
- ~~Не враховуються інтереси конкретного користувача u_0~~
- Проблема «холодного старту»
- Потрібно зберігати всю матрицю F
- Нічого рекомендувати нетиповим/новим користувачам

Фільтрація по подібності елементів (item-based)

«Разом з об'єктами, які купляв клієнт u_0, часто купують $R(u_0)$». В цьому алгоритмі степінь подібності елемента, що аналізується, до інших може бути обрахована у відкладеному режимі, за розкладом. Оскільки

вектори рейтингів усіх елементів доступні до моменту формування рекомендації. Таким чином цей алгоритм є більш ефективним з точки зору часу формування рекомендацій завдяки можливості проведення відкладеної обробки даних.

1) $R(u_0) := \{ u \in R \mid \exists r_0 : f_{u_0 r_0} \neq \emptyset, B(r) = corr(r, r_0) > \alpha \}$;
   де $corr(r, r_0)$ – одна із можливих мір близькості r до r_0;
2) Відсортувати $r \in R(u_0)$ по спаданню $B(r)$, взяти top N;

Недоліки:

- Рекомендації часто тривіальні (немає колаборативності)
- Проблема «холодного старту»
- Потрібно зберігати всю матрицю F
- ~~Нічого рекомендувати нетиповим/новим користувачам~~

Для описаних методів є необхідність у зберіганні усієї матриці даних, тобто інтересів користувачів. У зв'язку з цим виникають труднощі при прогнозуванні інтересів для нових користувачів або при появі нових елементів, оскільки для них ще немає оцінок. Також обмежується можливість методів при обробці великих об'ємів даних. У багатьох випадках зберігання усієї матриці інтересів є надлишковим: зазвичай користувачі та елементи діляться на групи з аналогічними характеристиками. Наприклад, багато науково-фантастичних фільмів подобатися схожим між собою групам користувачів. Тому виникає задача пониження розмірності матриці оцінок. Такі задачі вирішують Model-based методи.

У такому випадку можливий варіант об'єднання користувачів (елементів) в кластери (профілі) за допомогою деякого індексу подібності. Елементи і оцінки, які дали користувачі з одного кластера, використовуються для обрахунку рекомендацій. Кластерні моделі краще масштабуються, оскільки звіряють профіль користувача з відносно невеликою кількістю сегментів, а не з цілою користувацькою базою.

Складний та об'ємний кластерний підрахунок ведеться в офлайн режимі. Ця задача може бути виконана на основі різних математичних підходів.

Обчислення прогнозів рекомендацій

Найважливішим кроком у системі рекомендацій за допомогою колаборативної фільтрації є створення вихідного інтерфейсу з точки зору прогнозування.

Як тільки виділено набір найбільш подібних елементів на основі, наступним кроком є вивчення рейтингів цільових користувачів та використання техніки отримання прогнозів.

Тут ми розглянемо два таких прийоми.

- Зважена сума

Як випливає з назви, цей метод обчислює передбачення за елементом i для користувача u шляхом обчислення суми оцінок надані користувачем щодо предметів, подібних до i. Кожна оцінка зважується відповідною подібністю s_{ij} між елементами i та j.

Формально, використовуючи поняття, наведене на рисунку 1, можемо позначати передбачення P_{ui} як

$$P_{u,i} = \frac{\sum_{all\ similar\ items,N}(s_{i,N} * R_{u,N})}{\sum_{all\ similar\ items,N}(|s_{i,N}|)}$$

В основному, такий підхід намагається визначити, як активний користувач оцінює подібні елементи. Зважена сума масштабується сумою показників подібності, щоб переконатися чи прогноз знаходиться в межах заданого діапазону.

- Регресія

Цей підхід подібний до методу зваженої суми, але замість того, щоб безпосередньо використовувати рейтинги подібних предметів, які він використовує наближення рейтингів на основі регресійної моделі.

На практиці подібність, обчислена за допомогою косинусів або кореляційних вимірювань, може ввести в оману в тому сенсі, що два

рейтингові вектори можуть бути віддаленими (в евклідовому розумінні), але мають дуже високу схожість. У такому випадку використовуються необроблені рейтинги «так званого» подібного елемента може призвести до поганого прогнозування.

Основна ідея полягає у використанні тієї ж формули, що і метод зваженої суми, але замість використання подібного пункту N – рядку рейтингові значення $R_{u,N}$, ця модель використовує їх наближені значення $R_{u,N}$ на основі лінійної регресійної моделі. Якщо ми позначимо відповідні вектори цільового елемента i та аналогічний елемент N за R_i і R_N за моделлю лінійної регресії можна виразити як

$$\bar{R}'_N = \alpha \bar{R}_i + \beta + \epsilon$$

Параметри регресійної моделі α і β визначаються, оцінюючи обидва вектори рейтингу, ϵ – похибка регресійної моделі [6].

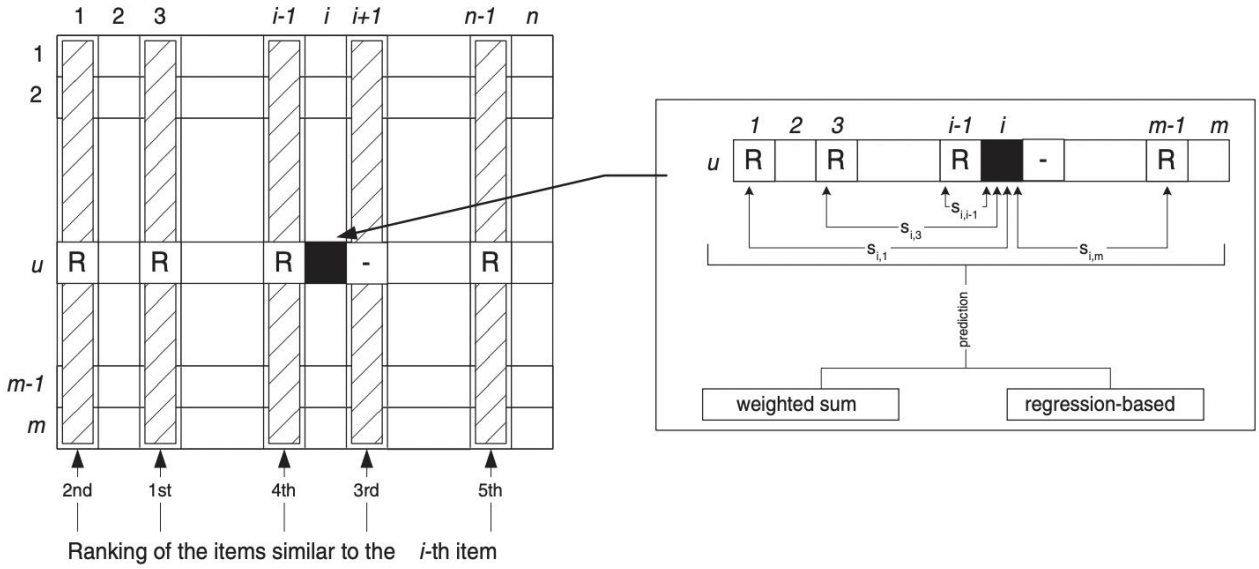

Рис. 1. Item-based алгоритм колаборативної фільтрації. Процес прогнозування на прикладі 5 сусідів

Висновки. У цій роботі був розглянутий спосіб побудови рекомендаційної системи за допомогою методів колаборативної фільтрації та способи підбору рекомендацій за допомогою Memory-based методу. Використовуючи розроблені системи показників, було порівняно

алгоритми підбору рекомендацій та проаналізано результати експериментів.Визначено, що перевагами описаних вище алгоритмів є очікуваність результатів, що є важливим аспектом рекомендаційних систем, просте створення і використання, кожен з користувачів отримає рекомендації.До недоліків можна віднести проблему «холодного старту», коли нові товари нікому не рекомендуються та необхідність зберігання великих об'ємів даних для навчання моделі, що негативно відбивається на швидкості підбору рекомендацій.

Література

1. Francesco R., Lior R., Bracha S., Paul K. B. Recommender Systems Handbook. Dordrecht: Springer, 2015. URL: https://www.cse.iitk.ac.in/users/nsrivast/HCC/Recommender_systems_handbook.pdf
2. Xiaoyuan Su and Taghi M. Khoshgoftaar "A Survey of Collaborative Filtering Techniques A Survey of Collaborative Filtering Techniques" // Hindawi Publishing Corporation, Advances in Artificial Intelligence archive, USA: 2009. С. 1-19. URL: https://www.hindawi.com/journals/aai/2009/421425/
3. Jones M. Recommender systems, Part 1.Introduction to approaches and algorithms. Learn about the concepts that underlie web recommendation engines / M. Jones. 2013. URL: https://www.ibm.com/developerworks/opensource/library/os-recommender1/index.html?s_tact=105agx99&s_cmp=cp
4. Meleshko E.V. Дослідження методів побудови рекомендаційних систем в мережі Інтернет / E.V. Meleshko, S.G. Semenov, V.D. Khokh // Системи управління, навігації та зв'язку. Збірник наукових праць. Полтава: ПНТУ, 2018. Т. 1 (47). С. 131-136. doi: https://doi.org/10.26906/SUNZ.2018.1.131.

5. Adomavicius G. На пути к новому поколению рекомендационных систем: обзор имеющихся систем и возможные инновации. IEEE Transactions on Knowledge and Data Engineering, Июнь 2005. Vol. 17. No. 6. URL: http://artpragmatica.ru/rs/in/pic/58-870-20061024072441-Toward_the_next_generation_of_recommender_systems.doc
6. Sarwar B., Karypis G., Konstan J., Reidl and J. "Item-based collaborative filtering recommendation algorithms," in Proceedings of the 10th international conference on World Wide Web. ACM New York, NY, USA, 2001. P. 285–295. URL: http://www.ra.ethz.ch/cdstore/www10/papers/pdf/p519.pdf