

ЗАСТОСУВАННЯ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ДАНИХ ПРИ КОЛАБОРАТИВНІЙ ФІЛЬТРАЦІЇ

Тамара Савчук¹, Антон Сакалюк²

Вінницький національний технічний університет
Хмельницьке шосе, 95, Вінниця, 21021, Україна, E-Mail: ²santei90@gmail.com

Анотація

Дана доповідь представляє собою огляд існуючих рекомендаційних систем і дає опис сучасних методик вироблення рекомендацій, що ґрунтуються на принципі колаборативної фільтрації. У доповіді також розглядаються можливі проблеми, які є характерними для рекомендаційних систем та пропозиції до їх усунення за рахунок використання в колаборативних фільтрах інтелектуального аналізу даних.

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.

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 $s_j \in S$ для користувача $c_i \in C$, де i – порядковий номер користувача, j – порядковий номер об'єкта [1]:

$$u_{ij} : c_i \times s_j \rightarrow R,$$

(1)

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

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

Якщо багато користувачів дають позитивну оцінку певному об'єкту, то ймовірно, що користувач, погодившись на рекомендації інших, обере даний об'єкт. Традиційні системи колаборативної фільтрації забезпечують узагальнені рекомендації шляхом підсумовування оцінок спільноти в цілому.

Колаборативні фільтри мають широку область застосування: оскільки вони не потребують аналізу самого змісту даного елемента, то легко можуть бути застосовані до будь-якого типу об'єкта.

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

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

де s_j – об'єкт;

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

Алгоритм лінійного методу знаходить найближчого користувача зі спільним інтересом, обчислюючи відстані між користувачами, основані на їх інформації про вподобання та рекомендації. 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 по-

шук таких користувачів у потужних базах даних є тривалим процесом, практичні алгоритми використовують евристичні для розв'язання даної задачі.

Фільтрація «об'єкт-об'єкт» застосовує ту ж ідею, але використовуються подібності між досліджуваними об'єктами, а не користувачами. Невідомий рейтинг об'єкту для тестового користувача може бути спрогнозований шляхом усереднення оцінок інших аналогічних об'єктів.

Обидва варіанти забезпечують високий рівень персоналізації у своїх рекомендаціях, і більшість систем, які використовують ці технології мають відповідні показники як швидкодії, так і точності прогнозування [1].

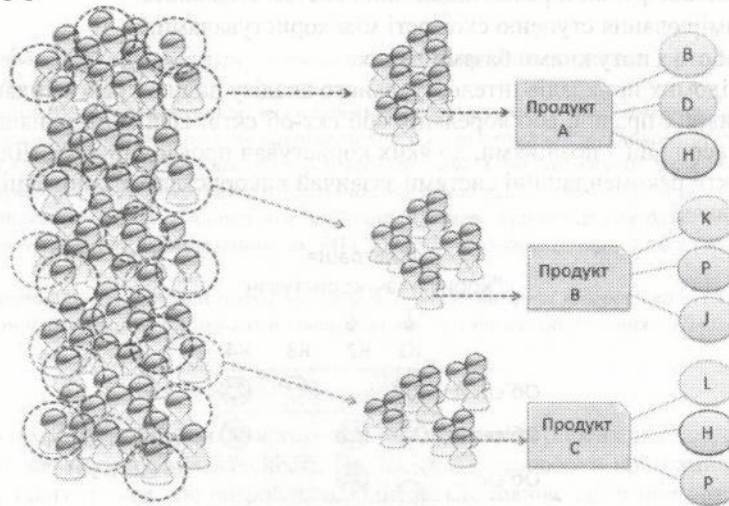


Рис. 1. Ілюстрація ідеї колаборативної фільтрації

Основними проблемами колаборативної фільтрації є проблеми на етапі “cold start” [2]:

— 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.

— Розрідженість оцінок. У будь-якій рекомендаційній системі кількість оцінок, які необхідно передбачити, перевищує кількість заданих 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.

Одним із варіантів розв'язання наведених проблем є використання класифікації, що можна реалізувати за допомогою байєсівських мереж. В байєсівських мережах створюються моделі за допомогою дерева рішень, в кожному вузлі якого представляється інформація про користувачів. Аналіз областей застосування байєсівських мереж [3] показав, що, з одного боку, вони можуть виявитися доцільними при необхідності синтезу моделі об'єктів в обмежених термінах, а з другого – не підходять для умов, в яких моделі споживчого уподобання часто оновлюються, що є характерним для колаборативної фільтрації.

Іншим підходом до розв'язання поставлених проблем є застосування методу кластеризації. Колаборативні фільтри, на відміну від контент-фільтрів, не беруть до уваги зміст самого об'єкта. Замість того, щоб витратити час на аналіз об'єктів розробники колаборативних фільтрів можуть зосередитися на дослідженні алгоритмів фільтрації інформації.

Замість того, щоб зберігати потужні масиви даних про кожного користувача – профілі користувачів визначаються за рейтингами інших користувачів зі спільним інтересом, так як дані рейтингу повинні зберігатися тільки один раз, з посиланням на користувачів.

Застосування кластерного підходу до аналізу даних про користувачів може зменшити потужність їх множини, як наслідок – збільшиться продуктивність фільтрації. Поділ множини користувачів на кластери може негативно відбитися на точності рекомендацій для користувачів, що знаходяться на межах кластера [4].

Рекомендаційна система повинна враховувати такі вхідні дані:

- контент (представлений у вигляді міток категорій);
- користувацькі рейтинги.

Серед основних функцій рекомендаційних систем виділяють:

- вимірювання ступеню схожості між користувачами;
- робота з потужними базами даних.

Одним з відомих прикладів інтелектуального аналізу даних у рекомендаційних системах є виявлення асоціативних правил, або кореляція «об'єкт-об'єкт». Цей метод виявляє об'єкти, які часто зустрічаються в асоціації з позиціями, до яких користувач проявив інтерес. Для обчислення кореляції «об'єкт-об'єкт» рекомендаційні системи зазвичай використовують поточні потреби, а не історичні дані користувача.



Рис.2. Порівняння фільтрації «користувач-користувач» та «об'єкт-об'єкт»

Аналіз традиційних методів колаборативної фільтрації показав, що їх використання базується на правилі, яке вимагає, щоб користувачі давали явні відповіді. Рекомендаційні системи використовуючи засоби інтелектуального аналізу даних можуть отримувати з інформації про соціальну активність спільноти користувачів потрібні для системи дані. Слід зауважити, що при наявній тенденції ігнорування традиційними рекомендаційними системами зміни об'єкту з плином часу, вони здатні прогнозувати попит користувача на нього. Системи, що застосовують методи інтелектуального аналізу даних, прогнозують час пропозиції певного об'єкта користувачу, а також рекомендований час для оцінювання об'єктів.

Отже, застосування інтелектуального аналізу даних в рекомендаційних системах, що передбачають включення контекстної інформації в означений процес, підвищує ефективність рекомендаційних систем за рахунок можливості використання мультикритеріальних оцінок.

Література:

- [1] Bardul M. Sarwar, George Karypis, Joseph A. Konstan, and John T. Riedl, "Analysis of recommendation algorithms for e-commerce," in Electronic Commerce, 2008.
- [2] Breese, J., Heckerman, D., & Kadie, C. "Empirical Analysis of Predictive Algorithms for Collaborative Filtering". Conference on Uncertainty in Artificial Intelligence, pp 43-52, 2006.
- [3] Konstan, J. A., Miller, B. N., Maltz, D., Herlocker, J. L., Gordon, L. R., and Riedl, J. "GroupLens: Applying collaborative filtering to usenet news". Communications of the ACM, pp 77–87, 2007.
- [4] G. Adomavicius and A. Tuzhilin, "Expert-Driven Validation of Rule-Based User Models in Personalization Applications," Data Mining and Knowledge Discovery, vol. 5, nos. 1 and 2, pp. 33-58, 2001.