УДК 004.031.43, 004.048

РОЗРОБКА КОЛАБОРАТИВНОГО СЕРЕДОВИЩА ДЛЯ ТУРИСТИЧНОЇ ГАЛУЗІ

І.А. Жирякова, Я.І. Партицький
Черкаський національний університет ім. Б. Хмельницького
e-mail: yaroslavprt@gmail.com

У сучасному світі людям постійно доводиться стикатися з проблемою раціонального вибору. Приймаючи рішення стосовно досить простих речей, наприклад: вибір фільму, який подивитись; музики, яку послухати; книги, яку почитати; новин, на які варто звернути увагу тощо; людина майже не замислюється над тим наскільки раціональним є прийняте рішення. Але як тільки «ціна» питання зростає людина починає більш серйозно замислюватись щодо правильності прийнятого рішення. Для формалізації цього процесу починаючи з 1992 року стали розвиватись алгоритми інтелектуального аналізу даних, методи кластеризації та методи колаборативної фільтрації [1].

За результатом вивчення наукових джерел, над розвитком рекомендаційних систем працюють досить багато, як закордонних так і вітчизняних дослідників, зокрема Д. Лемайр, А. Маклахлан, П. Мелвілл, Р. Дж. Муні, Р. Нагарян, Х.-Т. Торстен, А. Маршан, П. Маркс, Д. Янач, М. Занкер, О. Фелферніг, Ф. Герард, А. А. Правиков, М. М. Глибовець, С. С. Гороховський, А. А. Піка, та інші.

Для формування рекомендацій розрізняють наступні підходи: на підставі змісту (Memory-based) та на основі транзакцій (Model-based).

Суть підходу на підставі змісту полягає в тому, що рекомендації формуються для об’єктів, схожих на ті, які вже обирались користувачем, або на ті, які обирались іншими «схожими» користувачами. Ступінь «схожості» оцінюється на підставі характеристик об’єктів і користувачів. Цей підхід показав високу точність на практиці та дозволяє інкрементально враховувати нові дані (нові транзакції просто додаються в базу даних і враховуються при формуванні прогнозу поряд з уже внесеними), але він не може надати описовий аналіз існуючих залежностей та пояснити прогноз.

Суть підходу на основі транзакцій полягає в тому, що рекомендації формуються на підставі користувацької поведінки, тобто об’єкти вважаються «схожими», якщо часто входять разом в одну транзакцію, а користувачі вважаються «схожими», якщо обирають «схожі» об’єкти. Перевагою такого підходу є наявність моделі, що дає можливість краще зрозуміти сформовані рекомендації завдяки встановленим взаємозв'язкам між даними. В даному підході процес формування рекомендацій розбитий на два етапи: ресурсномістке навчання моделі в режимі off-line та досить просте обчислення рекомендацій на її основі в режимі реального часу (on-line). Однак, такі моделі не підтримують інкрементального навчання (поява нових даних вимагає оновлення всієї моделі). Точність прогнозу дещо поступається Memory-based підходу.

Дане дослідження спрямоване на вибір найбільш оптимального методу колаборативної фільтрації для створенні системи персоналізованих ревалентних рекомендацій по вибору місця проживання туриста, який подорожує територією України.

Розглянемо докладніше основну ідею формування рекомендацій CF-системами.

Будемо вважати, що «схожі» користувачі, які здійснюють «схожий» вибір об’єктів, а «схожі» об’єкти обираються користувачами спільно. В такому випадку можна розрізняти наступні способи до спільної фільтрації: фільтрацію по користувачам (User-centric) та фільтрацію по об’єктам (Item-centric).

Фільтрація по користувачам (User-centric). У цьому випадку невідомий рейтинг об’єкту виставляється на підставі рейтингів, які були проставлені тому ж об’єкту користувачами, «схожими» на даного. Цей підхід реалізується у два кроки: 1. Знайти користувачів, які здійснили вибір аналогічних об’єктів, як і даний користувач. 2.

Запропонувати користувачу об'єкти з максимальним рейтингом серед усіх об'єктів, які обирались схожими користувачами.

Фільтрація по об'єктам (Item-centric). Невідомий рейтинг будь-якого об'єкта виставляється на підставі рейтингів інших «схожих» об'єктів, уже обраних користувачем. Цей підхід реалізується також у два кроки: 1. Побудувати матрицю об'єктів для визначення ступеню «схожості» між ними. 2. Використовуючи ступінь «схожості» запропонувати об'єкти, «схожі» на ті, що вже обрані даними користувачем.

Для більш детального розгляду кожного способу фільтрації введемо наступні позначення. Нехай ми маємо дані про m товарах і n транзакціях. Тоді, позначимо кожну і-ту транзакцію m-мірним вектором $X_i := (x_{i1}, \ldots, x_{im})$, де $x_{ij}, i \in \{1, \ldots, n\}; j \in \{1, \ldots, m\}$ – рейтинг j-того об'єкту, обраного в і-тій транзакції, а кожен j-тий об'єкт n-мірним вектором $Y_i := (y_{1j}, \ldots, y_{nj})$ рейтингів j-того об'єкту в усіх транзакціях.

Тоді, фільтрація по користувачам відбуватиметься наступним чином. Якщо ми розглядаємо набір транзакцій $X_i, i \in \{1, \ldots, n\}$ як множину однаково розподілених незалежних випадкових величин, то кращим (з точки зору мінімізації середньоквадратичної помилки) прогнозом вибору користувачем j-того товару на k-тій транзакції, де $k > \mathrm{n}$ константою буде значення математичного сподівання відповідної випадкової величини, тобто $MX_{kj} = MX_{1j}$, а оцінкою цього прогнозу – середнє арифметичне за вибіркою згідно (1):

$$\overline{x_{kj}} = \frac{\sum_{i=1}^{n} x_{ij}}{n} \tag{1}$$

Сума в (1) містить кількість одиниць обраного j-того об'єкту з однаковою вагою для кожної і-тої транзакції, що відповідає припущенню про рівноцінність кожної транзакції. Метод фільтрації по користувачам передбачає, що «схожі» транзакції необхідно враховувати при прогнозуванні з більшою вагою ніж менш «схожі». Отже, замість формули (1) при прогнозуванні розміру замовлення j-того об'єкту в k-тій новій транзакції будемо використовувати формулу (2):

$$\overline{x_{kj}} = \frac{\sum_{i=1}^{n} s_{trans}(X_i, X_k) x_{ij}}{\sum_{i=1}^{n} \left| s_{trans}(X_i, X_k) \right|} \tag{2}$$

Згідно (2) чим більше «схожості» між транзакціями X_i, X_k тим з більшою вагою входить кількість одиниць обраного j-того об'єкту у і-тій транзакції у зважену суму при прогнозі.

Далі розглянемо можливі алгоритми реалізації сформульованого принципу в залежності від способу обчислення ступеню «схожості» між транзакціями $s_{trans}(X_i, X_k)$.

Одним з найбільш поширених алгоритмів, що реалізує даний принцип є алгоритм найближчого сусіда. Його суть полягає у визначенні r найближчих до даної транзакцій і використанні середнього значення їх рейтингів для прогнозування невідомих рейтингів в даній транзакції. У якості ступеню «схожості» між транзакціями може обиратись будь-яка відома метрика на просторі m-мірних векторів рейтингів об'єктів, наприклад, індукована евклідова норма (3):

$$d(X_k, X_r) = \sqrt{\sum_{l=1}^{m} (x_{kl} - x_{rl})^2} \tag{3}$$

де X_k – транзакція, яка розглядається; $X_i, i \in \{1,\ldots,n\}$ – збережена історія завершених транзакцій; а матриця X розмірності nxm з рядками у вигляді транзакцій X_i і стовпцями у вигляді рейтингу об'єктів $Y_j, j \in \{1,\ldots,m\}$.

Отже, прогноз j-того рейтингу в k-тій транзакції обчислюється як середнє значення j-того рейтингу в r найближчих до X_k транзакціях згідно (4):

$$\overline{x_{kj}} = \frac{\sum_{i \in I_r(x_k)}^{n} x_{ij}}{r}, I_r(x_k) \subset \{1,\ldots,n\} \tag{4}$$

А ступінь «схожості» між транзакціями обчислюється згідно (5):

$$s_{trans}(X_k, X_i) = \begin{cases} 1, i \in I_r(x_k) \\ 0, i \notin I_r(x_k) \end{cases} \tag{5}$$

Фільтрація по об'єктам полягає у виставленні невідомого рейтингу об'єкту в аналізованій транзакції на підставі зважених рейтингів інших об'єктів, що входять в цю транзакцію. Рейтинг об'єкту буде тим більше, чим більше рейтинг у інших об'єктів в аналізованій транзакції, які зазвичай обираються спільно з ним. Отримуємо формулу (6), аналогічну (2):

$$\overline{x_{kj}} = \frac{\sum_{i=1}^{m} s_{items}(Y_j, Y_i) x_{kj}}{\sum_{i=1}^{m} \left| s_{items}(Y_j, Y_i) \right|} \tag{6}$$

При прогнозуванні рейтингу як і у випадку фільтрації по користувачам слід використовувати його відхилення від середнього значення по всім транзакціям для даного об'єкту згідно (7):

$$\overline{x_{kj}} = \overline{y_j} + \frac{\sum_{i=1}^{m} s_{items}(Y_j, Y_i) \cdot (x_{ki} - \overline{y_i})}{\sum_{i=1}^{m} \left| s_{items}(Y_j, Y_i) \right|}, \overline{y_i} = \frac{\sum_{r=1}^{n} x_{ri}}{n} \tag{7}$$

де $\overline{y_i}$ – середній рейтинг і-того об'єкта по всім транзакціям.

На відміну від фільтрації по користувачам, де обчислення ступеня «схожості» довільної транзакції до всіх інших транзакцій може проводитися тільки в реальному часі, фільтрація по об'єктам дає можливість обчислити ступінь «схожості» аналізуємого об'єкту до всіх інших в режимі off-line за розкладом, адже вектори рейтингів усіх об'єктів доступні до моменту формування рекомендацій. Отже, процес формування рекомендацій за другим способом дає можливість розділити його на дві стадії: в режимі off-line (обчислення ступеня «схожості» об'єктів один до одного) і в реальному часі (обчислення рейтингів об'єктів).

Крім описаних способів при розробці CF-систем ще застосовується гібридний спосіб фільтрації, який полягає в отриманні оцінки невідомого рейтингу як зваженої суми оцінок на підставі фільтрації по користувачам, фільтрації по об'єктам і змішаної фільтрації (на підставі рейтингів «схожих» об'єктів у «схожих» транзакціях).

Даний спосіб передбачає отримання рейтингу на основі визначення спільного ступеня «схожості» між парами транзакція – об'єкт. Для цього використовується величина, що враховує ступінь «схожості» окремо для відповідних транзакцій і окремо для об'єктів. Комбінована ступінь «схожості» не перевищуватиме окремих ступенів «схожості» між транзакціями і об'єктами та обчислюватиметься за формулою (8):

$$s_{trans-items}\left(x_{kr}, x_{ij}\right) = \frac{1}{\sqrt{\left(1/s_{trans}\left(X_k, X_i\right)\right)^2 + \left(1/s_{items}\left(Y_r, Y_j\right)\right)^2}} \quad (8)$$

Отже, в розробленій рекомендаційній системі використовується саме гібридний спосіб фільтрації, який є універсальним.

Для зменшення складності обчислення ступеня «схожості» векторів об'єктів або транзакцій в системі використовується підхід пониження розмірності матриці M_{kr}, який базується на розкладанні цієї матриці по сингулярним значенням (SVD – Singular Value Decomposition) [2].

На рис. 1 представлено алгоритм роботи системи з урахуванням вище викладеного підходу.

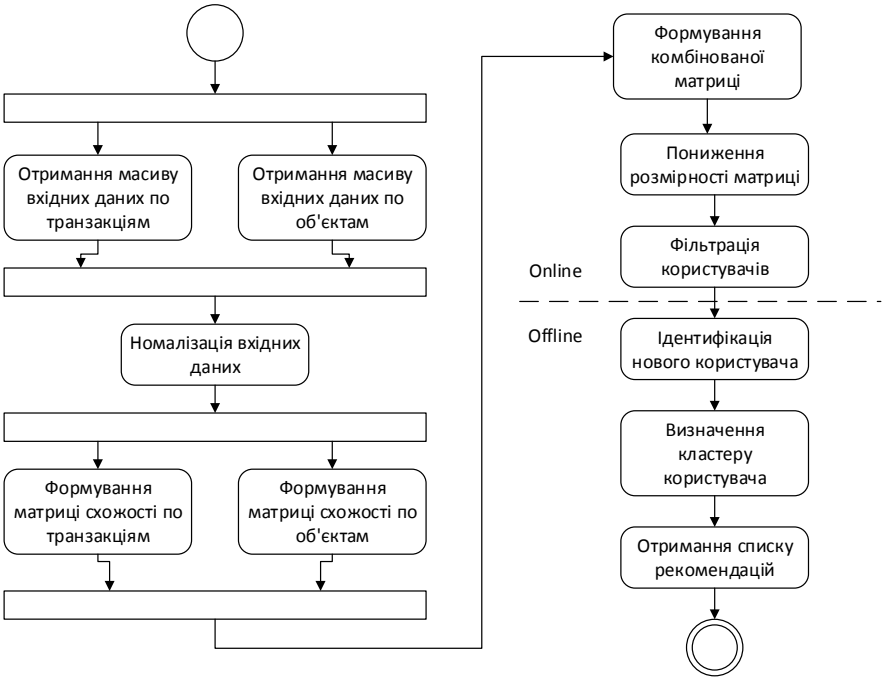

Рис. 1. Формування рекомендацій

Розроблена система дає можливість вдосконалити роботу туристичних агентств, завдяки застосуванню системного підходу для формалізації процесу роботи з клієнтами, що досить актуально в сучасних умовах.

Список літератури

1. Goldberg D. Using collaborative filtering to weave an information Tapestry / D. Goldberg, D. Nichols, B. M. Oki, D. Terry // Communications of the ACM. –1992. – № 35 (12). – P. 61–71.

2. Zhang W. Using Singular Value Decomposition approximation for collaborative filtering / W. Zhang, J. Wang, F. Ford Makedon, J. Pearlman // CEC 2005 : Seventh IEEE International Conference on E-Commerce Technology, 19-22 July, 2005. – Munich, Germany, 2005. – P. 257-264.