УДК 519.683.5

Рейда Олександр Миколайович,
к.т.н., доцент кафедри програмного забезпечення
Вінницького національного технічного університету, Україна

Боднар Роман Валентинович*, студент групи ПЗ 15-мі*
факультету інформаційних технологій та комп'ютерної інженерії
Вінницького національного технічного університету, Україна

МЕТОДИ ВИЯВЛЕННЯ ПОДІБНОСТІ МІЖ ОБ'ЄКТАМИ

Розглянуто методи і алгоритми, що використовуються при розробці рекомендаційних систем на основі колаборативної фільтрації .

Ключові слова: рекомендаційна система, колаборативна фільтрація, подібність між множинами.

There are considered methods and algorithms which are used in development recommender systems based on collaborative filtering.

Keywords: recommender system, collaborative filtering, iterative algorithm, similarity between sets.

Вступ. В сучасних умовах застосування нових альтернативних підходів до формування і прийняття високоякісних рішень в різних сферах суспільного життя неможливе без використання інформаційних систем. Причинами впровадження інформаційних технологій є зростаюча роль таких технологій практично в кожній галузі діяльності суспільства та зростаюча потужність таких технологій. Застосування сучасних підходів на основі інформаційних технологій дає можливість використовувати обчислювальні потужності комп'ютерів для виконання розрахунків, обробки, аналізу і прогнозування даних в режимі реального часу, для допомоги у прийнятті рішень. Серед сучасних напрямів розробки людино-машинних систем є рекомендаційні системи, експертні системи та системи підтримки прийняття рішень

Метою роботи є огляд методів, що використовуються при розробці рекомендаційних систем на основі колаборативної фільтрації.

Головною задачею є визначення оптимальності і точності роботи алгоритмів, що використовуються при розробці рекомендаційних систем на основі колаборативної фільтрації.

Об’єктом дослідження є алгоритми для визначення подібності між множинами об’єктів.

Предметом дослідження є методи і засоби проведення дослідження визначення подібності між множинами об’єктів.

Визначення подібності між об’єктами. Для визначення подібності між користувачами чи об’єктами можна використовувати такі підходи:

- відстань Евкліда, Хеммінга;
- кореляція Пірсона;
- коефіцієнт Жаккара;

Розрахунок евклідової відстані - один з найпростіших способів обчислення оцінки подібності. Об’єкти представляються у виді координатних осей. В кожній системі координат розташовуються точки, які відповідають користувачам. Близькість точок відображає схожість вподобань користувачів.

Чим ближче розташовуються два користувача на системі координат, тим більш схожі їх вподобання. Евклідова відстань між двома користувача знаходиться за формулою [1]:

$$w_{u,a} = \sqrt{\sum_{i \in I} (r_{a,i} - r_{u,i})^2} \quad (1)$$

де:

$w_{u,a}$ - міра схожості користувачіваі u,

I – множина об'єктів, які оцінені як користувачем a, так і користувачему;

$r_{u,i}$, $r_{a,i}$ -середня оцінка користувачів a і и відповідно.

Відстань, обчислена за цією формулою, відображає схожість між користувачами, а саме, чим менша відстань, тим більше користувачі схожі між собою. Для зручності обчислень, використовують модифіковану формулу,

значення за якої збільшується, коли користувачі більш схожі між собою [2].

$$w_{u,a} = \frac{1}{1 + \sqrt{\sum_{i \in I}(r_{a,i} - r_{u,i})^2}}, \quad (2)$$

де:

$w_{u,a}$ - міра схожості користувачіаі u,

I – множина об'єктів, які оцінені як користувачем a, так і користувачему;

$r_{u,i}$, $r_{a,i}$ -середня оцінка користувачів a і и відповідно.

Функція повертає значення від 0 до 1, якщо значення дорівнює 1, вподобання двох користувачів повністю співпадають.

Однак, якщо кластери добре нероздільні за однією ознакою і не роздільні по іншому, то після нормування дискримінування можливості першої ознаки будуть зменшені у зв'язку з посиленням «шумового ефекту» другого.

Колаборативна фільтрація на основі подібності користувачів має високу точність. Проте, недоліком даного підходу є висока ресурсомісткість (вимога до пам’яті) і складність (кількість обчислень, потрібне для отримання рекомендації). До того ж обчислення ступеня схожості може вираховуватись тільки в реальному часі, оскільки дані про поточну транзакцію є доступними тільки в момент обчислення рекомендації. Тому даний підхід може застосовуватись до відносно невеликих баз даних.

Коефіцієнт кореляції - це міра того, на скільки добре два набора даних лягають на пряму.

$$x^m = (x_1, \cdots, x_m),\ y^m = (y_1, \cdots, y_m); \quad (3)$$

де: x^m - перша вибірка даних,

y^m - друга вибірка даних

Коефіцієнт кореляції Пірсона розраховується за наступною формулою:

$$r_{xy} = \frac{\sum_{i=1}^{m}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{m}(x_i - \bar{x})^2 \sum_{i=1}^{m}(y_i - \bar{y})^2}} = \frac{cov(x,y)}{\sqrt{s_x^2 s_y^2}}, \quad (4)$$

де:

$\bar{x},\bar{y}$ – вибірково середнє x^m та y^m,

s_x^2, s_y^2 – вибіркові дисперсії,

$r_{xy} \in [-1,1]$.

Коефіцієнт кореляції Пірсонатакож називають тіснотою лінійного зв'язку:

- $|r_{xy}| = 1 \Rightarrow x,y$ - лінійно залежні
- $r_{xy} = 0 \Rightarrow x,y$ - лінійно незалежні

Для візуалізації такого методу, зобразимо на діаграмі оцінки, виставлені двома критиками. На діаграмі зображена пряма лінія - це лінія найкращого приближення, яка проходить на стільки близько до всіх точок координатної площини, на скільки це можливо. Якщо два критика виставляють всім об’єктам однакові оцінки, то лінія найкращого приближення буде діагональна і проходити через всі точки. В такому випадку отримаємо ідеальну кореляцію з коефіцієнтом 1 [2].

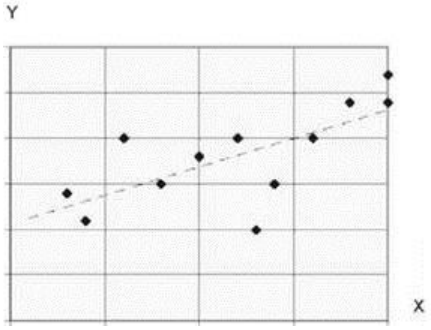

Рисунок 1 - Діаграма оцінок виставлених критиками

Передбачення рейтинга (оцінки) користувачавизначається за формулою:

$$\hat{r}_{u,i} = \frac{\sum_{j \in S} r_{u,i} * w_{i,j}}{\sum_{j \in S} |w_{i,j}|}, \quad (5)$$

де:

$r_{u,i}$ – оцінка об'єкта I користувачами u і a відповідно

S – множина користувачів найбільш близьких до множини I

Даний алгоритм відображає теоретичну базу метода, але на практиці ряд факторів потребує корегування розрахунків. Як правило, більшість оцінок невідомо, і розрідженість матриці оцінок достатньо висока. З іншого боку, дані, представлені в матриці оцінок є достатньо суб'єктивними. Деякі користувачі - оптимісти, і їх оцінки постійно високі, інші користувачі - песимісти, їх оцінки завжди занижені. Крім того, існують об'єкти, які подобаються усім.

Істотними недоліками можна вважати коефіцієнту кореляції Пірсона:

- Нестійкість до викидів.
- За допомогою коефіцієнта кореляції Пірсона можна визначити силу лінійної залежності між величинами, інші види взаємозв'язків виявляються методами регресійного аналізу.
- Необхідно розуміти різницю понять "незалежність" і "некоррельованність". З першого слідують другі, але не навпаки.
- Для того, щоб з'ясувати ставлення між двома змінними, часто необхідно позбутися від впливу третьої змінної. Розглянемо приклад 3-х змінних.

Міра Жаккара — бінарна міра подібності, запропонована Полем Жаккаром в 1901 році.[3] Запропонований метод здобув поширення і нині використовується для оцінки подібності скінченних множин, в інформатиці, для пошуку подібних документів, плагіату, тощо.

(Класичний)Коефіцієнт подібності Жаккара обчислюють за формулою:

$$K_J = \frac{c}{a+b-c} \tag{6}$$

де:

a — кількість видів на першому пробному майданчику,

b — кількість видів на другому пробному майданчику,

c — кількість видів, спільних для 1-ого та 2-ого майданчиків.

Коефіцієнт Жаккара двох множин X та Y дорівнює відношенню кількості елементів перетину множин до кількості елементів їхнього об'єднання.

В інформатиці для випадків порівняння двох множин за спільними характеристиками наприклад подібності фільмів, використовують таку форму запису:

$$sim(x,y) = \frac{|X \cap Y|}{|X \cup Y|} \tag{7}$$

Проте, виникає досить вагома проблема, якщо у вибірці присутній об’єкт з досить великою sim (x, y) серед респондентів. Система буде рекомендувати даний об’єкт усім респондентам, тим самим не демонструючи, інші подібні варіанти, що мають меншу «популярність» проте більше відповідають подібності.

Наприклад, розглянемо зазначену вище формулу для випадку, коли об'єкт у дуже популярний (наприклад, книга «Я, Робот»). Так як книга дуже популярна і її прочитало багато людей, то sim (x, y) буде наближатися до 1 майже для всіх фільмів x. Це означає, що книга у буде схожа на всі книги, а це в більшості випадків погано. Навряд чи книга «Я, Робот» буде схожою на книгу «Злочин і кара».

Таким чином для нормальної роботи слід модифікувати формулу:

$$sim(x,y) = \frac{\frac{|X \cap Y|}{|X|}}{\frac{|!X \cap Y|}{|!X|}} \tag{8}$$

де !X - це множина користувачів, які не переглянули фільм x [4].

Якщо y - дуже популярний об'єкт, то знаменник у формулі буде великим. Тоді значення схожості буде менше, а рекомендації будуть більш релевантними.

Висновок. Проаналізувавши алгоритми для виявлення подібності між об’єктами: кореляція Пірсона, Евклідова відстань і коефіцієнт Жаккара, було виявлено що дані методи не є оптимальними для застосування в рекомендаційних системах, як незалежні методи, і матимуть результатами лише за умови застосування гібридних систем з участю даних методів.

Література

1. Відстань між об'єктами (кластерами) і міра близькості [Електронний ресурс] //Режим доступу до матеріалу: http://bibliograph.com.ua/economicheskaya-statistika-2/11.htm
2. Toby Segaran. Programming Collective Intelligence. − Sebastopol: O’Reilly Media, Inc., 2007. − 360 c.
3. Коефіцієнт Жаккара [Електронний ресурс] //Режим доступу до матеріалу: https://uk.wikipedia.org/wiki/Коефіцієнт_Жаккара
4. Неперсонализированные рекомендации: метод ассоциаций [Електронний ресурс] //Режим доступу до матеріалу: https://habrahabr.ru/company/ivi/blog/247813/