УДК 681.3.01

К.т.н., доцент Соколова Н.А., студентка Родіонова Ю.С.

Національний технічний університет України
«Київський політехнічний інститут»

ОПТИМІЗАЦІЯ ПОШУКОВОЇ СИСТЕМИ ЕЛЕКТРОННОЇ НАУКОВОЇ БІБЛІОТЕКИ ШЛЯХОМ ВПРОВАДЖЕННЯ РЕКОМЕНДАЦІЙНОЇ СИСТЕМИ НОВОГО ПОКОЛІННЯ

Abstract

Nadiya A. Sokolova, assoc. prof., PhD; Yuliya Rodionova, student
Search engine optimization of electronic scientific library by introducing new generation recommended system

This paper proposes a novel, unified approach that integrates two available filtering paradigms to learn a usefulness function that gives user preference prediction. The key ingredient of hybrid method is the design of similarity function between user-item pairs that allows simultaneous generalization across the user and item dimensions in electronic scientific library. The comparative analysis of efficiency of both the classical and the modified algorithms is fulfilled. The ways for further research are proposed as well.

Вступ

Сучасні рекомендаційні системи (РС) застосовують різноманітні методи дослідження знань для вирішення проблеми надання персональних рекомендацій користувачам щодо інформації, продукції чи сервісів у режимі on-line. У зв'язку з стрімким ростом об'єму доступної інформації і числа користувачів у Web-просторі перед РС постають такі задачі: надання найбільш валідних рекомендацій, забезпечення масштабованості систем незалежно від кількості їх користувачів та об'єктів, досягнення високого рівня охоплення цільової аудиторії в умовах розрідженості даних [1].

Існує багато методів розв'язання цих задач: коллаборативна фільтрація, контентна фільтрація, гібридна – що вбирає в себе ідеї двох вищеописаних. В даній статті пропонується варіація гібридного методу, яка підвищує степінь відповідності наданих рекомендацій і пошукових запитів користувачів.

Постановка задачі

Задача полягає в модифікації існуючих методів фільтрації РС та їх комбінуванні в один таким чином, щоб функція корисності для надання рекомендацій щодо об'єктів включала в якості параметрів імпліцитні данні системи – тобто рейтинги, та експліцитні – деякі соціальні параметри, тобто гарантувала більшу степінь відповідності рекомендованих об'єктів і пошукових запитів користувачів.

Огляд існуючих рекомендаційних систем

Формально проблема надання рекомендацій може бути представлена наступним чином: нехай *C* – група всіх користувачів, *S* – множина об'єктів, що пропонуються до вибору. Нехай *u* – функція корисності, що описує корисність предмета *S* для *C*, тобто u: $C \times S \rightarrow R$, де *R* – кількість обраних об'єктів. Тоді для кожного користувача $c \in C$, ми хочемо вибрати такий об'єкт:

$$s_c' = \arg\max_{s \in S} \mathbf{u}(\mathbf{c}, \mathbf{s})$$

Функція корисності може бути визначена користувачем (в системах, що засновані на користувацьких оцінках) або програмою (що враховує прибуток). Загалом, корисність *u* не визначена для всієї сукупності $C \times S$, тому рекомендаційний механізм повинен вміти передбачати оцінки для ще неоціненних об'єктів за допомогою машинного самонавчання, евристичних методів або теорії апроксимації [2].

В контентних РС висновок про корисність об'єкта $u(c, s)$ робиться на основі раніше присвоєних оцінок, що надавалася користувачем об'єктам, схожим з *s*. Тобто маємо контент профілю користувача *c* – що містить інформацію про його смаки і побажання, а також контент документу *s* – що може бути представлений як вектори прямої чи оберненої частотності ваги ключових слів. Тоді *u(s,c)* залежна від цих двох параметрів, і може бути обчислена за допомогою міри лінійної подібності – косинусу кута між векторами, які в якості координат містять дані з відповідних контентів:

$$u(s, c) = \text{бали}(\text{Контент профілю } c, \text{Контент } (s))$$

Головною проблемою такого виду фільтрації є неможливість отримати рекомендації на нові, зовсім незнайомі користувачеві об'єкти.

Коллаборативні РС надають рекомендації користувачу на об’єкти, що могли бути ним не помічені. Вони робляться на основі оцінок, що дали такому об’єкту схожі з ним користувачі. Для обчислення коефіцієнта подібності між користувачами використовуються різні підходи. В основному вони орієнтовані на те, які оцінки користувачі надали одним і тим самим об’єктам. Найбільше розповсюдження отримали кореляційний метод, в якому для обчислення подібності використовують коефіцієнт Пірсона, та метод лінійної подібності, де в якості такого показника обчислюється косинус кута між векторами [3]. Слід звернути увагу на те, що різні РС можуть використовувати різні підходи для якнайбільш ефективного обчислення подібності між користувачами й аналізу зроблених оцінок. Вважається, що найбільш ефективно спочатку вирахувати всі значення подібності між користувачами системи, а потім лише періодично їх перераховувати.

Невирішеними проблемами коллаборативної фільтрації є проблема нового користувача, проблема нового об’єкту, проблема розрідженості даних. Остання виникає через те, що в будь-якій РС кількість оцінок, які треба надати, в більшості випадків перевищує кількість даних оцінок [1].

Опис підходу до побудови гібридної рекомендаційної системи

Проаналізувавши існуючі методи побудови РС, було виявлено ряд недоліків, а також шляхи їх подолання у електронній бібліотеці наукових і учбових матеріалів. Створювана мною гібридна РС спирається на коллаборативну фільтрацію, але також використовує контентні профілі користувачів, що дозволяє об’днати їх переваги.

Використання контентної методики має обмеження щодо неможливості розрізнити два об’єкта, що описані ідентичним набором ключових слів. В електронній науковій бібліотеці я його вирішила за рахунок структурованого опису об’єктів, за зразком набору метаданних Дублінського ядра (DC). В такому випадку в більшій мірі гарантуються різні значення в векторах контенту об’єкта за рахунок розширення його профілю.

Проблема нового користувача у методі коллаборативної фільтрації вирішена за рахунок анкетування користувачів. Тобто в модифікованому методі відтепер два користувачі вважатимуться схожими не лише за рахунок спільних смаків (що вираховуються згідно даних оцінок), але й за приналежністю до спільних груп, наприклад, факультет, кафедра, академічна група тощо.

Для мого методу також вирішена проблема розрідженості оцінок, оскільки в електронній наковій бібліотеці контентні користувацькі профілі

використовуються для встановлення близькості між користувачами. Ця методика використовується як перший крок для звуження списку найближчих сусідів (так звана k-кластеризація).

Емпірично було показано, що точність надання рекомендацій буде вищою, якщо присвоювати неоціненним об'єктам деяку оцінку за замовчанням. В гібридному методі я використовувала кореляційну методику не для встановлення схожості між користувачами, а для обчислення подібності між об'єктами та отримання гіпотетичних оцінок. Проте формула з використанням простої лінійної міри косинуса має один суттєвий недолік – в даному випадку не береться до уваги різниця в шкалі рейтингів між різними користувачами. Скоректований коефіцієнт подібності між об'єктами *i* та *j* обчислюється за формулою:

, де

$\overline{R_u}$ - середній рейтинг користувача *u*; $R_{u,i}, R_{u,j}$ – оцінки, надані користувачем *u* об'єктам *i* та *j* відповідно.

Висновки

В роботі було запропоновано ряд підходів щодо вдосконалення існуючих методів фільтрації РС та комбінування їх таким чином, щоб забезпечити найбільшу степінь відповідності знайдених об'єктів та персональних запитів користувачів. Однією з оцінок якості наданих рекомендації слугує величина MAE (Mean Absolut Error) – відношення рейтингу, даного користувачем до рейтингу, передбаченого системою. За проведеними аналітичними підрахунками для створеного гібридного методу у порівнянні з класичними вона є меншою близько на 11%, що є значним виграшем.

В подальших дослідженнях найактуальнішим є питання про визначення критеріїв оцінювання валідності наданих рекомендацій, в результаті аналізу яких можливе майбутнє удосконалення створених РС.

Література

1. *Карауш А.С.* Распределенные технологии использования библиотечных систем в МИБС г. Томска / А.С. Карауш, А.С.

Макаревич // Информационный бюллетень РБА. -СПб.: Издательство Российской национальной библиотеки. - 2006. - N40. - С. 112-114.
2. *Gediminas Adomavicius, Alexander Tuzhilin* IEEE Transactions on Knowledge and Data Engineering, Vol. 17, No. 6, June 2005.
3. *Balabanovic M., & Shoham, Y.* (1997). Fab: Content-based, collaborative recommendation. Communications of the ACM, 40, 6-72.