УДК 004.021

А.Й. Савицький, Д.В. Попович

МЕТОДИ КОЛАБОРАТИВНОЇ ФІЛЬТРАЦІЇ ДЛЯ НЕПРЯМИХ РЕЙТИНГІВ

Анотація: Робота присвячена методам колаборативної фільтрації на основі сусідства. Визначені їх ключеві етапи та варіанти модифікації для використання з непрямими рейтингами. Проведений порівняльний аналіз для фільтрації за користувачами та об'єктами. Дослідження проводились з даними для більш ніж 1 мільйона користувачів, які були опубліковані в рамках конкурсу Million Data Song Challenge.

Ключові слова: рекомендаційна система, колаборативна фільтрація, Million Data Song Challenge.

Вступ

Стрімке зростання кількості інформації та сервісів доступних в інтернеті, ускладнює процес вибору та підвищує ймовірність прийняття помилкового рішення. Для подолання проблеми надлишкової інформації застосовують рекомендаційні системи.

Рекомендаційні системи – це програмні засоби, які намагаються передбачити, які об'єкти будуть цікаві користувачу, якщо є апріорна інформація про його вподобання [1].

Об'єкт — це загальний термін, який використовується для позначення того, що система рекомендує користувачу. Зазвичай рекомендаційна система розробляється для конкретного типу об'єктів (наприклад фільмів чи новин) і, відповідно, її архітектура, підходи до рекомендацій та навіть графічний інтерфейс спрямовані на забезпечення ефективної роботи саме для цього типу об'єктів [1].

Рекомендація — це об'єкт або група об'єктів, які система рекомендує користувачу. Рекомендаційні системи спрямовані на осіб, які не мають достатнього досвіду або компетенції для оцінки конкретного об'єкту, серед різноманіття вибору альтернатив, які може запропонувати сервіс [1].

Більшість створених комерційних системи, в своїй основі, використовують методи колаборативної фільтрації. Вони базуються на припущені, що схожі користувачі мають схожі вподобання, а схожі об'єкти використовуються користувачами спільно. При цьому, схожість об'єктів та користувачів визначається не за їх змістом, а за історією їх взаємодій. Така історія може бути подана у вигляді матриці, стовпці якої відповідатимуть об'єктам, а рядки — користувачам. Значення, в свою чергу, це певна кількісна міра (оцінка) рівня інтересу користувача до об'єкта. Оцінки поділяють на прямі (оцінка об'єкта користувачем за певною шкалою) та непрямі (різноманітні дані про поведінку користувача: кількість переглядів, час взаємодії з об'єктом і т.д.) [2].

Прямі оцінки більш інформативні, оскільки виражають безпосереднє ставлення користувача до об'єкта. Однак, це вимагає активної участі користувачів у процесі збору інформації, що в рамках реальних інформаційних та комерційних сервісів велика рідкість. Це призводить до проблеми розрідженості матриці оцінок, що впливає на точність колаборативних методів. Альтернативним підходом є використання непрямих оцінок, які можна отримати фіксуючи поведінку користувачів. В той же час, непрямі оцінки мають ряд особливостей (відсутність негативних оцінок, наявність шуму, іншу семантику числових значень) [2], які не дозволяють аналізувати їх методами, що розроблялися для прямих рейтингів. В рамках даної роботи ми спробуємо розглянути та порівняти методи колаборативної фільтрації, які придатні для аналізу непрямих оцінок.

Рекомендаційні системи є одним з важливих розділів інтелектуального аналізу даних — Data Mining. Підтвердженням актуальності досліджень рекомендаційних систем та методів побудови рекомендацій, є їх активне використання всесвітньовідомим комерційними проектами, до яких входять: Amazon, YouTube, LastFm, Pandora, Google Adsense та інші [3].

Постановка задачі

Методи колаборативної фільтрації можна поділити на три групи: методи на основі сусідства (neighborhood-based), методи на основі моделі (model-based) та гібридні методи. В свою чергу, у методах на основі сусідства виділяють два підходи: фільтрація за користувачами (user-based) та фільтрація за об'єктами (item-based) [1]. В рамках даної роботи, ми зосередимось на методах на основі сусідства.

Для тестування обраних нами алгоритмів, скористаємось експериментом на статистичних даних (of fline experiment). Він дозволяє порівняти роботу різних методів на вибірці, яка містять інформацію про поведінку цільової аудиторії за певний період часу. Приховуючи частину даних матриці оцінок, вибірку розділяють на тренувальну та контрольну. Задача — маючи дані з тренувальної вибірки, передбачити дані контрольної вибірки.

Для проведення досліджень використаємо дані, що були підготовані в рамках конкурсу Million Song Dataset Challenge [3]. Учасникам доступна повна історія прослуховувань пісень для 1 мільйона користувачів та половина історії для 110 тисяч користувачів. Алгоритми учасників перевірялись за допомогою прихованої половини. Характеристики даних наведені у таблиці 1.

Більшість алгоритмів, які працюють з прямими оцінками, ставлять перед собою задачу мінімізувати середньоквадратичну похибку у прогнозуванні оцінок прихованої частини. При роботі з непрямими оцінками, які не мають чіткого діапазону значень, задача ал-

горитмів зводиться до прогнозування факту взаємодії користувача з об'єктом.

Таблиця 1

Характеристика вхідних даних

Кількікість об'єктів	384 546
Кількість користувачів	1 019 318
Кількість непрямих оцінок	48 373 586
Заповненість матриці оцінок	0.01%

Результати роботи рекомендаційного алгоритму зазвичай подають у вигляді впорядкованого списку, тому важливо оцінити якість такого ранжування. Виходячи з особливостей роботи з непрямими рейтингами, ми обрали в якості головного критерію оцінки алгоритмів — середнє значення середньої точності.

Середнє значення середньої точності (Mean Average Precision) — основний показник точності роботи рекомендаційного алгоритму. Активно використовується в теорії інформаційного пошуку. Ця характеристика одночасно поєднує в собі оцінку точності прогнозування взаємодії та ранжування.

Нехай, $M \in \{0,1\}$ це факт взаємодії користувача u з об'єктом i, а $y(j) = i$ рекомендаційни алгоритм, який визначає, що об'єкт i був рекомендований у позиції j. Запишемо оцінку точності рекомендації списку з k об'єктів для користувача u алгоритмом y:

$$P_k(u, y) = \frac{1}{k} \sum_{j=1}^{k} M_{u,y(j)} \tag{1}$$

Тоді середню точність рекомендацій для користувача u алгоритмом y на списку довжиною τ можна записати як:

$$AP(u, y) = \frac{1}{n_u} \sum_{k=1}^{\tau} P_k(u) M_{u,y(k)}, \tag{2}$$

де n_u — кількість об'єктів з якими насправді взаємодіяв користувач u. І, нарешті, усереднимо це значення для всіх користувачів:

$$MAP = \frac{1}{m} \sum_{u} AP(u, y_u), \tag{3}$$

де m — кількість користувачів, для яких надаються рекомендації.

Результати досліджень

Методи колаборативної фільтрації на основі сусідства складаються з трьох основних пунктів, які є спільними, як для фільтрації за користувачами, так і для фільтрації за об'єктами. А саме:

1. Нормалізація матриці оцінок. Одні користувачі схильні давати вищі рейтинги ніж інші, тому необхідно врахувати особливості їх поведінки. Для матриць з прямими рейтингами запропоновано багато методів, найбільш популярні: mean-centering та z-score [1]. Для нашого випадку, з непрямими оцінками, їх використання не принесло бажаного результату, тому ми зупинились на звичайній бінаризації.

2. Формування міри близькості для об'єктів/користувачів. Виходячи з основного припущення методів колаборативної фільтрації про схожість інтересів для схожих користувачів, необхідно представити "схожість" у чисельному вигляді. Формально це пошук відстані між векторами оцінок об'єктів/користувачів. Найбільш поширені міри близькості у методах колаборативної фільтрації: косинусна (cosine) та кореляція Пірсона (Pearson Correlation) [4]. Виконавши бінаризацію, ми не можемо використовувати кореляцію Пірсона, оскільки дисперсія значень буде нульовою. Тому ми зупинились на косинусній мірі. Відстань між векторами оцінок користувачів u та v, можна записати як:

$$d\left(\overrightarrow{r_u}, \overrightarrow{r_v}\right) = \frac{\overrightarrow{r_u} \cdot \overrightarrow{r_v}}{|\overrightarrow{r_u}| \cdot |\overrightarrow{r_v}|} \tag{4}$$

Аналогічно можна визначити відстань між векторами оцінок об'єктів.

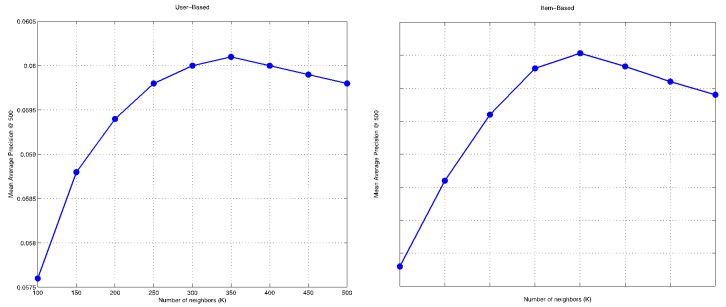
3. Прогнозування оцінки. Отримавши інформацію про те наскільки користувачі/об'єкти схожі між собою, можна заповнити пропущенні значення вектору оцінок. Для прямих рейтингів застосовують зважену суму оцінок сусідів [4]. У випадку з бінаризованими непрямими рейтингами, це зводиться до суми знайдених відстаней:

$$\hat{r}_{ui} = \sum\nolimits_{v \in N} d\left(\overrightarrow{r_u}, \overrightarrow{r_v}\right) \cdot r_{vi} = \sum\nolimits_{v \in N} d\left(\overrightarrow{r_u}, \overrightarrow{r_v}\right), \tag{5}$$

де N — множина користувачів.

Однак, на практиці, при великій розмірності матриці оцінок (в нашому випадку $1.2 \cdot 10^6 \times 3.8 \cdot 10^5$), множину користувачів N обмежують k найближчими сусідами, щоб зменшити обчислювальну складність алгоритму [1]. Вплив параметру k, для обраних вхідних даних, на середнє значення середньої точності методів колаборативної фільтрації за користувачами та об'єктами зображено на рисунку 1.

Доцільність використання методів колаборативної фільтрації для непрямих оцінок можна побачити у порівнянні їх результатів з найпростішим рекомендаційним алгоритмом — списком найбільш популярних об'єктів (Таблиця 2).



Таблиця 2

Порівняння точності досліджуваних алгоритмів

Рекомендаційний алгоритм	Максимальна MAP
За популярністю об'єктів	0.0223
Колаборативна фільтрація за користувачами	0.0602
Колаборативна фільтрація за об'єктами	0.1453

Висновки

Непрямі оцінки накладають певні обмеження на використання методів колаборативної фільтрації. В рамках даної роботи, ми спробували визначити ключові етапи методів на основі сусідства та адаптувати їх для обробки непрямих рейтингів. Прийняття рішень, на кожному з етапів цих алгоритмів (нормалізація, вибір міри близькості, прогнозування оцінки) суттєво впливає на результат їх роботи. Тому, важливо проводити їх адаптацію для кожної конкретної прикладної задачі.

Методи колаборативної фільтрації на основі сусідства відносно прості у реалізації, однак, для побудови списку рекомендованих об'єктів, вимагають обробки всіє матриці оцінок. Зважаючи на те, що більшість комерційних проектів надають рекомендації користувачам у реальному часі, це може викликати значні складнощі у побудові програмно-технічного забезпечення для такої системи. Однак, варто зазначити, що існують певні модифікації цих алгоритмів, які зменшують вплив великої розмірності матриці оцінок, на продуктивність рекомендаційної системи [5].

Список використаних джерел

1. Recommender Systems Handbook / Francesco Ricci, Lior Rokach, Bracha Shapira, Paul B. Kantor. — NY, UAS : Springer Science+Business Media, 2011. — 845 p.
2. Yifan Hu. Collaborative Filtering for Implicit Feedback Datasets / Yifan Hu, Yehuda Koren, Chris Volinsky // Proceedings of

the 2008 Eighth IEEE International Conference on Data Mining (ICDM '08), (Pisa, Italy, December 15-19, 2008). — IEEE, 2008. — P. 263-272.

3. Brian McFee. The Million Song Dataset / Thierry Bertin-Mahieux, Daniel P.W. Ellis, Brian Whitman, Paul Lamere // Proceedings of the 21st international conference companion on World Wide Web (WWW '12 Companion), (Lyon, France, April 16-20, 2012). — ACM, 2011. — P. 909-916.
4. Recommender Systems An Introduction / Dietmar Jannach, Markus Zanker, Alexander Felfernig, Gerhard Friedrich. — NY, UAS : Cambridge University Press, 2011. — 335 p.
5. Sebastian Schelter. Scalable Similarity-Based Neighborhood Methods with MapReduce / Sebastian Schelter, Christoph Boden, Volker Markl // Proceedings of the sixth ACM conference on Recommender systems (RecSys '12), (Dublin, Irland, September 9-13, 2012). — ACM, 2012. — P. 163-170.

Отримано 12.09.2014 р.