Menggabungkan Penyaringan Kolaboratif Berbasis Pengguna dan Berbasis Item Menggunakan Pembelajaran Mesin

Priyank Thakkar, Krunal Varma, Vijay Ukani, Sapan Mankad dan Sudeep Tanwar

Abstrak **Pemfilteran** kolaboratif (CF) biasanya digunakan untuk merekomendasikan item-item yang disukai oleh pengguna yang sama di masa lalu. Penyaringan kolaboratif berbasis pengguna (UbCF) dan penyaringan kolaboratif berbasis item (IbCF) adalah dua jenis CF dengan tujuan yang sama yaitu memperkirakan peringkat pengguna target untuk item target. Makalah ini mengeksplorasi berbagai cara untuk menggabungkan prediksi dari UbCF dan IbCF dengan tujuan untuk meminimalkan kesalahan prediksi secara keseluruhan. Dalam makalah ini, kami mengusulkan sebuah pendekatan untuk menggabungkan prediksi dari UbCF dan IbCF melalui regresi linier berganda (MLR) dan regresi vektor pendukung (SVR). Hasil dari pendekatan yang diusulkan dibandingkan dengan hasil dari pendekatan fusi lainnya. Perbandingan tersebut menunjukkan keunggulan pendekatan yang diusulkan. Semua pengujian dilakukan pada dataset besar yang tersedia untuk umum.

Kata kunci Pemfilteran kolaboratif berbasis pengguna - Pemfilteran kolaboratif berbasis item Pembelajaran mesin - Regresi linier berganda - Regresi vektor pendukung

P. Thakkar (✉) - K. Varma - V. Ukani - S. Mankad - S. Tanw ar
Institute of Technology, Nirma University, Ahmedabad 382481, India
e-mail: priyank.thakkar@nirmauni.ac.in

K. Varma
e-mail: krunalvarma@nirmauni.ac.in

V. Ukani
e-mail: vijay.ukani@nirmauni.ac.in

S. Mankad
e-mail: sapanmankad@nirmauni.ac.in

S. Tanwar
e-mail: sudeep.tanwar@nirmauni.ac.in

ger Nature Singapore Pte Ltd. 2019
S. C. Satapathy dan A. Joshi (eds.), *Teknologi Informasi dan Komunikasi untuk Sistem Cerdas*, Inovasi Cerdas, Sistem dan Teknologi 107,
https://doi.org/10.1007/978-981-13-1747-7_17

1 Pendahuluan

Sistem rekomendasi menjadi semakin penting karena meningkatnya penggunaan web untuk bisnis dan transaksi e-commerce [2]. Sistem rekomendasi film [7, 8], sistem rekomendasi buku [9], sistem rekomendasi tag [13], sistem rekomendasi teman Facebook adalah beberapa contoh sistem rekomendasi. CF mods adalah salah satu model dasar dari sistem rekomendasi yang dapat mengeksploitasi data interaksi pengguna dan barang seperti rating. UbCF dan IbCF adalah dua jenis penyaringan kolaboratif yang banyak digunakan baik di industri maupun akademis untuk mengatasi masalah kelebihan informasi.

Langkah pertama dalam UbCF adalah menemukan sekumpulan pengguna yang paling mirip dengan pengguna target. Peringkat pengguna target untuk item target kemudian diprediksi dengan menggunakan peringkat yang diberikan kepada item target oleh tetangga/pengguna yang paling mirip. Di sisi lain, tetangga terdekat di IbCF adalah item yang paling mirip dengan item target. Peringkat yang diberikan oleh pengguna target kepada item-item yang paling mirip ini kemudian digunakan untuk menghitung peringkatnya untuk item target. Jumlah tetangga terdekat adalah parameter desain dan harus disetel dengan benar.

Makalah ini berfokus pada penggabungan prediksi dari UbCF dan IbCF untuk mencapai prediksi akhir. Kontribusi baru dari makalah ini adalah penggunaan MLR dan SVR untuk menggabungkan prediksi dari UbCF dan IbCF. Hasil dari pendekatan yang kami usulkan dibandingkan dengan pendekatan fusi lainnya sebagai tambahan dari hasil UbCF dan IbCF ketika diimplementasikan secara individual.

Susunan makalah ini adalah sebagai berikut. Tinjauan literatur yang ada disajikan di Bagian 2, sedangkan Bagian. 3 membahas dasar-dasar penyaringan kolaboratif. Pendekatan fusi yang diusulkan dan hasil eksperimen dibahas di Bagian 4 dan 5. 4 dan 5 masing-masing. Makalah ini diakhiri dengan kata penutup di Bagian 6.

2 Pekerjaan Terkait

Ada beberapa upaya untuk menggabungkan prediksi dari sistem rekomendasi yang berbeda. Dalam [5], pemfilteran berbasis konten dan kolaboratif digabungkan dengan bantuan pendekatan hibrida. Sebuah pendekatan yang menggunakan dekomposisi nilai tunggal dan hibridisasi berbasis konten dan IbCF untuk merekomendasikan program di TV diusulkan dalam [3].

Salah satu upaya pertama untuk menggabungkan pendekatan UbCF dan IbCF dijelaskan dalam [12]. Pendekatan ini didasarkan pada fusi kemiripan dan kerangka kerja fusi bersifat probabilistik. Pendekatan yang dijelaskan dalam makalah ini terinspirasi dari pekerjaan yang dilakukan di [11] dan [6].

Thakkar dkk. [11] juga mencoba menggabungkan prediksi dari UbCF dan IbCF. Namun, pendekatan mereka sederhana dan mengandalkan rata-rata tertimbang dari prediksi untuk menghasilkan prediksi akhir. Mereka menemukan bobot untuk rata-rata melalui validasi silang lima kali lipat dari dataset pelatihan.

Para penulis di [6] telah menggabungkan prediksi dari UbCF dan IbCF menggunakan regresi bertumpuk. Sejauh yang kami ketahui, mereka menyelesaikan masalah regresi dengan menggabungkan prediksi dari UbCF dan IbCF sebagai masalah optimasi kuadrat terkendala. Kami telah mencoba menyelesaikan masalah tersebut dengan menggunakan pendekatan sederhana regresi linier berganda (tanpa kendala) selain pendekatan yang melibatkan regresi vektor pendukung. Dataset yang digunakan dan metodologi eksperimental yang digunakan juga berbeda.

3 Penyaringan Kolaboratif

UbCF dan IbCF dibahas dalam bagian ini. Jika kita mengasumsikan m pengguna dan n item, dimensi matriks rating pengguna-item X adalah $m \times n$. Elemen $x_{i,j} = r$ mengindikasikan bahwa pengguna ke-i telah memberikan rating r *pada* item ke-j, di mana $r = R$. $x_{i,j} = \varphi$ menggambarkan bahwa item *ke-j* belum diberi rating oleh pengguna ke-i. Baris dan kolom dalam X berhubungan dengan pengguna dan
profil item, masing-masing.

3.1 Penyaringan Kolaboratif Berbasis Pengguna (UbCF)

Seperti yang telah disebutkan sebelumnya, langkah pertama dalam UbCF adalah mencari tahu tetangga terdekat pengguna target. Hal ini dapat dicapai dengan mencari kesamaan antara pengguna target dan semua pengguna lainnya. N pengguna yang paling mirip kemudian dapat dipilih untuk membentuk satu set N tetangga terdekat. Ada beberapa cara untuk menemukan kemiripan antara pengguna dan korelasi Pearson adalah salah satu metode yang digunakan dalam makalah ini. Korelasi Pearson antara pengguna u_1 dan u_2 seperti yang dibahas dalam [1, 11] adalah:

$$sim(u_1, u_2) = \frac{\sum_{i \in I_{u1\,u2}} (x_{u1,i} - \bar{x}_{u1})(\bar{x}_{u2,i} - \bar{x}_{u2})}{\sqrt{\sum_{i \in I_{u1\,u2}} (x_{u1,i} - \bar{x}_{u1})^2 \sum_{i \in I_{u1\,u2}} (x_{u2,i} - \bar{x}_{u2})^2}} \quad (1)$$

Di sini, $I_{u1\,u2}$ digunakan untuk menunjuk satu set item yang dikorelasikan oleh u_1 dan u_2 . $\bar{x}_{u1}$ menunjukkan peringkat rata-rata pengguna u_1 .

Ada beberapa cara yang dapat digunakan untuk menentukan peringkat pengguna i untuk item j.

Dalam makalah ini, kami telah menggunakan Eq. 2 untuk menyelesaikan tugas ini [1, 11].

$$x_{i,j} = \bar{x}_i + \frac{\sum_{u \in \hat{U}} sim(i, u^r) \times (x_{u^r,j} - \bar{x}_{u^r})}{\sum_{u \in \hat{U}} |sim(i, u^r)|} \quad (2)$$

Di sini, $\hat{U}$ menyatakan himpunan N tetangga/pengguna terdekat dari pengguna i

yang telah memberi nilai pada item
j .

3.2 Penyaringan Kolaboratif Berbasis Item (IbCF)

Dalam IbCF, item-item yang memiliki profil yang mirip dengan item target dianggap sebagai tetangga terdekat dari item target. Seperti pada UbCF, korelasi Pearson seperti yang disebutkan pada Eq. 3 [10, 11] dapat digunakan untuk menemukan kemiripan antar item.

$$sim(i_1, i_2) = \frac{\sum_{u \in U} (x_{u,i_1} - \bar{x}_{i_1})(x_{u,i_2} - \bar{x}_{i_2})}{\sum_{u \in U} (x_{u,i_1} - \bar{x}_{i_1})^2 \sum_{u \in U} (x_{u,i_2} - \bar{x}_{i_2})^2} \quad (3)$$

Di sini, U mewakili sekumpulan pengguna yang telah memberi peringkat pada i_1 dan i_2 . $\bar{x}_i$ menggambarkan peringkat rata-rata dari item i_1 . Ada beberapa cara untuk menghitung peringkat pengguna i untuk item j. Makalah ini menggunakan Persamaan 4 untuk menyelesaikan tugas tersebut [1, 11].

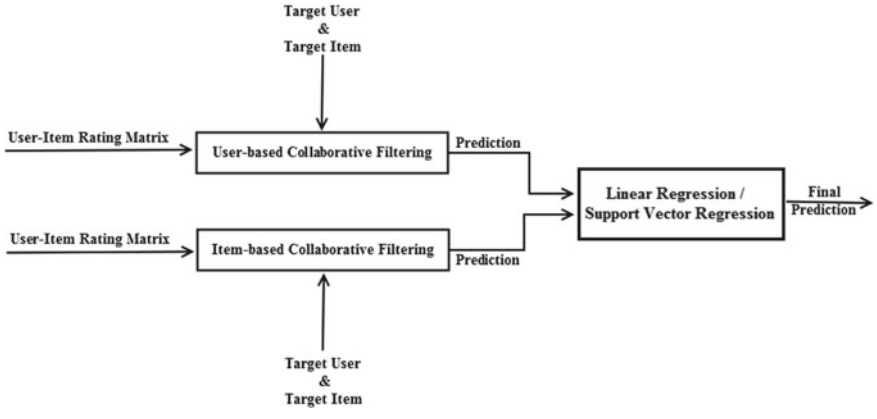
$$x_{i,j} = \bar{x}_j + \frac{\sum_{i' \in \hat{I}} sim(j, i') \times (x_{i,i'} - \bar{x}_{i'})}{\sum_{i' \in \hat{I}} |Sim(j, i')|} \quad (4)$$

Di sini, $\hat{I}$ merepresentasikan kumpulan N item yang paling mirip dengan item j dan telah dinilai oleh pengguna i.

4 Pendekatan yang Diusulkan

Makalah ini mengusulkan untuk menggabungkan prediksi dari UbCF dan IbCF melalui regresi linier berganda dan model regresi vektor pendukung. Pendekatan ini terinspirasi dari pekerjaan yang dilakukan di [6, 11]. Ide tersebut digambarkan pada Gambar 1.

Sangat penting untuk dicatat bahwa training set diperlukan untuk mempelajari model UbCF dan IbCF. Training set yang digunakan untuk mempelajari UbCF dan IbCF dilambangkan sebagai $Train_{CF}$



Gbr. 1 Pendekatan yang diusulkan

selanjutnya dalam makalah ini. Sangat mudah untuk memahami bahwa $Train_{CF}$ untuk UbCF dan IbCF masing-masing terdiri dari profil pengguna dan item.

Model regresi linier berganda dan regresi vektor pendukung juga membutuhkan training set untuk dilatih. Kami menamakan training set ini sebagai $Train_{ML}$ untuk selanjutnya dalam makalah ini. Jelaslah bahwa $Train_{ML}$ terdiri dari prediksi-prediksi dari UbCF dan IbCF sebagai data training. Prediksi-prediksi yang membentuk $Train_{ML}$ ini dibuat oleh UbCF dan IbCF melalui validasi silang lima kali lipat dari $Train_{CF}$.

Setelah $Train_{ML}$ terbentuk, ia digunakan untuk melatih model regresi linier berganda dan regresi vektor pendukung. Model-model yang telah dilatih ini kemudian digunakan untuk membuat prediksi akhir berdasarkan prediksi dari UbCF dan IbCF.

5 Evaluasi Eksperimental

Bagian ini dimulai dengan pembahasan mengenai dataset. Metodologi yang digunakan untuk berbagai eksperimen dan hasilnya juga dibahas.

5.1 Dataset

Semua percobaan dilakukan pada dataset Hetrec2011-movielens-2k [4], (http://www.imdb.com/, http://www.rottentomatoes.com/) yang dipublikasikan oleh sebuah kelompok penelitian yang dikenal sebagai GroupLens (http://www.grouplens.com/). Dataset tersebut dirangkum dalam Tabel 1. Matriks peringkat film pengguna dibuat dengan melakukan prapemrosesan terhadap dataset ini.

5.2 Langkah-langkah Evaluasi

Untuk mengevaluasi kinerja pendekatan UbCF, IbCF dan fusi, digunakan mean absolute error (MAE), mean absolute percentage error (MAPE) dan mean squared error (MSE) seperti yang telah didiskusikan dalam [11].

Tabel 1 Ringkasan dataset

Jumlah pengguna	2113
Jumlah film/item	10197
Jumlah peringkat	855,598
Rentang peringkat	0.5, 1.0, ..., 5.0
Jumlah rata-rata peringkat per pengguna	405
Jumlah rata-rata peringkat per film/item	85

5.3 Metodologi Eksperimental

Untuk semua percobaan, pengguna yang telah memberi peringkat antara 100 dan 120 film dipilih sebagai pengguna target. Terdapat 87 pengguna seperti itu dalam kumpulan data. Untuk setiap pengguna target, dipilih secara acak 25 item film sebagai item target. Hal ini memberikan kami satu set pengujian yang terdiri dari 87 pengguna dan 25 film. Prediksi dibuat untuk setiap pasangan pengguna-film dalam dataset uji. Dalam matriks peringkat pengguna-item yang sebenarnya, 87 × 25 peringkat set uji disamarkan untuk membangun $Train_{CF}$. Eksperimen yang dilakukan termasuk UbCF, IbCF
dan empat pendekatan fusi. Pendekatan fusi 1 menggunakan rata-rata sederhana sementara pendekatan fusi 2 menggunakan rata-rata tertimbang seperti yang dibahas dalam [11]. Pendekatan fusi 3 dan 4 adalah kontribusi baru dari makalah ini, dan mereka menggunakan regresi linier dan regresi vektor pendukung untuk menggabungkan prediksi UbCF dan IbCF seperti yang dibahas di Bagian. 4.

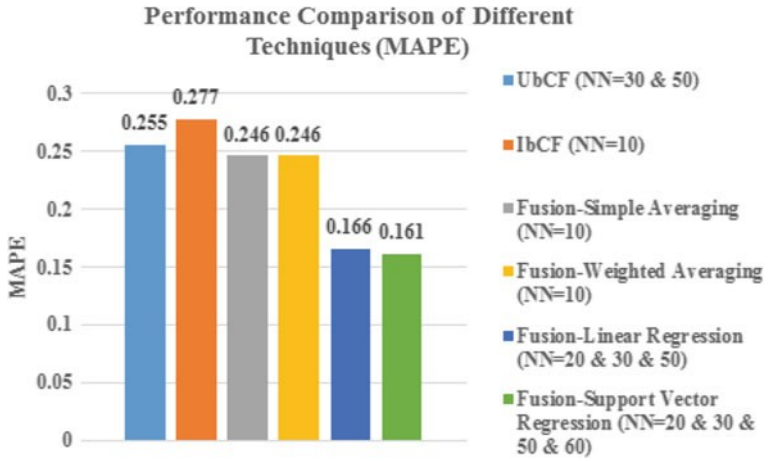
5.4 Hasil dan Pembahasan

Hasil dari berbagai teknik yang berbeda digambarkan pada Tabel 2. Untuk setiap teknik, percobaan dilakukan untuk 12 nilai tetangga terdekat (NN) yang berbeda. Terlihat bahwa hanya MAPE yang dilaporkan dalam hasil. Perlu disebutkan bahwa MAE dan MSE juga diukur tetapi belum dilaporkan karena keterbatasan ruang.

MAPE minimum yang dicapai dengan berbagai teknik dirangkum dalam Gbr. 2. MAPE minimum yang dicapai dengan fusi melalui pendekatan rata-rata tertimbang dan sederhana

Tabel 2 MAPE (nilai yang dilaporkan × 100%) dalam pendekatan UbCF, IbCF, dan fusi

Sr.	NN	UbCF	IbCF	Fusi menggunakan sederhana rata-rata [11]	Fusion menggunakan rata-rata tertimbang [11]	Fusi menggunakan regresi linier berganda	Penggabungan menggunakan regresi vektor pendukung
1	1	0.337	0.320	0.278	0.278	0.181	0.173
2	2	0.229	0.305	0.263	0.263	0.175	0.169
3	5	0.270	0.285	0.250	0.250	0.170	0.165
4	10	0.260	**0.277**	**0.246**	**0.246**	0.167	0.162
5	20	0.256	0.278	0.247	0.247	**0.166**	**0.161**
6	30	**0.255**	0.278	0.248	0.248	**0.166**	**0.161**
7	50	**0.255**	0.281	0.249	0.249	**0.166**	**0.161**
8	60	0.256	0.282	0.250	0.250	0.167	**0.161**
9	70	0.257	0.284	0.251	0.251	0.167	0.162
10	80	0.258	0.285	0.252	0.252	0.168	0.163
11	90	0.258	0.285	0.252	0.252	0.168	0.163
12	100	0.258	0.285	0.252	0.252	0.168	0.163



Gbr. 2 Perbandingan kinerja dari berbagai teknik yang berbeda

adalah 0,246. Ini jelas lebih baik jika dibandingkan dengan UbCF dan IbCF, dengan MAPE minimum masing-masing 0,255 dan 0,277.

Dapat dilihat bahwa penggabungan melalui regresi linier berganda dan regresi vektor pendukung telah meningkatkan MAPE menjadi 0,166 dan 0,161, yang secara signifikan lebih baik daripada pendekatan lainnya. Hal ini dikarenakan pada kedua pendekatan ini, bobot optimal yang diperlukan untuk prediksi rata-rata tertimbang dari UbCF dan IbCF dipelajari melalui LR dan SVR dan masalahnya ditangani sebagai masalah pembelajaran.

6 Kesimpulan

Makalah ini berfokus pada penggabungan prediksi dari UbCF dan IbCF dengan tujuan untuk meminimalisir kesalahan dalam prediksi. Eksperimen yang dilakukan meliputi UbCF, IbCF dan empat pendekatan fusi. Dua pendekatan fusi pertama mengandalkan rata-rata sederhana dan berbobot dari prediksi dari UbCF dan IbCF untuk menghasilkan prediksi akhir. Kontribusi utama dari makalah ini adalah sebuah pendekatan yang menggabungkan prediksi dari UbCF dan IbCF melalui regresi linier berganda dan regresi vektor pendukung. Keunggulan dari pendekatan ini terlihat jelas dari hasilnya. Dapat dilihat bahwa peningkatan kinerja sekitar 8% jika dibandingkan dengan penggabungan melalui simple dan weighted averaging. Peningkatan ini sekitar 9% dan 11% jika dibandingkan dengan UbCF dan IbCF. Peningkatan kinerja ini cukup menggembirakan dan menunjukkan arah ke depan di mana ketangguhan pendekatan yang diusulkan dapat divalidasi melalui pengujian pada dataset lain.

Referensi

1. Adomavicius, G., Tuzhilin, A.: Menuju sistem pemberi rekomendasi generasi berikutnya: sebuah survei tentang keadaan mutakhir dan kemungkinan perluasannya. IEEE Trans. Knowl. Data Eng. **17**(6), 734-749 (2005)
2. Aggarwal, C.C.: Sistem Pemberi Rekomendasi. Springer, Berlin (2016)
3. Barragáns-Martínez, A . B., Costa-Montenegro, E ., Burguillo, J . C., Rey-López, M., Mikic-Fonte, F.A., Peleteiro, A.: Pendekatan penyaringan kolaboratif berbasis konten dan berbasis item untuk merekomendasikan program TV yang ditingkatkan dengan dekomposisi nilai tunggal. Inf. Sains. **180 (**22), 4290-4311 (2010)
4. Cantador, I., Brusilovsky, P., Kuflik, T.: Lokakarya kedua tentang heterogenitas informasi dan fusi dalam sistem pemberi rekomendasi (hetrec2011). In: RecSys. pp. 387-388 (2011)
5. De Campos, LM, Fernández-Luna, JM, Huete, JF, Rueda-Morales, MA: Menggabungkan rekomendasi berbasis konten dan kolaboratif: pendekatan hibrida berdasarkan jaringan bayesian. Int. J. Perkiraan Alasan. **51**(7), 785-799 (2010)
6. Liu, Q., Xiong, Y., Huang, W.: Menggabungkan model berbasis pengguna dan berbasis item untuk pemfilteran kolaboratif menggunakan regresi bertumpuk. Chin. J. Elektron. **23**(4), 712-717 (2014)
7. Miller, B.N., Albert, I., Lam, S.K., Konstan, J.A., Riedl, J.: Movielens unplugged: pengalaman dengan sistem rekomendasi yang kadang-kadang terhubung. Dalam: Prosiding Konferensi Internasional ke-8 tentang Antarmuka Pengguna Cerdas, hal. 263-266. ACM (2003)
8. Patel, R., Thakkar, P., Kotecha, K.: Meningkatkan sistem rekomendasi film. In: Jurnal Internasional Penelitian Lanjutan dalam Rekayasa dan Teknologi (IJARET), ISSN pp. 0976-6499 (2014)
9. Rich, E.: Pemodelan pengguna melalui stereotip. Cogn. Sains. **3**(4), 329-354 (1979)
10. Sarwar, B., Karypis, G., Konstan, J., Riedl, J.: Algoritma rekomendasi penyaringan kolaboratif berbasis item. Dalam: Prosiding Konferensi Internasional ke-10 tentang World Wide Web,
    Hal. 285-295. ACM (2001)
11. Thakkar, P., Varma, K., Ukani, V.: Pendekatan berbasis fusi hasil untuk penyaringan kolaboratif berbasis pengguna dan berbasis item. Dalam: Konferensi Internasional Teknologi Informasi dan Komunikasi untuk Sistem Cerdas, pp. 127-135. Springer (2017)
12. Wang, J., De Vries, A.P., Reinders, M.J.: Menyatukan pendekatan penyaringan kolaboratif berbasis pengguna dan berbasis item dengan fusi kemiripan. Dalam: Prosiding Konferensi Internasional ACM SIGIR Tahunan ke-29 tentang Penelitian dan Pengembangan dalam Temu Kembali Informasi, hal. 501-
    508. ACM (2006)
13. Yagnik, S., Thakkar, P., Kotecha, K.: Merekomendasikan tag untuk sumber daya baru dalam sistem penandaan buku sosial. Int. J. Data Min. Knowl. Manag. Proses **4**(1), 19 (2014)