INTISARI

Sistem rekomendasi merupakan suatu sistem yang digunakan dalam banyak domain di lingkungan *online* untuk membantu pengguna dalam memperoleh *item* yang mereka sukai. Salah satu metode sistem rekomendasi yang terkenal adalah *Memory-Based Collaborative Filtering* (MBCF) yang memanfaatkan algoritme similaritas untuk membangkitkan rekomendasi. Baru-baru ini, penelitian sistem rekomendasi telah mengembangkan algoritme similaritas dengan mengombinasikan similaritas berdasarkan nilai rating pengguna dan nilai perilaku pengguna. Nilai rating pengguna merupakan skor rating yang diberikan secara langsung terhadap *item*. Sementara itu, nilai perilaku pengguna merupakan akumulasi skor yang diperoleh secara tidak langsung dalam mengakses data genre. Masalah dengan algoritme similaritas ini hanya mempertimbangkan data genre untuk menangkap nilai perilaku pengguna, tanpa mengakomodasi data profil pengguna yang merupakan faktor yang mempengaruhi keputusan pengguna dalam memilih/membeli *item*. Masalah lainnya adalah proses penggabungan similaritas mengakibatkan kompleksitas algoritme menjadi meningkat, sehingga membutuhkan waktu yang lebih lama untuk menghasilkan rekomendasi. Algoritme similaritas yang mempertimbangkan data genre membutuhkan langkah tambahan untuk menghitung similaritas berdasarkan perilaku pengguna.

Berdasarkan masalah di atas, penelitian ini bertujuan untuk mengembangkan algoritme similaritas baru - yang disebut *User Profile Correlation-based Similarity* (UPCSim) - yang mengkaji data genre dan data profil pengguna, yaitu usia, jenis kelamin, pekerjaan, dan lokasi. Data profil pengguna digunakan untuk mencari bobot similaritas berdasarkan nilai rating pengguna dan nilai perilaku pengguna. Bobot kedua similaritas tersebut diperoleh dengan menghitung koefisien korelasi antara data profil pengguna dan nilai rating atau perilaku pengguna dengan menggunakan metode *multiple linear regression*. Selain itu, penelitian ini juga bertujuan untuk mengombinasikan metode pengelompokan data dan UPCSim, yang dikenal dengan *Clustering-Based* UPCSim (CB-UPCSim). Pengelompokan data didasarkan pada kemiripan profil pengguna untuk meningkatkan akurasi prediksi dan mereduksi waktu komputasi yang kompleks akibat penggabungan kedua similaritas. Evaluasi dilakukan dengan mengukur kinerja prediksi, kinerja klasifikasi, dan waktu pemrosesan. Kinerja prediksi diukur menggunakan *Mean Absolute Error* (MAE) dan *Root Mean Square Error* (RMSE). Sedangkan, kinerja klasifikasi diukur menggunakan *precision*, *recall*, dan F_1 *score*.

Hasil eksperimen pada *dataset* MovieLens 100K menunjukkan bahwa algoritme UPCSim mampu mengungguli algoritme similaritas UPCF (*User Probability score Collaborative Filtering*), dengan mereduksi nilai MAE dan RMSE sebesar 1,64% dan 1,4%. Dalam hal kinerja klasifikasi, algoritme UPCSim mampu mengungguli algoritme UPCF dengan meningkatkan nilai *precision* sebesar 0,05, *recall* sebesar 0,07, dan F_1 *score* sebesar 0,06. Namun, algoritme UPCSim membutuhkan waktu lebih lama 0,54 detik dibandingkan algoritme UPCF. Sementara itu, algoritme CB-UPCSim selain mampu mengungguli algoritme UPCF juga lebih unggul dari algoritme UPCSim, baik dalam hal kinerja prediksi, kinerja klasifikasi, dan waktu pemrosesan. Algoritme CB-UPCSim menghasilkan waktu pemrosesan 5,5 kali lebih cepat (dibandingkan algoritme UPCF) dan 6,1 kali lebih cepat (dibandingkan algoritme UPCSim). Uji hipotesis terhadap perbedaan kinerja algoritme similaritas *baseline* (UPCF) dengan algoritme similaritas yang diusulkan (UPCSim dan CB-UPCSim) menunjukkan bahwa ada perbedaan kinerja yang signifikan antara algoritme similaritas *baseline* (UPCF) dengan algoritme similaritas yang diusulkan (UPCSim dan CB-UPCSim).

Penurunan nilai MAE dan RMSE serta peningkatan nilai *precision*, *recall*, dan F_1 *score* tersebut memberikan dampak pada akurasi sistem rekomendasi yang meningkat, sehingga pengguna akan semakin nyaman dalam mencari *item* yang sesuai dengan keinginannya tanpa menghabiskan banyak waktu untuk memilih *item*. Keunggulan lain dari kedua algoritme yang diusulkan (UPCSim dan CB-UPCSim) adalah proses perhitungan similaritas yang bekerja berdasarkan metrik jarak tanpa menangkap perilaku fisik sistem, sehingga hanya bergantung pada karakteristik datanya. Oleh karena itu, UPCSim dan CB-UPCSim dapat menambahkan variabel baru tanpa mengubah model sistem saat atribut profil pengguna meningkat, sehingga kedua algoritme similaritas yang diusulkan dapat menjadi alternatif model similaritas untuk sistem rekomendasi pada domain lain yang memiliki karakteristik *dataset* yang sama. Sebagai contoh, implementasi model bisa digunakan dengan kombinasi 3 atribut profil pengguna (usia, jenis kelamin, dan pekerjaan) atau 4 atribut profil pengguna (usia, jenis kelamin, pekerjaan, dan lokasi). Hasil evaluasi model menunjukkan bahwa nilai MAE yang dihasilkan pada 4 atribut profil pengguna lebih kecil dibandingkan pada 3 atribut profil pengguna. Dari kedua algoritme yang diusulkan, algoritme CB-UPCSim menunjukkan kinerja yang lebih unggul dibandingkan algoritme UPCSim, baik dalam hal kinerja prediksi, kinerja klasifikasi, dan waktu pemrosesan. Hal ini terjadi karena algoritme CB-UPCSim menerapkan metode *clustering*, yaitu pengguna dengan preferensi yang sama berada dalam satu *cluster*, dan perhitungan similaritas hanya mempertimbangkan data pada satu *cluster* tanpa mengolah data pada *cluster* lainnya. Maka dari itu, algoritme CB-UPCSim menjadi metode unggulan yang diusulkan dalam penelitian ini.

Kata kunci: *Memory-Based Collaborative Filtering*, *clustering*, nilai rating pengguna, nilai perilaku pengguna, UPCSim, CB-UPCSim

ABSTRACT

A recommendation system is a system that is used in many domains in the online environment to assist users in finding the items they like. One of the well-known recommendation system methods is Memory-Based Collaborative Filtering (MBCF) *which utilizes similarity algorithms to generate recommendations. Recently, recommendation system research has developed a similarity algorithm by combining similarity based on user rating and behavior values. The user rating value is the rating score given directly to the item. Meanwhile, the user behavior value is an accumulated score obtained indirectly in accessing genre data. The problem with this similarity algorithm is that it only considers genre data to capture user behavior values without accommodating user profile data which is a factor that influences user decisions in choosing/buying items. Another problem is that combining similarities increases algorithm complexity, so it takes longer to generate recommendations. The similarity algorithm that considers genre data requires an additional step to calculate similarity based on behavior.*

Therefore, this study aims to develop a new similarity algorithm - called User Profile Correlation-based Similarity (UPCSim) - *which examines genre and user profile data, namely age, gender, occupation, and location. User profile data is used to find similarity weights based on user rating and behavior values. The weight of the two similarities is obtained by calculating the correlation coefficient between the user profile data and the rating value or user behavior using the multiple linear regression method. In addition, this study also aims to combine the clustering method and* UPCSim*, known as Clustering-Based* UPCSim (CB-UPCSim). *The data clustering is based on the similarity of user profiles to improve prediction accuracy in the* UPCSim *algorithm and reduce complex computational time due to combining the two similarities. Evaluation is done by measuring prediction performance, classification performance, and processing time. Prediction performance was measured using Mean Absolute Error* (MAE) *and Root Mean Square Error* (RMSE). *While the classification performance was measured using precision, recall, and* F_1 *score.*

The experimental results on the MovieLens 100K dataset show that the UPCSim *algorithm can outperform the* UPCF (*User Probability score Collaborative Filtering*) *similarity algorithm by reducing the* MAE *and* RMSE *values by* 1.64% *and* 1.4%*, respectively. In terms of processing time, the* UPCSim *algorithm takes* 0.54 *seconds longer than the UPCF algorithm. In terms of classification performance, the UPCSim algorithm can outperform the UPCF algorithm by increasing the precision value by* 0.05*, recall by* 0.07*, and* F_1 *score by* 0.06*. Meanwhile, apart from exceeding the UPCF algorithm, the* CB-UPCSim *algorithm is also superior to the* UPCSim *algorithm, both in prediction performance, classification performance, and processing time. The* CB-UPCSim *algorithm results in* 5.5 *times faster processing times* (*compared to the* UPCF *algorithm*) *and* 6.1 *times faster* (*compared to the* UPCSim *algorithm*)*.The validation of the performance difference between the baseline similarity algorithm* (UPCF) *and the proposed similarity algorithm* (UPCSim *and* CB-UPCSim) *shows that there is a significant performance difference between the baseline similarity algorithm* (UPCF) *and the proposed similarity algorithm* (UPCSim *and* CB-UPCSim).

The decrease in MAE *and* RMSE *values and the increase in precision, recall, and* F_1 *score impact increasing the accuracy of the recommendation system so that users will be more satisfied finding items that match their desires without spending a lot of time choosing items. Another advantage of the two proposed algorithms* (UPCSim and CB-UPCSim) *is the similarity calculation process that works based on distance metrics*

without capturing the physical behavior of the system, so it only depends on the data characteristics. Therefore, UPCSim *and* CB-UPCSim *can add new variables without changing the system model when the user profile attribute increases. Consequently, the two proposed similarity algorithms can be an alternative model for recommendation systems in other domains with the same dataset characteristics. For example, a model implementation can use three user profile attributes* (*age, gender, and occupation*) *or four user profile attributes* (*age, gender, occupation, and location*)*. The model evaluation results show that the MAE value generated for the four user profile attributes is smaller than the three user profile attributes. Of the two proposed algorithms, the* CB-UPCSim *algorithm shows superior performance to the* UPCSim *algorithm in prediction performance, classification performance, and processing time. The superiority of* CB-UPCSim *algorithms occurs because the* CB-UPCSim *algorithm applies the clustering method, where users with the same preferences are in one cluster, and the similarity calculation only considers data in one cluster without processing data in the other clusters. Therefore, the* CB-UPCSim *algorithm is the superior method proposed in this study.*

Keywords*: Memory-Based Collaborative Filtering, clustering, user rating value, user behavior value,* UPCSim*,* CB-UPCSim