# Sistem Rekomendasi Film Menggunakan Metode *Hybrid Collaborative Filtering* Dan *Content-Based Filtering*

1st Hilmi Hidayat Arfisko
*Fakultas Informatika*
*Universitas Telkom*
Bandung, Indonesia
hilmiha@students.telkomuniversity.ac.id

2nd Agung Toto Wibowo
*Fakultas Informatika*
*Universitas Telkom*
Bandung, Indonesia
agungtoto@telkomuniversity.ac.id

**Abstrak**

**Sistem rekomendasi pada dasarnya merupakan sistem yang berguna untuk menyaring dan mengidentifikasi item berupa produk, layanan atau informasi yang memiliki potensi besar untuk dipilih, dibeli ataupun digunakan oleh pengguna. Terdapat beberapa metode yang dapat digunakan dalam membangun sistem rekomendasi, seperti *collaborative filtering* yang merekomendasikan *item* berdasar kemiripan pengguna dalam hal memilih atau memberi nilai kepada *item* dan *content-based filtering* yang merekomendasikan *item* berdasarkan kemiripan *item* dalam hal isi atau konten *item* yang disukai oleh pengguna. Namun perlu diketahui, masing-masing metode ini memiliki kelemahan dan kelebihan. Untuk menutupi kelemahan masing-masing metode ini, pendekatan *hybrid* dapat dilakukan dimana kedua metode ini digabungkan dengan harapan dapat mengurangi kelemahan dari satu metode melalui kelebihan dari satu metode lainnya dan menghasilkan rekomendasi yang lebih baik. Oleh karena itu dalam penelitian ini dilakukan pembuatan sistem rekomendasi film menggunakan metode *hybrid collaborative filtering* dan *content-based filtering*. Dalam pengujiannya, hasil rekomendasi metode ini dibandingkan dengan hasil rekomendasi ketika hanya menggunakan metode *collaborative filtering* saja, metode *content-based filtering* saja dan metode *hybrid* dengan kedua metode tersebut dibalik. Dapat disimpulkan bahwa hasil pengujian yang dilakukan menggunakan metode hybrid collaborative filtering dan content-based filtering menghasilkan list rekomendasi item film yang lebih baik dibandingkan 3 metode lainnya yang diujicobakan terhadap keseluruha pengguna dalam dataset pengujian.**

**Kata kunci: sistem rekomendasi film, pendekatan *hybrid*, *collaborative filtering*, *content-based filtering***

***Abstract***

***The recommendation system is basically a system that is useful for filtering and identifying items in the form of products, services or information that have great potential to be selected, purchased or used by users. There are several methods that can be used in building a recommendation system, such as collaborative filtering which recommends items based on user similarity in terms of selecting or assigning value to items and content-based filtering which recommends items based on item similarity in terms of content or item content liked by users. However, it should be noted that each of these methods has advantages and disadvantages. To cover the weaknesses of each of these methods, a hybrid approach can be used where the two methods are combined in the hope of reducing the weaknesses of one method through the advantages of one method and producing better recommendations. Therefore, in this study, a film recommendation system was developed using hybrid collaborative filtering and content-based filtering methods. In the test, the recommended results of this method are compared with the recommendations when using only the collaborative filtering method, the content-based filtering method and the hybrid method with the two reversed. It can be concluded that the results of the tests carried out using the hybrid collaborative filtering method and content-based filtering resulted in a list of recommended film items that was better than the other 3 methods that were tested on all users in the test dataset.***

***Keywords: movie recommendation system, hybrid approach, collaborative filtering, content-based filtering***

## I. PENDAHULUAN

### A. Latar Belakang

Budaya menonton acara TV dan film sekarang ini dipermudah dengan adanya internet. Platform *streaming* seperti Netflix, HBO Max dan Disney+

memberikan lebih banyak fleksibilitas kepada pengguna untuk menonton acara TV dan film favorit mereka, kapan saja dan di perangkat apa pun. Jumlah acara TV atau film yang disediakan oleh masing-masing platform ini pun bisa dibilang cukup besar, dengan salah satu platform seperti netflix memiliki katalog berkisaran 6000 *item* acara TV dan film. Dengan banyaknya jumlah *item* yang disediakan, sistem rekomendasi menjadi fitur penting untuk dibangun dan memiliki peran besar dalam membantu pengguna menemukan *item* relevan yang mungkin mereka sukai. Selain itu, dari sudut pandang bisnis, sistem rekomendasi dapat membantu meningkatkan waktu aktfitas pengguna di dalam suatu platform dengan cara menampilkan *item-item* relevan kepada pengguna, yang mana nantinya seiring waktu dapat meningkatkan pendapatan untuk pelatform itu sendiri. Oleh karena itu, sistem rekomendasi merupakan salah satu fitur penting yang harus dimiliki oleh platform-platform seperti layanan *streaming* acara TV dan film.

Sistem rekomendasi pada dasarnya merupakan sistem yang berguna untuk menyaring dan mengidentifikasi *item* berupa produk, layanan atau informasi yang memiliki potensi besar untuk dipilih, dibeli ataupun digunakan oleh pengguna [1], [2]. Terdapat beberapa metode yang dapat digunakan dalam membangun sistem rekomendasi, seperti *collaborative filtering* (CF) yang merekomendasikan *item* berdasar kemiripan pengguna dalam hal memilih atau memberi nilai kepada *item* dan *content-based filtering* (CBF) yang merekomendasikan item berdasarkan kemiripan *item* dalam hal isi atau konten *item* yang disukai oleh pengguna [1], [3], [4]. Namun perlu diketahui, masing-masing metode ini memiliki kelemahan. Untuk menutupi kelemahan masing-masing metode ini, pendekatan *hybrid* dapat dilakukan dimana kedua metode ini digabungkan dengan harapan dapat mengurangi kelemahan dari satu metode melalui kelebihan dari satu metode lainnya dan menghasilkan sistem rekomendasi yang lebih baik [1], [3], [4].

Oleh karena itu, pada penelitian ini dibangun sistem rekomendasi menggunakan pendekatan *hybrid* metode *collaborative filtering* (CF) dan metode *content-based filtering* (CBF) terhadap *item* film. Adapun dataset yang digunakan dalam penelitian ini adalah dataset *movielens* yang disediakan oleh *grouplens.org* yang berisikan informasi *ratings* untuk film oleh pengguna serta informasi konten item film berupa genre dan tag mengenai film oleh pengguna.

B. Topik dan Batasannya

Dalam penelitian ini akan dibahas bagaimana mengimplementasikan metode *hybrid collaborative filtering* dan *content-based filtering* (*Hybrid* CF-CBF) dalam membangun sistem rekomendasi film. dengan harapan menghasilkan hasil rekomendasi lebih baik dibandingkan dengan hanya menggunakan masing-masing metode secara tersendiri.

Adapun dataset yang digunakana adalah dataset movielens yang telah diproses sebelumnya. Dataset ini berupa dataset 100.836 rating *item* film dengan 610 pengguna yang minimal telah memberikan rating terhadap 20 *item* film dan 9.742 *item* film. Selain itu tedapat pula dataset konten *item* film yang berisikan 32.977 tag untuk semua *item* film. Batasan pembahasan dari pembuatan tugas akhir ini tidak membahas dan juga membandingkan pembangunan system rekomendasi dengan metode lain selain penggunaan metode CF dan CBF, serta hanya berfokus pada pembangunan system rekomendasi menggunakan metode *hybrid* CF-CBF. Dikarenakan sifat dari kedua metode yang digunakan dalam pendekatan hybrid ini, sistem rekomendasi yang dibangun tidak dapat menangani *cold-start problem* pada pengguna baru.

C. Tujuan

Tujuan dari pembuatan tugas akhir ini adalah membangun sistem rekomendasi film menggunakan pendekatan metode *hybrid* CF-CBF dan bagaimana dampak pendekatan *hybrid* terhadap keakuratan hasil rekomendasi.

D. Organisasi Tulisan

Pada bab 2 akan dibahas mengenai dasar teori terkait penelitian yang dilakukan. Bab 3 akan dibahas mengenai perancangan sistem rekomendasi yang dibangun. Bab 4 akan membahas evaluasi dari sistem yang dibangun dan bab 5 akan membahas kesimpulan dari penelitian ini.

## II. KAJIAN TEORI

A. Sistem Rekomendasi

Sistem rekomendasi adalah perangkat lunak atau metode yang menghasilkan usulan berupa *item-item* spesifik yang mungkin menarik bagi suatu pengguna tertentu. Dalam sistem rekomendasi, terdapat dua pendekatan dalam menghasilkan usulan kepada pengguna, yaitu sistem rekomendasi yang bersifat *personalized* dan *non-personalized* [4].

Sistem rekomendasi yang bersifat *non-personalized* menghasilkan usulan dengan mengevaluasi keseluruhan *item* dalam sistem secara sekaligus tanpa mempertimbangkan preferensi pengguna [3]. Contoh implementasinya seperti menampilkan *item-item* yang paling populer atau *item-item* yang baru saja rilis.

Sedangkan sistem rekomendasi yang bersifat *personalized* menghasilkan usulan dengan mempertimbangkan preferensi setiap pengguna [4]. Dalam mengimplementasikan sistem rekomendasi yang bersifat *personalized*, terdapat 3 metode utama yang dapat dilakukan, yaitu *collaborative filtering*

(CF), *content-based filtering* (CBF) dan pendekatan *hybrid* [3].

*B. Collaborative Filtering*

Sistem rekomendasi dengan metode *collaborative filtering* (CF) bekerja dengan mengumpulkan umpan balik pengguna dalam bentuk *rating item-item* dan kemudian memanfaatkan kesamaan perilaku antar pengguna dalam memprediksi *rating* untuk menentukan bagaimana suatu *item* direkomendasikan [1], [3], [4]. Salah satu pendekatan dalam membangun sistem rekomendasi CF adalah *neighborhood-based method* [3]. Dalam pendekatan ini dilakukan prediksi *rating* terhadap target *item* dengan cara menghitung kemiripan (*similarity*) menggunakan koleksi *rating item-item* yang sebelumnya telah dinilai oleh pengguna. Pendekatan *neighborhood-based method* dalam CF dapat dibedakan menjadi dua, yaitu *user-based CF* dan *item-based CF* [3].

Pada penelitian ini akan berfokus pada penggunaan *item-based* CF, yang bekerja dengan mencari hubungan kemiripan *item* berdasar pada tabel *rating* untuk membuat rekomendasi *item* kepada pengguna [3]. Metode ini mengasumsikan jika beberapa *item* diberi nilai mirip oleh sekelompok pengguna, maka seorang target pengguna di dalam kelompok akan memberikan nilai terhadap *item-item* secara serupa dengan kebanyakan pengguna lain dalam kelompok.

Salah satu cara menghitung kemiripan *item* dalam *item-based* CF adalah dengan menggunakan perhitungan *pearson correlation* sesuai pada persamaan (1) [1], [3], [5], [6].

$$PC(i,j) = \frac{\sum_{u \in U_{ij}} (r_{ui} - \bar{r}_i)(r_{uj} - \bar{r}_j)}{\sqrt{\sum_{u \in U_{ij}} (r_{ui} - \bar{r}_i)^2 \sum_{u \in U_{ij}} (r_{uj} - \bar{r}_j)^2}} \quad (1)$$

Di mana:

| | | |
|---|---|---|
| $PC(i,j)$ | = | Nilai kemiripan item *i* dengan item *j* |
| $U_{ij}$ | = | Kumpulan pengguna yang memberi rating terhadap item *i* dan *j* |
| $r_{ui}$ dan $r_{uj}$ | = | Nilai rating yang diberikan oleh pengguna *u* dan terhadap item *i* dan *j* |
| $\bar{r}_i$ dan $\bar{r}_j$ | = | Nilai rata-rata rating item *i* dan item *j* |

Setelah nilai kemiripan *item* didapatkan, dipilih *subset item* (*neighbor*) berdasarkan nilai kemiripannya untuk memprediksi *rating* yang akan diberikan suatu pengguna terhadap *item*. Salah satu metode perhitungan yang digunakan untuk memprediksi rating baik untuk *item-based* CF adalah *weighted average* sesuai dengan persamaan (2) [2], [3], [6].

$$P(u,i) = \frac{\sum_{n \in N} r_{un} w_{in}}{\sum_{n \in N} |w_{in}|} \quad (2)$$

Di mana:

| | | |
|---|---|---|
| $P(u,i)$ | = | Nilai prediksi pengguna *u* terhadap ite |
| $N$ | = | Kumpulan neighbor item i |
| $r_{un}$ | = | Nilai rating item *n* oleh pengguna *u* |
| $w_{in}$ | = | Nilai similarity item *i* dengan item *n* |

Nantinya item-item yang memiliki nilai prediksi tertinggi dalam *neighbor* akan dijadikan rekomendasi.

Hal lain yang perlu diketahui mengenai sistem rekomendasi CF adalah sistem dengan metode ini memiliki beberapa kelebihan dan kekurangan yang dapat dilihat pada *Table 1*.

Table 1. Kelebihan dan kekurangan metode *collaborative filtering*

| Sistem Rekomendasi *Collaborative Filtering* | |
|---|---|
| Kelebihan | Kekurangan |
| • Hasil rekomendasi yang beragam dan bersifat s*erendipitous* (relevan dan baru) | • *Cold-start problem* (tidak dapat menghasilkan rekomendasi dikarenakan tidak adanya informasi preferensi) untuk pengguna baru dan *item* baru<br>• *Sparse problem* (matriks *rating* pengguna-item yang jarang/banyak yang kosong dapat mempengaruhi keakuratan algoritma) |

*C. Content-Based Filtering*

Sistem rekomendasi dengan metode *content-based filtering* (CBF) bekerja dengan menggunakan informasi profil preferensi pengguna terhadap *item* untuk dicari *item* dengan jenis yang mirip sebagai hasil rekomendasi [1], [3], [4]. Sistem rekomendasi CBF sebagian besar dirancang untuk merekomendasikan *item* berbasis teks, sehingga dalam hal ini konten dapat berupa kata kunci seperti kategori *item*, tag, dan genre [3].

Salah satu pendekatan dalam membangun sistem rekomendasi CBF adalah membangun profil pengguna dan *item* menggunakan konten yang telah diberi nilai [7]. Dalam hal ini profil pengguna memuat nilai bobot yang mewakili minat pengguna untuk setiap konten di dalam *item-item* yang pernah

dinilai pengguna. Sedangkan profil *item* memuat nilai bobot yang mewakili seberapa relevan konten-konten yang ada terhadap *item*.

Untuk membangun matriks profil *item* dapat digunakan TF-IDF (*term frequency-inverse document frequency*) yang merupakan statistik numerik yang mencerminkan nilai kepentingan sebuah kata pada dokumen di dalam koleksi [1], [3].

Nilai TF-IDF didapat dari perhitungan TF yang melambangkan frekuensi dari istilah subjek muncul dalam dokumen, dan IDF yang melambangan frekuensi dokumen yang mengandung istilah subjek dalam koleksi. Adapun nilai TF untuk suatu istilah dalam dokumen dapat dihitung menggunakan persamaan (3).

$$tf(t,d) = 1 + \log\left(f_{td}\right) \tag{3}$$

Di mana:

$tf(t,d)$ = Nilai TF untuk istilah *t* dalam dokumen *d*
$f_{td}$ = Jumlah istilah *t* dalam dokumen *d*

Nilai IDF untuk suatu istilah di dalam koleksi dapat dihitung menggunakan persamaan (4).

$$idf(t) = \log\left(\frac{N}{df_t}\right) \tag{4}$$

Di mana:

$idf(t)$ = Nilai IDF untuk istilah *t* dalam koleksi
$df_t$ = Jumlah dokumen yang mengandung istilah *t*
N = Total jumlah dokumen dalam koleksi

Bobot TF-IDF untuk istilah t di dalam dokumen d dapat dihitung menggunakan persamaan (5).

$$tf - idf(t,d) = \mathrm{tf}(t,d) \times idf(t) \tag{5}$$

Di mana:

$tf - idf(t,d)$ = Nilai TF-IDF untuk istilah *t* untuk dukumen d dalam koleksi
$tf(t,d)$ = Nilai TF untuk istilah *t* dalam dokumen *d*
$idf(t)$ = Nilai IDF untuk istilah *t* dalam koleksi

Setelah matriks profil *item* dibuat, profil pengguna dapat dibangun. Dikarenakan profil pengguna menunjukkan tingkat preferensi pengguna terhadap *item*, profil pengguna dapat dibangun menggunakan jumlah total vektor fitur/konten untuk semua *item* yang dinilai positif atau relevan oleh pengguna [8].

Menggunakan vektor profil item dan pengguna yang telah dibangun, dilakukan prediksi nilai ketertarikan pengguna terhadap konten *item* apakah sesuai dengan preferensi untuk direkomendasikan [8], [9]. Salah satu cara menghitung prediksi antara dua variabel ini adalah dengan menggunakan *dot product* antara vektor profil *item* dan pengguna [8]. Nantinya *item-item* yang memiliki nilai prediksi tertinggi akan dijadikan rekomendasi.

Hal lain yang perlu diketahui mengenai sistem rekomendasi CBF adalah sistem dengan metode ini memiliki beberapa kelebihan dan kekurangan yang dapat dilihat pada *Table 2*.

Table 2. Kelebihan dan kekurangan metode *content-based filtering*

| Sistem Rekomendasi *Content-Based Filltering* | |
|---|---|
| Kelebihan | Kekurangan |
| • *User-independent* sehingga tidak mengalami *sparse problem*<br>• Tidak mengalami *cold-start problem* untuk item baru | • *Cold-start problem* untuk pengguna baru<br>• *Over-specialization problem* (pengguna terbatas mendapatkan rekomendasi yang mirip dengan yang sudah dikenal dalam profilnya) |

*D.* Pendekatan *Hybrid*

Pendekatan *hybrid* menggabungkan beberapa metode sistem rekomendasi yang berbeda untuk menghasilkan sistem yang lebih baik dengan cara mengurangi kelemahan dari satu metode melalui kelebihan dari satu metode lain [1], [3], [4]. Misalnya dengan penggunaan metode CBF untuk menangani *cold-start problem item* baru pada

metode CF [4]. Atau penggunaan metode CF untuk menangani *over-specialization problem* pada metode CBF [4]. Dan atau mengatasi permasalahan sparsity data pada CF menggunakan CBF yang hanya memerlukan preferensi konten item yang dikonsumsi pengguna saja [4].

Pendekatan *hybrid* pada sistem rekomendasi dibagi menjadi tiga kategori utama, yang masing-masing memiliki beberapa pendekatan. Adapun ketegori tersebut adalah *monolithic*, *parallel* dan *pipline hybridization* [10].

Pada penilitian ini, difokuskan pada pembuatan sistem rekomendasi dengan kategori *pipline hybridization* dengan pendekatan *cascade* terhadap metode CF dan CBF. Pendekatan ini merupakan proses rekomendasi yang bertahap dimana metode rekomendasi pertama diminta untuk menghasilkan *list* kasar kandidat *item* rekomendasi dan metode rekomendasi kedua dan seterusnya menyaring kembali *list* kandidat awal untuk dijadikan *list* rekomendasi akhir [10], [11]. Dikarenakan sifatnya ini, semua metode kecuali metode pertama dalam pendekatan ini hanya dapat mengubah urutan dan menghilangkan *item* dalam daftar rekomendasi serta tidak dapat memperkenalkan *item* baru atau mengembalikan *item* yang telah dihilangkan dari metode sebelumnya ke dalam daftar rekomendasi [10]. Selain itu, pendekatan *hybrid* ini peka terhadap urutan *item* dalam *list* rekomendasi, dimana hasil rekomendasi dengan metode *hybrid* CF-CBF akan berbeda dengan *hybrid* CBF-CF.

E. Pengujian Sistem Rekomendasi

Dalam konteks sistem rekomendasi yang biasanya dibuat untuk menghasilkan *top-N item* teratas kepada pengguna, diperlukan metrik yang memberi sistem penghargaan untuk merekomendasikan *item* yang tidak hanya benar dan relevan, namun berhasil menempatkan *item* paling relevan di posisi atas dan yang kurang relevan di posisi bawah dalam *list* rekomendasi. Oleh karena itu, dalam penelitian digunakan metrik pengujian *Hit Rate* dan *Mean Average Precision* (MAP).

*F. Hit Rate*

Hit Rate (HR) dalam konteks rekomendasi merupakan metrik ya atau tidak, yang melihat apakah ada *item* relevan dalam *list* rekomenadasi untuk pengguna tertentu. Jika ada *item* relevan dalam *list* rekomendasi untuk perngguna tersebut, maka dihitung 1 hit [12]. Adapun perhitungan *hit rate* terhadap sekumpulan pengguna dalam satu set pengujian dapat dihitung menggunakan persamaan (6).

$$HR = \frac{|U^k_{\underline{hi}t}|}{|U_{all}|} \quad (6)$$

Di mana:

$HR$ = Nilai *hit rate*

$U^k_{hit}$ = Jumlah pengguna dalam pengujian dimana item relevan berada dalam *top*-k hasil rekomendasi

$U_{all}$ = Jumlah pengguna dalam pengujian

*G. Mean Average Precision*

*Mean Average Precision* (MAP) merupakan metrik untuk mengukur kinerja model yang melakukan tugas pencarian dokumen/informasi dan deteksi objek [12]. Penggunaan metrik MAP cocok untuk algoritma yang mengembalikan urutan peringkat *item* dalam *list*, dimana setiap item bisa *hit* atau *miss* (relevan atau tidak relevan) untuk suatu pengguna dan *item* yang rendah dalam *list* cenderung tidak digunakan [12]. Sebelum MAP dapat dihitung, diperlukan informasi mengenai *precision*, *precision at k* dan *average precision*.

*Table 3.* Confusion matrix

| | | Reality | |
|---|---|---|---|
| | | Relevant | Not Relevant |
| Prediction | Recomended | True Positive (TP) | False Positive (FP) |
| | Not Recomended | False Negative (FN) | True Negative (FN) |

*Precision* adalah rasio banyak item relevan yang direkomendasikan di dalam satu set item yang direkomendasikan [12]–[15]. Menggunakan *comfusion matrix* yang dapat dilihat pada *Table 3*, precision dapat dihitung menggunakan persamaan (7).

$$P = \frac{TP}{TP + FP} \tag{7}$$

Di mana:

| | | |
|---|---|---|
| P | = | Nilai precision |
| TP | = | Jumlah item relevan yang direkomendasikan |
| TP + FP | = | Jumlah item yang direkomendasikan |

*Precision at k* atau bisa disebut dengan P(k), merupakan perhitungan *precision* dengan mempertimbangkan subset daftar rekomendasi dari peringkat satu hingga *k* [13], [14]. Adapun perhitungan *precision at k* dapat dicontohkan seperti pada *Gambar 4.*

*Gambar 4.* Contoh perhitungan *precision at k*

| rank | prediction | result |
|---|---|---|
| 1 | Item 1 | True Positive |
| 2 | Item 6 | False Positive |
| 3 | Item 4 | False Positive |
| | Item 7 | False Positive |
| | Item 3 | True Positive |
| | Item 2 | True Positive |
| | Item 5 | False Positive |

| rank | prediction | result |
|---|---|---|
| 1 | Item 1 | True Positive |
| 2 | Item 6 | False Positive |
| 3 | Item 4 | False Positive |
| 4 | Item 7 | False Positive |
| 5 | Item 3 | True Positive |
| 6 | Item 2 | True Positive |
| | Item 5 | False Positive |

$k$=3

$$P(k) = \frac{1}{3}$$

$k$=6

$$P(k) = \frac{3}{6}$$

*Average precision* (AP) merupakan gambaran lebih mengenai kemampuan sistem untuk mengurutkan item di dalam *list* [13], [14]. Adapun nilai *average precision* untuk *N* jumlah item rekomendasi dapat dihitung menggunakan persamaan (8).

$$AP@N = \frac{1}{\min(m,N)} \sum_{k=1}^{N} P(k) \times rel(k) \quad if\ m \neq 0$$
$$AP = 0 \quad if\ m = 0 \tag{8}$$

Di mana:

| | | |
|---|---|---|
| $AP@N$ | = | Nilai average precision untuk N jumlah item yang direkomendasikan |
| $N$ | = | Jumlah item yang direkomendasikan |
| $m$ | = | Jumlah item relevan |
| $P(k)$ | = | Nilai *Precision at k* |
| $rek(k)$ | = | Indikator jika item ke-*k* relevan ($rek(k) = 1$) atau tidak relevan ($rek(k) = 0$) |

*Mean Average Precision* (MAP) merupakan gambaran *average precision* untuk semua pengguna [13], [14]. Adapun nilai MAP untuk *N* jumlah *item* yang direkomendasikan dapat dihitung menggunakan persamaan (9).

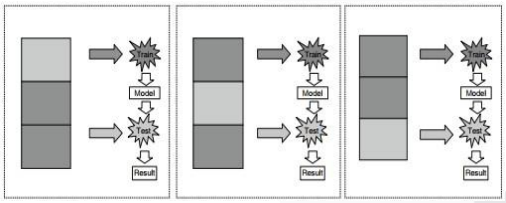
$$MAP@N = \frac{1}{|U|}\sum_{u \in U}(AP@N)_u \quad (9)$$

Di mana:

| | | |
|---|---|---|
| $MAP@N$ | = | Nilai mean average precision untuk N jumlah item yang direkomendasikan terhadap seluruh pengguna. |
| $N$ | = | Jumlah item yang direkomendasikan |
| $U$ | = | Kumpulan pengguna |

H. K-Fold Cross Validation

*Cross validation* merupakan pendekatan statistik yang digunakan untuk mengevaluasi dan memperkirakan keterampilan model algoritma pembelajaran, dengan cara membagi dataset yang digunakan menjadi dua segmen, di mana satu segmen digunakan dalam pelatihan terhadap model dan segmen sisanya digunakan dalam validasi terhadap model [16]. Bentuk dasar dari *cross validaton* adalah *k-fold cross validation*, dimana dataset dibagi menjadi k buah segmen yang sama atau hampir sama ukurannya untuk kemudian dilakukan sebanayak k iterasi pelatihan dan validasi sehingga di dalam setiap iterasi, segmen dataset yang berbeda digunakan dalam validasi model dan sisa k-1 segmen lainnya digunakan dalam pelatihan model. Pada *Gambar 1* dapat dilihat ilustrasi contoh *k*=3, dimana segmen data berwarna gelap digunakan sebagai dataset pelatihan mode, dan segmen dataset yang terang digunakan untuk validasi.

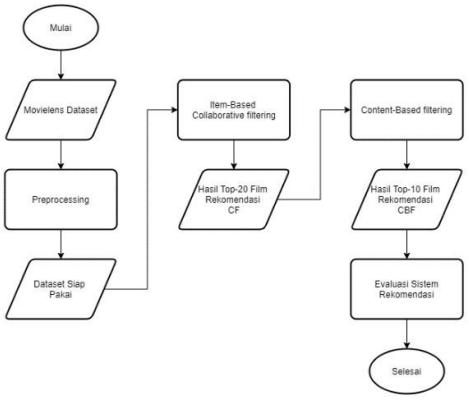



Gambar 1. Ilustrasi pembagian dataset dengan k=3 dalam k-fold cross validation

Untuk menghitung kinerja setiap algoritma pembelajaran pada setiap segmen, digunakan beberapa perhitungan akurasi seperti metrik evaluasi yang telah ditentukan. Setelah seluruh iterasi dilakukan, sample metrik evaluasi akan tersedia untuk setiap algoritma dan penggunaan metodologi seperti penggunaan rata-rata dapat digunakan dalam menghasilkan nilai agregat sample metrik evaluasi untuk setiap algoritma [16].

## III. METODE

Pada penelitian ini, metode yang digunakan adalah metode *Hybrid* CF-CBF dengan pendekatan *cascade*. Pada pendekatan *hybrid* ini, sistem rekomendasi yang dibuat diminta untuk menghasilkan *list* 20 *item* rekomendasi menggunakan metode *item-based* CF, yang kemudian dilakukan pengurutan ulang menggunakan metode CBF untuk diambil 10 *item* teratas sebagai hasil rekomendasi akhir, sesuai dengan alur pada Gambar 2.




*Gambar 2.* Alur sistem rekomendasi hybrid CF-CBF yang dibangun

A. Preprocessing Data

Dalam penelitian ini, digunakan dataset movielens yang disediakan oleh grouplens.org yang berisikan *ratings item* film oleh pengguna serta informasi konten *item* berupa judul, genre dan *tag* mengenai *item* film. Dataset-dataset ini di proses menjadi dataset rating film-pengguna dan dataset konten item film.

Dataset rating pengguna-film, yang berisi daftar rating-rating item film yang diberi oleh pengguna sesuai dengan representasinya pada *Table 5*. Dalam proses pembuatan dataset ini, telah dilakukan pengecekkan terhadap dataset movielens untuk menghindari adanya data yang kosong.

*Table 5*. Representasi dataset rating film-pengguna

| userId | movieId | rating |
|---|---|---|
| 1 | 1 | 4.0 |

| | | |
|---|---|---|
| 1 | 3 | 4.0 |
| : | : | : |
| 610 | 168252 | 5.0 |
| 610 | 170875 | 3.0 |

Sedangkan untuk dataset konten *item* film, berisi daftar *item* film dan kontennya yang berupa kata yang didapatkan dari judul, genre dan *tag* mengenai *item* film dengan representasinya sesuai dengan *Table 6*. Dalam proses pembuatan dataset ini, konten-konten item yang ada telah dilakukan proses pemotongan kalimat menjadi kata. Selain itu dilakukan pula penghilangan tanda baca, karakter non-ascii, nomor dan *stopwords* dari daftar konten *item* film.

*Table 6.* Representasi dataset konten *item* film

| movieId | konten |
|---|---|
| 1 | "toy" |
| 1 | "story" |
| 1 | "adventure" |
| : | : |
| 176329 | "comedy" |
| 145994 | "soviet" |
| 145994 | "classics" |

B. Pembuatan *List* Rekomendasi Film Metode *Item-Based* CF

Dalam proses ini dilakukan pembuatan *list* rekomendasi film menggunakan metode *item-based* CF dengan cara mencari hubungan kemiripan antar *item* berdasarkan tabel *rating*. Untuk mencari kemiripan, digunakan perhitungan *pearson correlation* sesuai dengan persamaan (1).

Tahap selanjutnya adalah memilih *subset* 15 buah *item* film (*neighbor*) berdasarkan nilai kemiripannya untuk digunakan dalam prediksi rating suatu pengguna terhadap item film. Nilai prediksi *rating* nantinya dihitung menggunakan *weighted sum* sesuai dengan persamaan (2) yang kemudian untuk setiap pengguna dipilih 20 *item* dengan nilai prediksi rating terbesar untuk dijadikan hasil rekomendasi tahap pertama. Hasil rekomendasi tahap pertama ini akan dilajutkan ke metode CBF untuk diurutkan ulang.

C. Pembuatan Profil Item dan Profil Pengguna untuk Metode CBF

Dalam proses ini, dilakukan pembuatan profil *item* film dan pengguna. Untuk membangun profil *item* film, dilakukan perhitungan TF-IDF terhadap konten untuk *item* film menggunakan persamaan (5). Untuk pembuatan profil pengguna, dapat dilakukan dengan menjumlah total vektor fitur/konten dari semua item yang dinilai positif atau relevan oleh pengguna (diberi rating >=3).

D. Pembuatan List Rekomendasi Film Metode CBF Menggunakan Hasil Rekomendasi CF

Dalam tahap ini dilakukan pengurutan ulang 20 *item* dalam *list* rekomendasi metode CF menggunakan CBF, untuk diambil 10 item teratas sebagai hasil rekomendasi akhir. Pada tahap ini, vektor profil *item* hasil rekomendasi CF dan vektor profil pengguna digunakan untuk memprediksi konten *item* film apakah sesuai dengan preferensi pengguna dan cocok untuk diletakkan pada posisi 10 item teratas menggunakan perhitungan *dot product* antara kedua vector profile. *Item* kemudian disusun secara *descending* berdasarkan hasil perhitungan prediksinya dan diambil 10 *item* teratas sebagai hasil rekomendasi akhir.

**a. Melakukan Evaluasi Hasil Rekomendasi Akhir**

Pengujian sistem yang dibangun dilakukan menggunakan *k-fold cross validation* untuk membagi dataset *rating* film-pengguna menjadi

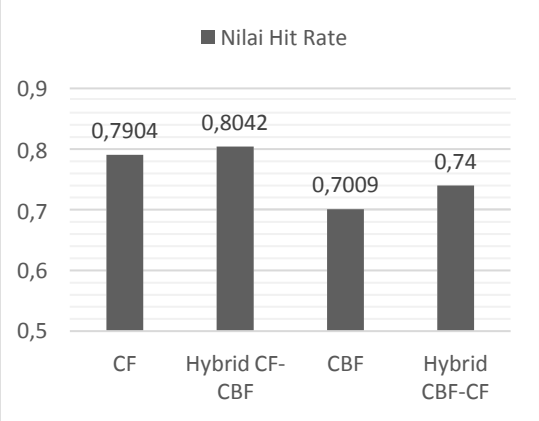
dataset *training* dan *testing*. Digunakan nilai k=5 untuk membagi dataset menjadi 5 bagian, dimana natinya 4 dari 5 bagian dataset tersebut digunakan dalam proses *training* dan 1 sisanya digunakan dalam proses *testing* yang dilakukan secara perulangan sebanyak 5 kali.

Dalam proses *training*, sistem rekomendasi akan menghasilkan *list* 20 *item* film sebagai rekomendasi untuk setiap pengguna dalam dataset *training*, dan hanya 10 *item* teratas didalam *list* yang dijadikan hasil rekomendasi akhir yang kemudian dibandingkan dengan dataset *testing* apakah relevan atau tidak untuk setiap pengguna menggunakan metrik evaluasi *Hit Rate* dan MAP@N dengan nilai N pada MAP=10. Selain itu dilakukan pula pengujian terhadap 10 pengguna paling aktif dalam memberikan rating (rata-rata memberi rating terhadap 1313,6 *item*) dan 10 pengguna paling tidak aktif (rata-rata memberi rating terhadap 14,14 *item*) yang bertujuan untuk mengetahui seberapa baik masing-masing metode dalam menghasilkan rekomendasi dalam dua kondisi tersebut.

Hasil rekomendasi yang diuji dan dibandingkan untuk dievaluasi dalam penelitian ini adalah hasil rekomendasi metode CF saja tanpa proses *hybrid* (CF) dan CF dengan proses *hybrid* CBF (*Hybrid* CF-CBF), serta rekomendasi ketika kedua metode dibalik dalam proses pendekatan hybridnya, sehingga dilakukan pengujian terhadap hasil rekomendasi CBF saja tanpa proses *hybrid* (CBF), dan CBF dengan proses *hybrid* CF (*Hybrid* CBF-CF).

## IV. HASIL DAN PENGUJIAN

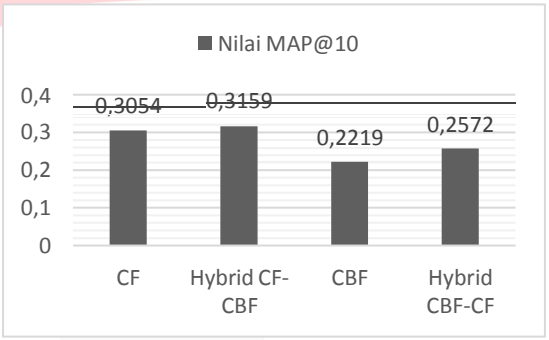
Pengujian pertama dilakukan untuk menghasilkan list rekomendasi untuk semua pengguna dari 4 metode, yaitu metode CF, *Hybrid* CF-CBF, CBF dan *Hybrid* CBF-CF. Berdasarkan 5 kali pengulangan *k-fold cross validation* untuk keempat metode, didapatkan nilai rata-rata metrik evaluasi *hit rate* seperti yang ditampilkan pada *Gambar 3* dan nilai rata-rata metrik evaluasi MAP@10 pada *Gambar 4*.




*Gambar 3.* Hasil perhitungan hit rate percobaan pertama

Pada *Gambar 3*, dari empat metode yang dijalankan, metode *Hybrid* CF-CBF memiliki nilai *hit rate* tertinggi diantara metode lain dengan nilai 0,8042. Hal ini menunjukkan metode tersebut berhasil menampilkan setidaknya satu item relevan dalam *list* rekomendasi terhadap 80,4% keseluruhan pengguna dalam dataset *training* disetiap iterasi *k-fold cross validation*.

Selain itu dalam pengujian ini dapat dilihat pula nilai *hit rate* yang meningkat untuk hasil metode CF dan CBF setelah dilakukannya pendekatan *hybrid*. Hal ini menunjukkan bahwa dengan adanya penyusunan ulang 20 *item* hasil rekomendasi CF atau CBF menggunakan pendekatan *hybrid*, dapat memunculkan *item-item* relevan yang tersembunyi di dalam *list* ke posisi lebih atas sehingga masuk ke dalam 10 *item* rekomendasi akhir. Hal ini pula yang menjadi alasan mengapa nilai MAP yang meningkat untuk metode CF dan CBF setelah dilakukannya pendekatan *hybrid*, sesuai dengan hasil pengujian pada *Gambar 4*.



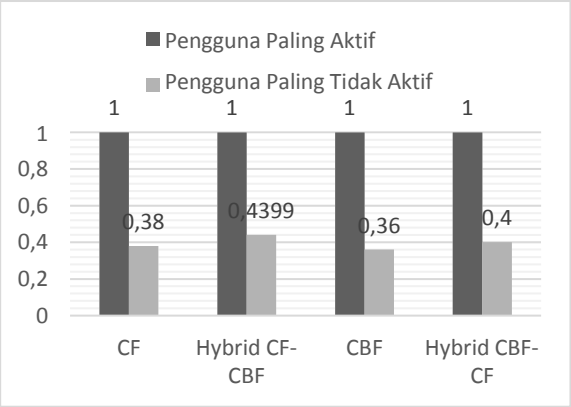

*Gambar 4.* Hasil perhitungan MAP@10 percobaan pertama

Pada *Gambar 4*, dapat dilihat pula bahwa metode *Hybrid* CF-CBF memiliki nilai MAP tertinggi dengan nilai MAP 0,3159 yang menunjukkan bahwa metode tersebut dapat menghasilkan urutan hasil rekomendasi yang lebih baik dibandingkan dengan 3 metode lainnya, dimana *item* yang mungkin relevan untuk pengguna berada di posisi teratas dan *item* yang mungkin kurang relevan berada di posisi bawah dalam *list* rekomendasi akhir.

Dari kedua perhitungan evaluasi *hit rate* dan MAP terhadap keempat metode dalam percobaan pertama ini, didapatkan metode *Hybrid* CF-CBF menghasilkan *list* rekomendasi *item* film yang paling baik dibandingkan dengan metode lainnya. Adapun penyebab mengapa metode *Hybrid* CF-CBF memiliki nilai *hit rate* dan MAP yang lebih besar dibandingkan metode *Hybrid* CBF-CF dikarenakan sifat pendekatan *hybrid cascade* yang digunakan peka terhadap urutan *item* dalam *list* rekomendasi, dimana *list* rekomendasi metode pertama akan menjadi *list* rekomendasi kasar yang hanya akan diubah urutan atau dihilangkan *item*nya dari *list* oleh metode rekomendasi selanjutnya menjadi *list*

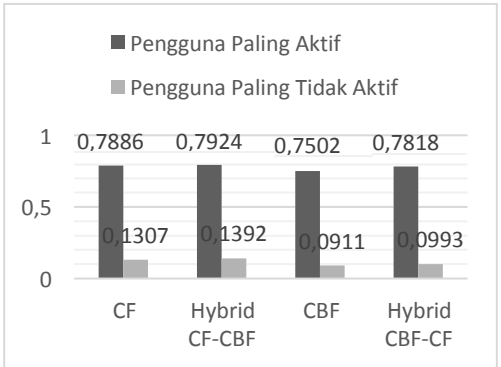
rekomendasi akhir. Hal ini menyebabkan kualitas hasil rekomendasi akhir dari pendekatan *hybrid* ini akan dipengaruhi oleh seberapa baik metode pertama dalam dalam menghasilkan *list* rekomendasi kasarnya.

Selain pengujian pertama, dilakukan pula pengujian kedua terhadap 10 pengguna paling aktif dalam memberikan *rating* dan 10 pengguna paling tidak aktif. Dari 5 kali perulangan *k-fold cross validation* untuk keempat metode, didapatkan nilai rata-rata metrik evaluasi *hit rate* seperti yang ditampilkan pada *Gambar 5* dan nilai rata-rata metrik evaluasi MAP@10 pada *Gambar 6*.




*Gambar 5*. Hasil perhitungan hit rate percobaan kedua

Berdasarkan hasil pengujian yang dilakukan untuk menghasilkan rekomendasi untuk kasus 10 pengguna paling aktif pada *Gambar 5*, keempat metode memiliki nilai *hit rate* sempurna dengan nilai 1,00. Hal ini menunjukkan bahwa keempat metode tersebut dapat menampilkan setidaknya satu *item* relevan dalam *list* rekomendasi untuk seluruh 10 pengguna paling aktif disetiap iterasi *k-fold cross validation*. Namun hal ini berubah ketika keempat metode diminta menghasilkan rekomendasi untuk kasus 10 pengguna paling tidak aktif seperti yang dapat dilihat pada *Gambar 5*, dimana nilai *hit rate* untuk setiap metode menurun drastis jika dibandingkan dengan hasil rekomendasi untuk kasus 10 pengguna paling aktif.




*Gambar 6*. Hasil perhitungan MAP@10 percobaan kedua

Pola yang sama dapat pula dilihat pada nilai MAP untuk hasil rekomendasi percobaan ini, sesuai dengan hasil pada *Gambar 6*. Hal ini menunjukkan bahwa jumlah *item* yang dirating oleh pengguna dapat mempengaruhi kualitas hasil rekomendasi keempat metode. Semakin banyak item yang telah dirating oleh pengguna, maka semakin baik pula kualitas hasil rekomendasi yang dihasilkan oleh masig-masing metode diuji.

Berdasarkan *Gambar 5* dan *Gambar 6* ketika dihadapkan dengan kasus 10 pengguna paling tidak aktif, dapat dilihat pola nilai *hit rate* dan MAP yang meningkat untuk hasil metode CF dan CBF setelah dilakukan pendekatan *hybrid*. Hal ini menunjukkan meskipun dengan keadaan pengguna baru memberikan *rating* terhadap sedikit *item* film, dengan adanya penambahan pendekatan *hybrid* terhadap hasil rekomendasi dapat meningkatkan kualitas hasil rekomendasi jika dibandingkan dengan hanya menggunakan kedua metode tersebut secara tersendiri.

## V. KESIMPULAN

Penulisan tugas akhir ini berfokus pada bagaimana membangun sistem rekomendasi film menggunakan pendekatan *hybrid collaborative filtering* dan *content-based filtering* (Hybrid CF-CBF) serta bagaimana dampak pendekatan *hybrid* terhadap keakuratan hasil rekomendasi. Berdasarkan hasil pengujian yang telah dilakukan terhadap empat metode, didapatkan kesimpulan sebagai berikut:

1. Metode *hybrid* CF-CBF yang dibangun menghasilkan *list* rekomendasi *item* film yang paling baik dibandingkan tiga metode lain yang diujikan berdasarkan perhitungan evaluasi metrik *hit rate* dan MAP.
2. Jumlah item yang dirating oleh pengguna dapat mempengaruhi hasil rekomendasi. Semakin banyak item yang telah dirating oleh pengguna, maka semakin baik masing-masing metode menghasilkan rekomendasi yang berkualitas.
3. Dikarenakan sifat dari metode cascade dalam pendekatan *hybrid* yang dibangun, kualitas hasil rekomendasi akhir dari pendekatan *hybrid* CF-CBF ataupun *hybrid* CBF-CF akan dipengaruhi oleh seberapa baik metode pertama dalam pendekatan *hybrid* dalam menghasilkan *list* rekomendasi kasarnya.
4. Penambahan pendekatan *hybrid* terhadap hasil rekomendasi CF ataupun CBF dapat meningkatkan kualitas hasil rekomendasi jika dibandingkan dengan hanya menggunakan kedua metode tersebut secara tersendiri serta membantu dalam menghasilkan hasil rekomendasi yang lebih baik ketika dihadapkan dengan kasus

pengguna yang baru memberikan *rating* terhadap sedikit *item* film.

Saran yang dapat dipertimbangkan untuk penelitian selanjutnya adalah mecoba metode rekomendasi lain dan atau menggunakan pendekatan *hybrid* dengan tipe lain dalam menghasilkan *list* rekomendasi sebagai pembanding.

REFERENSI


[1] F. O. Isinkaye, Y. O. Folajimi, and B. A. Ojokoh, "Recommendation systems: Principles, methods and evaluation," *Egyptian Informatics Journal*, vol. 16, no. 3, pp. 261–273, 2015, doi: 10.1016/j.eij.2015.06.005.

[2] G. Geetha, M. Safa, C. Fancy, and D. Saranya, "A Hybrid Approach using Collaborative filtering and Content based Filtering for Recommender System," *Journal of Physics: Conference Series*, vol. 1000, no. 1, 2018, doi: 10.1088/1742-6596/1000/1/012101.

[3] P. Nagarnaik and A. Thomas, "Survey on recommendation system methods," *2nd International Conference on Electronics and Communication Systems, ICECS 2015*, vol. 137, no. 7, pp. 1603–1608, 2015, doi: 10.1109/ECS.2015.7124857.

[4] D. Das, L. Sahoo, and S. Datta, "A Survey on Recommendation System," *International Journal of Computer Applications*, vol. 160, no. 7, pp. 6–10, 2017, doi: 10.5120/ijca2017913081.

[5] C. S. M. Wu, D. Garg, and U. Bhandary, "Movie Recommendation System Using Collaborative Filtering," *Proceedings of the IEEE International Conference on Software Engineering and Service Sciences, ICSESS*, vol. 2018-Novem, pp. 11–15, 2019, doi: 10.1109/ICSESS.2018.8663822.

[6] K. Y. Jung, D. H. Park, and J. H. Lee, "Hybrid collaborative filtering and content-based filtering for improved recommender system," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 3036, pp. 295–302, 2004, doi: 10.1007/978-3-540-24685-5_37.

[7] S. Reddy, S. Nalluri, S. Kunisetti, S. Ashok, and B. Venkatesh, *Content-based movie recommendation system using genre correlation*, vol. 105, no. September. Springer Singapore, 2019. doi: 10.1007/978-981-13-1927-3_42.

[8] J. Jeon, "Data Science Series: Content-based Recommender System using Azure Databricks," *Visualbi*, 2018. https://visualbi.com/blogs/microsoft/azure/data-science-series-content-based-recommender-system-using-azure-databricks/

[9] K. Luk, "Introduction to TWO approaches of Content-based Recommendation System," *Towards Data Science Web page*, 2019. https://towardsdatascience.com/introduction-to-two-approaches-of-content-based-recommendation-system-fc797460c18c

[10] D. Jannach, M. Zanker, A. Felfernig, and G. Friedrich, *Recommender Systems*. Cambridge: Cambridge University Press, 2010. doi: 10.1017/CBO9780511763113.

[11] E. Çano and M. Morisio, "Hybrid recommender systems: A systematic literature review," *Intelligent Data Analysis*, vol. 21, no. 6. IOS Press, pp. 1487–1524, 2017. doi: 10.3233/IDA-163209.

[12] B. Wang, "Ranking Evaluation Metrics for Recommender Systems," *Towards Data Science*, 2021. https://towardsdatascience.com/ranking-evaluation-metrics-for-recommender-systems-263d0a66ef54

[13] R. J. Tan, "Breaking Down Mean Average Precision (mAP)," *Towards Data Science*, 2019. https://towardsdatascience.com/breaking-down-mean-average-precision-map-ae462f623a52#1a59

[14] S. Sawtelle, "Mean Average Precision (MAP) For Recommender Systems," 2016. sdsawtelle.github.io/blog/output/mean-average-precision-MAP-for-recommender-systems.html

[15] G. Shani and A. Gunawardana, "Evaluating Recommendation Systems," *Recommender Systems Handbook*, pp. 257–297, 2011, doi: 10.1007/978-0-387-85820-3_8.

[16] P. Refaeilzadeh, L. Tang, and H. Liu, "Cross-Validation," in *Encyclopedia of Database Systems*, Springer New York, 2016, pp. 1–7. doi: 10.1007/978-1-4899-7993-3_565-2.