TIERS Information Technology Journal
Vol.5, No.1, June 2024, pp. 13~19
ISSN: 2723-4533 / E-ISSN: 2723-4541
DOI: 10.38043/tiers.v4i1.5209

Marvel Movie Recommendation System Using Hybrid Item-Based and Content-Based Filtering Methods

Daffa Rizki Putra Noordi[1], Herliyani Hasanah[2], Sri Sumarlinda[3]
daffarizki7686@gmail.com[1], herliyani_hasanah@udb.ac.id[2], sri_sumarlinda@udb.ac.id[3]
[1,2,3] Informatics Engineering Study Program, Universitas Duta Bangsa Surakarta

ABSTRACT

Currently, there are so many movie genres available to the general public, making it difficult for viewers to choose a movie. One of the most popular movies is the “Marvel Movies” or MCU (Marvel Cinematic Universe), which has become the highest grossing franchise of all time with 90 movies released. The large number of movie titles makes it difficult for people to choose which movie to watch. Therefore, a Marvel movie recommendation system is needed using a hybrid item-based and content-based filtering method. The content-based method calculates the similarity between movies by identifying similar Marvel movies based on content such as genre, actor, director, and synopsis. Meanwhile, item-based completes content-based recommendations by considering user preferences. The reason for using the hybrid item-based and content-based filtering method is to be able to produce more accurate recommendations than a single method. The types and sources of data used are secondary data from journals and the internet (Imdb and Movielens), as well as datasets about Marvel movies. From the results of testing the hybrid model, the precision value is 0.8 or 80% which indicates that the model is accurate. In item-based filtering testing, the similarity result of 0.68 shows good item similarity. In the content-based filtering test, the highest similarity is 0.14 and the lowest similarity is 0.10 which shows that the similarity between the searched content and the generated content is relevant.

Keywords: Recommendation System; Marvel Film; Hybrid Filtering; Item-Based Filtering; Content-Based Filtering

Article Info

Accepted	:	20-06-2024
Revised	:	15-02-2024
Published Online	:	25-06-2025

Correspondence Author:

Daffa Rizki Putra Noordi
Informatics Engineering Study Program,
Universitas Duta Bangsa,
Gawanan, Colomadu, Karanganyar, Indonesia
Email: daffarizki7686@gmail.com

1. INTRODUCTION

In today's era there are countless movie genres available to the public, making it challenging for viewers to choose a movie. This trend coincides with the development of the modern era. The domestic and international film industry continues to produce feature films that would be a shame for today's movie buffs to miss. Film is also an art form and communication medium that uses moving images to tell a story or convey information [11].

Of the many films, there is one film that is very popular in the community, namely "Marvel Films" or often called MCU (Marvel Cinematic Universe). The MCU has become the highest grossing film franchise of all time, with a total of 90 films that have been released from the source [4]. Marvel movies are based on characters and stories from Marvel Comics. The number of viewers in Indonesia is increasing and trending positively after the COVID-19 pandemic, it is reported that cinema audiences rose to reach 14.5% YoY (Year on Year) and reached 114.5 million viewers in 2023 compared to 2022 which touched 100 million based on analysis from box office talk. The number of movie titles that have been aired makes it difficult for people to find which movie to watch [7].

Based on the background of the above problems to overcome these problems, it is necessary to have information about Marvel movies that can facilitate individuals in finding Marvel movies that match their preferences, so a system is needed that can provide marvel movie recommendations. As researchers of previous movie recommendation systems with the content-based filtering method, the consequences of tests directed at reviews involving three members with a total of 4000 movie titles obtained an accuracy value using the mean average precision (MAP) of 0.823254 for single query types and 0.7500556 for multiple seeds query types from these results it is found that single query types produce better recommendations than multiple seeds query types according to [6]. The results of this study indicate that the item-based collaborative filtering method provides recommendation results that are very close to the value preferences given by its users. This is shown in the results of system testing which obtained an accuracy value of 97%. The movie recommendation system on Apache mount uses item-based filtering according to [16]. The movie recommendation system with hybrid filtering and k-nearest neighbor from the results of the trials carried out shows that the KNN algorithm is successfully applied to the film selection recommendation system. Based on the user satisfaction test, it is known that user satisfaction with the recommendation system built reached 82.6%. The results of the reliability test using Cronbach Alpha reached 0.7 so it was concluded that the questionnaire distributed was reliable. Test The validity test conducted also shows that the questionnaire distributed is valid [5]. Movie recommendation System using the result is that the application can provide recommendations to active users by calculating the prediction rating with the condition that several other users have already rated the movie that will be predicted for active users. The author has an idea to create a marvel movie recommendation system using a hybrid item-based and content-based filtering method because it can find similar marvel movies and marvel movies that match user preferences, resulting in more diverse recommendations [1]. The content-based method is used to calculate the similarity between movies by identifying marvel movies that are similar to the user's preferred marvel movies based on content such as (genre, actor, director, synopsis), while item-based complements content-based recommendations by considering user preferences. This modeling integrates two or more recommendation algorithms in one recommendation system so as to reduce the weaknesses or limitations of each recommendation technique [9].

2. RESEARCH METHODS

2.1 Secondary Data

Is data that supports information from various sources. Secondary data for this research is collected through a literature review, namely looking for literature studies for basic information related to this research. Secondary data for this research includes journals and the internet (Imdb and Movielens) to complement the data needed, and datasets about marvel movies.

2.2 Data Collection

At this stage, data from the case study can be collected in the form of documents, data sets, or observations. The MovieLens Dataset and the IMDb website served as the sources for the datasets used in this study. The details of the datasets used in this study are as follows: marvel_movies.csv (Imdb) and user_rating.csv (Movielens). Examples of the contents of the datasets are shown in table 1 and table 2 as below.

.

Tabel 1. Dataset marvel_movies.csv

movieID	title	genre
101	Captain America	Action, Adventure, Sci-Fi, War
102	The Fantastic Four	Action, Adventure, Family, Sci-Fi
103	X-Men	Action, Adventure, Sci-Fi

Tabel 2. Dataset user_rating.csv

userId	title	rating
1	She-Hulk: Attorney at Law	1.22
2	Deadpool & Wolverine	2.69
3	Thor: Ragnarok	4.21

2.3 TF-IDF

Term Frequency (TF) is the greater the frequency of occurrence of a word in a document, the greater the weight value for that word. Inverse Document Frequency (IDF) is the greater the frequency of word occurrence, the smaller the weight value of the word. The formula for determining the TF-IDF value can be found below:

TF- IDF = TF * IDF

TF is the number of times a word appears in a document in TF-IDF.

$$TF = \frac{Number\ of\ occurrences\ of\ a\ word\ (x)}{Number\ of\ words\ in\ the\ document} \quad (1)$$

In TF-IDF, IDF is a computation to determine the occurrence of a word in all documents.

$$IDF = log \frac{Number\ of\ documents}{Number\ of\ occurrences\ of\ a\ word\ (x)} \quad (2)$$

3. RESULTS DAN DISCUSSION

3.1 *Pre-Processing* Data

Pre-processing is the stage of selecting raw data that will be processed in each document. This process is important in building an effective recommendation system. Datasets that have been compiled cannot be directly modeled. In this case it needs to include several processescleaning, case folding, stopword removal, tokenization, and lemmatization,

a. *Cleanning*

It is the process of removing punctuation marks.

Tabel 3. Process Results cleaning

no	Input	Result
1	The Fantastic Four	The Fantastic Four
2	The Amazing Spider-Man 2	The Amazing SpiderMan 2
3	Captain America: The Winter Soldier	Captain America The Winter Soldier

b. *Case Folding*

It is the process of converting each capital letter in each data into lowercase letters.

Tabel 4. Process Results Case Folding

no	Input	Result
1	The Fantastic Four	the fantastic four
2	The Amazing SpiderMan 2	the amazing spiderman 2
3	Captain America The Winter Soldier	captain America the winter soldier

c. *Tokenization*

The process of breaking words into smaller, more meaningful words.

Tabel 5. Process Result Tokenisasi

no	Input	Result
1	the fantastic four	['the', 'fantastic', 'four']
2	the amazing spiderman 2	['the', 'amazing', 'spiderman', '2']
3	captain america the winter soldier	['captain', 'america', 'the', 'winter', 'soldier']

d. *Stopword Removal*

The process of removing words that are deemed not to represent something.

Tabel 6. Process Results Stopword Removal

no	Input	Result
1	['the', 'fantastic', 'four']	['fantastic', 'four']
2	['the', 'amazing', 'spiderman', '2']	['amazing', 'spiderman', '2']
3	['captain', 'america', 'the', 'winter', 'soldier']	['captain', 'america', 'winter', 'soldier']

e. *Lemmatization*

Is the process of returning a word to its base form.

Tabel 7. Process Results Lemmatization

no	Input	Result
1	['fantastic', 'four']	['fantastic', 'four']
2	['amazing', 'spiderman', '2']	['amazing', 'spiderman', '2']
3	['captain', 'america', 'winter', 'soldier']	['captain', 'america', 'winter', 'soldier']

3.2 Item-Based Filtering

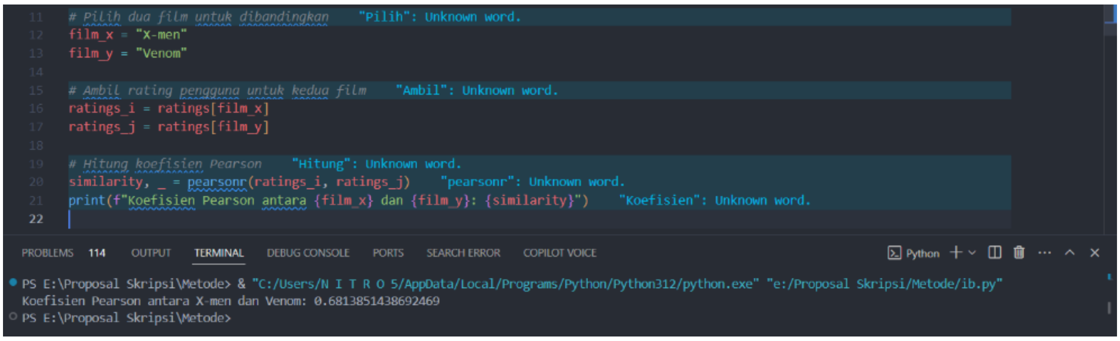
Item-based filtering using Pearson coefficient is a method used in recommendation systems to find similarities between items based on user preferences.

Tabel 8. *Koefisien Person*

userId	X-Men	Venom	Morbius
1	5	4	2
2	3	3	4
3	4	5	3
4	4	4	5
5	2	3	1

Here I will calculate the Pearson coefficient between the X-men movie and Venom with the formula below.

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}} \quad (3)$$



```
# Pilih dua film untuk dibandingkan    "Pilih": Unknown word.
film_x = "X-men"
film_y = "Venom"

# Ambil rating pengguna untuk kedua film    "Ambil": Unknown word.
ratings_i = ratings[film_x]
ratings_j = ratings[film_y]

# Hitung koefisien Pearson    "Hitung": Unknown word.
similarity, _ = pearsonr(ratings_i, ratings_j)    "pearsonr": Unknown word.
print(f"Koefisien Pearson antara {film_x} dan {film_y}: {similarity}")    "Koefisien": Unknown word.
```

```
PS E:\Proposal Skripsi\Metode> & "C:/Users/N I T R O 5/AppData/Local/Programs/Python/Python312/python.exe" "e:/Proposal Skripsi/Metode/ib.py"
Koefisien Pearson antara X-men dan Venom: 0.6813851438692469
PS E:\Proposal Skripsi\Metode>
```

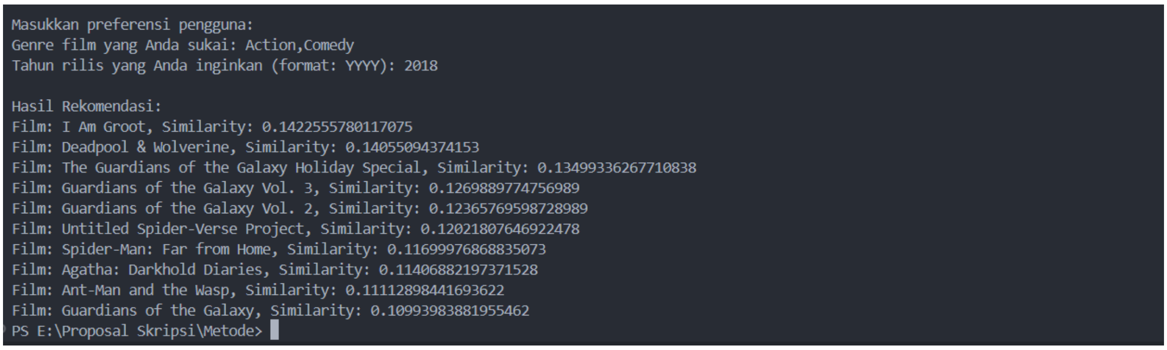
Figure 1. Coefficient between X-men and Venom

.

After calculating using the code above, the result of the coefficient between X-men and Venom is 0.68.

3.3 Content-Based Filtering

Content-based filtering with TF-IDF cosine similarity is used to recommend items to users based on the similarity between item features and user preferences.



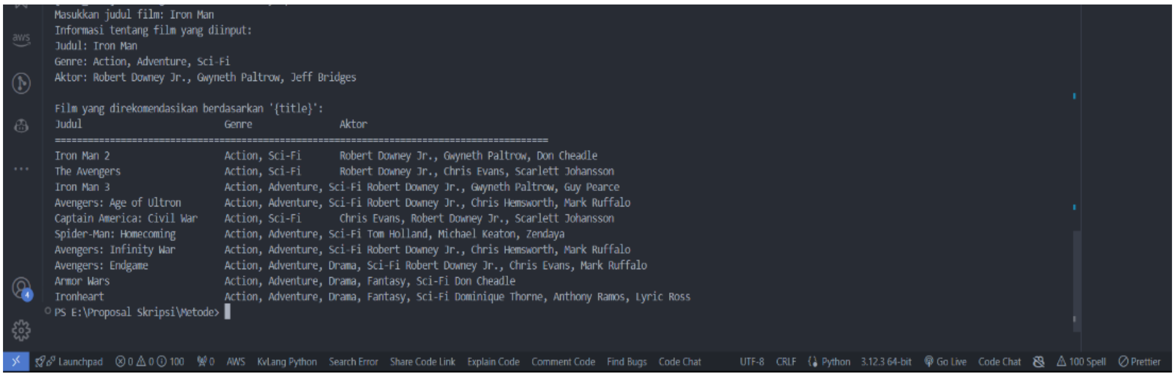
```
Masukkan preferensi pengguna:
Genre film yang Anda sukai: Action,Comedy
Tahun rilis yang Anda inginkan (format: YYYY): 2018

Hasil Rekomendasi:
Film: I Am Groot, Similarity: 0.1422555780117075
Film: Deadpool & Wolverine, Similarity: 0.14055094374153
Film: The Guardians of the Galaxy Holiday Special, Similarity: 0.13499336267710838
Film: Guardians of the Galaxy Vol. 3, Similarity: 0.1269889774756989
Film: Guardians of the Galaxy Vol. 2, Similarity: 0.12365769598728989
Film: Untitled Spider-Verse Project, Similarity: 0.12021807646922478
Film: Spider-Man: Far from Home, Similarity: 0.11699976868835073
Film: Agatha: Darkhold Diaries, Similarity: 0.11406882197371528
Film: Ant-Man and the Wasp, Similarity: 0.11112898441693622
Film: Guardians of the Galaxy, Similarity: 0.10993983881955462
PS E:\Proposal Skripsi\Metode> █
```

Figure 2. Content-based filterin

In the results above, the above recommendations totaled 10, based on genre: Action, Comedy and year: 2018 the movie i am groot gets the highest similarity value with a value of 0.14 and the guardians of the galaxy movie gets the lowest similarity value with a value of 0.10.

3.4 Evaluation Model

The precision value between the recommendation result and the Marvel movie for which the recommendation is sought is calculated during model evaluation. Using hybrid item-based and content-based filtering methods.



```
Masukkan judul film: Iron Man
Informasi tentang film yang diinput:
Judul: Iron Man
Genre: Action, Adventure, Sci-Fi
Aktor: Robert Downey Jr., Gwyneth Paltrow, Jeff Bridges

Film yang direkomendasikan berdasarkan '{title}':
Judul                         Genre               Aktor
=====================================================================================
Iron Man 2                    Action, Sci-Fi      Robert Downey Jr., Gwyneth Paltrow, Don Cheadle
The Avengers                  Action, Sci-Fi      Robert Downey Jr., Chris Evans, Scarlett Johansson
Iron Man 3                    Action, Adventure, Sci-Fi Robert Downey Jr., Gwyneth Paltrow, Guy Pearce
Avengers: Age of Ultron       Action, Adventure, Sci-Fi Robert Downey Jr., Chris Hemsworth, Mark Ruffalo
Captain America: Civil War    Action, Sci-Fi      Chris Evans, Robert Downey Jr., Scarlett Johansson
Spider-Man: Homecoming        Action, Adventure, Sci-Fi Tom Holland, Michael Keaton, Zendaya
Avengers: Infinity War        Action, Adventure, Sci-Fi Robert Downey Jr., Chris Hemsworth, Mark Ruffalo
Avengers: Endgame             Action, Adventure, Drama, Sci-Fi Robert Downey Jr., Chris Evans, Mark Ruffalo
Armor Wars                    Action, Adventure, Drama, Fantasy, Sci-Fi Don Cheadle
Ironheart                     Action, Adventure, Drama, Fantasy, Sci-Fi Dominique Thorne, Anthony Ramos, Lyric Ross
PS E:\Proposal Skripsi\Metode> █
```

Figure 3. Evaluation Model

It is clear from the 10 recommended results in the recommendation results above that there are 8 relevant marvel movies and 2 irrelevant marvel movies. Considered relevant because the title or actor has similarities, irrelevant because it has nothing in common with the title or actor.

$$Presisi = \frac{Number\ of\ relevant\ recommendations}{Total\ number\ of\ recommendations} = Result \tag{4}$$

$$Presisi = \frac{8}{10} = 0.8 \tag{5}$$

So, the precision of the marvel movie recommendation system created is 0.8 or 80%.

4. CONCLUSION

The creation of a marvel movie recommendation system using hybrid item-based and content-based filtering methods has proven effective in improving the personalization and relevance of marvel movie recommendations. By utilizing the advantages of each method. The content-based filtering method is used to calculate similarity between movies by identifying marvel movies that are similar to marvel movies that users like based on content such as (genre, actor, director, synopsis), while item-based complements content-based recommendations by considering user preferences. In the evaluation of the hybrid model produces a precision value of 0.8 or 80%. In the item-based filtering test results obtained a similarity result of 0.68. In the content-based filtering test results, the highest similarity result is 0.14 and the lowest similarity result is 0.10. The limitations of this method rely heavily on the availability of accurate and complete data. If the data used is incomplete or inaccurate, the recommendation results will also be unsatisfactory and Complexity in data processing the hybrid approach requires complex data processing and computational resources, which can be time-consuming and resource-intensive.

REFERENCES

[1]. Agustian, E. R., & Nugroho, E. P. (2020). Sistem Rekomendasi Film Menggunakan Metode Collaborative Filtering dan K-Nearest Neighbors. *JATIKOM: Jurnal Aplikasi Dan Teori Ilmu Komputer*, *3*(1), 18–21.

[2]. Amelia, T., & Pambudi, A. (2023). Rekomendasi Jurusan Kuliah Berdasarkan Minat dan Kemampuan Menggunakan Metode Content Based Filtering. *Technologia: Jurnal Ilmiah*, *14*(3), 245–253.

[3]. Arfisko, H. H., & Wibowo, A. T. (2022). Sistem Rekomendasi Film Menggunakan Metode Hybrid Collaborative Filtering Dan Content-Based Filtering. *EProceedings of Engineering*, *9*(3).

[4]. christiangarciacolon. (2023). *Marvel Films*. Https://Www.Imdb.Com/List/Ls000024621/.

[5]. Ciaputra, A. T., & Hansun, S. (2020). Rekomendasi Pemilihan Film Dengan Hybrid Filtering Dan Knearest Neighbor. *Jurnal Rekayasa Informasi*, *9*(2), 101–109.

[6]. Fajriansyah, M., Adikara, P. P., & Widodo, A. W. (2021). Sistem Rekomendasi Film Menggunakan Content Based Filtering. *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer*, *5*(6), 2188–2199.

[7]. Hadi, I., Santoso, L. W., & Tjondrowiguno, A. N. (2020). Sistem Rekomendasi Film menggunakan User-based Collaborative Filtering dan K-modes Clustering. *Jurnal Infra*, *8*(1), 228–234.

[8]. Hidayat, D., & Putri, A. M. (2022). Mitos dalam Film Gundala (Analisis Monomyth Joseph Campbell). *Desainpedia Journal of Urban Design, Lifestyle & Behaviour*, *1*(2), 50–59.

[9]. Kusuma, A. S., Pratiwi, D. C., & Atina, V. (2023). Sistem Rekomendasi Pemilihan Produk UMKM Berbasis Hybrid Recommendation. *Prosiding Seminar Nasional Teknologi Informasi Dan Bisnis*, 96–105.

[10]. Mukti, K. T., & Mardhiyah, I. (2022). SISTEM REKOMENDASI PEMBELIAN LISENSI FILM MENGGUNAKAN PENDEKATAN HYBRID FILTERING. *Jurnal Riset Sistem Informasi Dan Teknologi Informasi (JURSISTEKNI)*, *4*(3), 127–139.

[11]. Muni, A., & Ihwan, K. (2021). Perangcangan Sistem Informasi Film Berbasis WEB. *JUTI UNISI*, *5*(2), 28–33.

[12]. Nugraha, D., Purboyo, T. W., & Nugrahaeni, R. A. (2021). Sistem Rekomendasi Film Menggunakan Metode User Based Collaborative Filtering. *EProceedings of Engineering*, *8*(5).

[13]. Putraa, I. D. A. C., & Suhartanaa, I. K. G. (n.d.). *Sistem Rekomendasi Anime dengan Metode Content Based Filtering*.

[14]. Putri, M. W., Muchayan, A., & Kamisutara, M. (2020). Sistem Rekomendasi Produk Pena Eksklusif Menggunakan Metode Content-Based Filtering dan TF-IDF. *JOINTECS (Journal of Information Technology and Computer Science)*, *5*(3), 229–236.

[15]. Rizky, M. I., Asror, I., & Murti, Y. R. (2020). Sistem Rekomendasi Program Studi untuk Siswa SMA Sederajat Menggunakan Metode Hybrid Recommendation dengan Content Based Filtering dan Collaborative Filtering. *EProceedings of Engineering*, *7*(1).

[16]. Sari, K. R., Suharso, W., & Azhar, Y. (2020). Pembuatan Sistem Rekomendasi Film dengan Menggunakan Metode Item Based Collaborative Filtering pada Apache Mahout. *Jurnal Repositor*, *2*(6), 767–774.

[17]. Septiani, D., & Isabela, I. (2022). Analisis term frequency inverse document frequency (tf-idf) dalam temu kembali informasi pada dokumen teks. *Sistem Dan Teknologi Informasi Indonesia (SINTESIA)*, *1*(2), 81–88.

.

[18]. Silitonga, P. D. P., & Purba, D. E. R. (2021). Implementasi System Development Life Cycle Pada Rancang Bangun Sistem Pendaftaran Pasien Berbasis Web. *Jurnal Sistem Informasi Kaputama (JSIK)*, *5*(2), 196–203.

[19]. Ula, N., Setianingsih, C., & Nugrahaeni, R. A. (2021). Sistem Rekomendasi Lagu Dengan Metode Content Based Filtering Berbasis Website. *EProceedings of Engineering*, *8*(6).