

# A Study on Dataset of Marvel Movies

Mr. AshaRam Gurjar<sup>1</sup>, Dr. Ruchi Sharma<sup>2</sup>, Anmol Vijay<sup>3</sup>, Abhay Sharma<sup>4</sup>

<sup>1</sup>Assistant Professor Department of Information Technology, Jaipur Engineering College & Research Centre, Jaipur

<sup>2</sup>Professor Department of AI&DS, Jaipur Engineering College & Research Centre, Jaipur

<sup>3</sup>Student Department of Information Technology, Jaipur Engineering College & Research Centre, Jaipur

Email: anmolvijay.it25@gmail.com

<sup>4</sup>Student Department of Information Technology, Jaipur Engineering College & Research Centre, Jaipur

Email : abhaysharma.it25@gmail.com

**Abstract**— Today, there are so many movie genres out there for the general public to view, and it is hard for viewers to select a movie. One of the most favorite movies is the "Marvel Movies" or MCU (Marvel Cinematic Universe), which has become the highest grossing franchise ever with 90 movies released. The numerous movie titles make it hard for individuals to select which movie to watch. Thus, a Marvel movie recommendation system is required using a hybrid item-based and content-based filtering approach. The content-based approach computes similarity between movies by finding similar Marvel movies based on content like genre, actor, director, and synopsis. Item-based fills in content-based recommendations by taking into account user preferences. There as on behind using the hybrid item based and content based filtering approachistobeableto generate more precise recommendations than one approach. The types and sources of data employed are secondary data in journals and the internet (Imdb and Movie lens), and datasets regarding Marvel movies. From the test results of the hybrid model, precision value is 0.8 or 80% which is indicative that the model is precise. In item-based filtering test, the similarity value of 0.68 is representative of good item similarity. In the content-based filtering test, maximum similarity is 0.14 and minimum similarity is 0.10 which indicates that the similarity of the searched content and the generated content is relevant.

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

## I. 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. 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.

## II. 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 Movelens) 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.

| 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      | Star Trek: 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:

$$TF \cdot IDF = TF * IDF$$

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

$$TF = \frac{\text{number of times a word appears in a document}}{\text{total number of words in the document}} \quad (1)$$

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

$$IDF = \log \frac{\text{total number of documents}}{\text{number of documents containing the word}} \quad (2)$$

## III. INDENTATIONS AND EQUATIONS

### 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 processes cleaning, case folding, stop word removal, tokenization, and lemmatization, a. Cleaning 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

b. Case Folding

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

| 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'] |

### 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.



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.

$$Presisi = \frac{TP}{TP + FP} = \frac{8}{8 + 2} = 0.8$$

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

#### IV. 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.

#### V. REFERENCES

(Use APA Format)

- [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] Philip, E., & Philip, A. (2022). The influence of positive self-affirmation towards Malaysian ESL students at tertiary level of Education. *Journal of Humanities and Education Development*, 4(4), 09-17. doi:10.22161/jhed.4.4.2
- [5] Ciaputra, A. T., & Hansun, S. (2020). Rekomendasi Pemilihan Film Dengan Hybrid Filtering Dan Knearest Neighbor. *Jurnal Rekayasa Informasi*, 9(2), 101–109.