JNATIA Volume 1, Nomor 1, November 2022
Jurnal Nasional Teknologi Informasi dan Aplikasinya

Sistem Rekomendasi Anime dengan Metode Content Based Filtering

I Dewa Agung Cahya Putra[a1], I Ketut Gede Suhartana[a2]

[a]Program Studi Informatika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Udayana
Badung, Bali, Indonesia
[1]dewaa2880@gmail.com
[2]ikg.suhartana2@unud.ac.id

Abstract

Anime is a term for animated films or cartoons produced by the Japanese state. Currently the number of anime in circulation is very large, so anime lovers sometimes struggle to find an anime that suits their tastes. One of the reasons is the limited description and review translated from Japanese into other languages. Making an anime recommendation system with a content based filtering approach that utilizes TF-IDF and cosine similarity. The "genre" feature is used as a recommendation system parameter that will be processed by TF-IDF and cosine similarity. The training data uses data downloaded from Kaggle. Modeling begins by calculating the weight of the genre feature values using TF-IDF and looking for similarity values using cosine similarity. After that, the process carried out is sorting the similarity values on the recommendation system that will display the results of anime recommendations. There is an evaluation of the model, which results in a precision value of 88.1%. Testing the precision value is done again when the model is integrated into the website and gets a value of 72.8%.

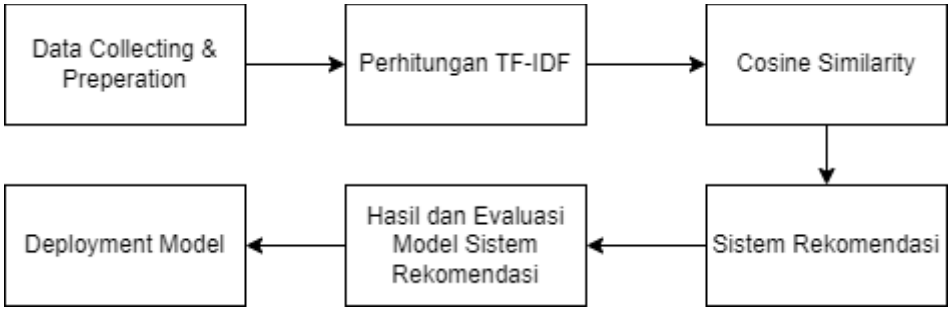
Keywords: *Recommendation System, TF-IDF, Cosine Similarity, Anime, Content Based*

1. Introduction

Anime adalah istilah film animasi atau kartun yang diproduksi oleh negara Jepang [1]. Saat ini jumlah anime yang beredar sangatlah banyak, sehingga para penikmat anime terkadang kesusahan untuk mencari anime yang cocok dengan selera mereka. Salah satu penyebabnya adalah terbatasnya deskripsi dan review yang diterjemahkan darim bahasa Jepang ke dalam bahasa lainnya. Berdasarkan pada masalah tersebut, maka dalam penelitian ini diusulkan sebuah sistem untuk memberi saran kepada para penggemar anime mengenai genre dan judul yang sekiranya cocok untuk mereka. Dari sekian banyaknya anime yang diproduksi membuat calon penonton kesulitan dalam menentukan anime yang akan ditontonnya. Untuk mencari film anime tentunya akan memakan waktu, selain itu anime yang sudah ditentukan untuk ditonton belum tentu sesuai dengan keinginan calon penonton setelah menontonnya, sehingga akan menghabiskan waktu lebih banyak lagi. Menonton anime melalui bioskop, platform penyedia layanan streaming, maupun penyewaan dan pembelian kaset DVD juga diperlukan biaya, akan terbuang sia-sia apabila film yang ditonton tidak sesuai keinginan. Berdasarkan masalah yang telah dijelaskan sebelumnya, penulis mengajukan penelitian dengan judul "Sistem Rekomendasi Anime dengan Metode Content-Based Filtering" dengan menggunakan dataset yang berisi informasi anime (anime.csv). Dataset ini berdasarkan data dari website myanimelist.net. Jumlah data dalam dataset anime terdiri 12294 data dengan kondisi dataset terdapat missing value yaitu pada kolom genre sebanyak 62 data, kolom type sebanyak 25 data, dan kolom rating sebanyak 230 data. Setelah dilakukan cleaning data, jumlah data yang digunakan dari dataset anime sebanyak 12015. Berikut merupakan tautan pengunduhan dari data yang digunakan pada proyek machine learning ini yang terdapat di website Kaggle (https://www.kaggle.com/CooperUnion/anime-recommendations-database) yang diunduh pada hari Jum'at, 23 September 2022 pada pukul 14:21:30. Fitur genre dari judul anime dari dataset yang diberi nilai bobot dengan metode pembobotan TF-IDF. Hasil dari pembobotan akan dicari kemiripannya dengan menggunakan metode cosine similarity dengan menghitung kemiripan fitur pada satu film dengan film lainnya. Perhitungan akan diakhiri dengan menampilkan hasil rekomendasi yang didapatkan oleh model content-based filtering. Metode content-based

filtering menganalisis preferensi dari perilaku pengguna dimasa lalu untuk membuat model. Model tersebut akan dicocokkan dengan serangkaian karakteristik atribut dari barang yang akan direkomendasikan. Barang dengan tingkat kecocokan tertinggi akan menjadi rekomendasi untuk pengguna.

2. Reseach Methods



Gambar 1. Alur Penelitian

2.1 Sistem Rekomendasi

Sistem rekomendasi merupakan program atau sistem penyaringan informasi yang menjadi solusi dalam masalah kelebihan informasi dengan cara menyaring sebagian informasi penting dari banyaknya informasi yang ada dan bersifat dinamis sesuai dengan preferensi, minat, atau perilaku pengguna terhadap suatu barang. Sistem rekomendasi dirancang untuk memahami dan memprediksi preferensi pengguna berdasarkan perilaku pengguna [2]. Terdapat beberapa metode yang dapat digunakan dalam membangun sebuah sistem rekomendasi antara lain *content-based filtering*, *collaborative filtering*, *hybrid filtering*, dan lain sebagainya [3]. Terdapat dua metode pendekatan pada sistem rekomendasi tes [3]:

a. *Content Based Filtering*
   Menggunakan kemiripan antar produk yang akan direkomendasikan dengan produk yang disukai pengguna.
b. *Collaborative Filtering*
   Menggunakan kemiripan kueri dengan item pengguna dengan pengguna lain.

2.2 Content Based Filtering

Sistem rekomendasi dengan metode content-based filtering merekomendasikan item yang mirip dengan item sebelumnya yang disukai atau dipilih oleh pengguna. Kemiripan item dihitung berdasarkan pada fitur-fitur yang ada pada item yang dibandingkan [4]. Metode ini bersifat user independence, tidak bergantung pada situasi apakah item tersebut merupakan item baru (yang belum pernah dipilih oleh pengguna manapun) maupun bukan item baru.

2.3 TF-IDF

TF-IDF adalah salah satu metode yang banyak digunakan dalam ranah *information retrieval* dan *text mining* untuk mengevaluasi hubungan setiap kata atau *term* pada sekumpulan dokumen [5]. Nilai TF-IDF yang tinggi bagi suatu kata menandakan bahwa kata tersebut terdapat pada sedikit dokumen namun dalam frekuensi yang tinggi sehingga dapat digunakan untuk mengetahui kata yang penting dari suatu dokumen. Berikut adalah rumus untuk menghitung nilai TD-IDF:

$$TF - IDF = TF * IDF \qquad (1)$$

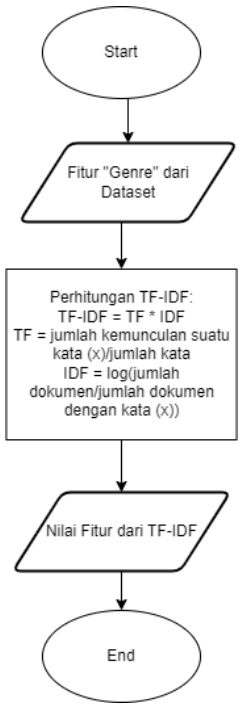
Pada TF-IDF, TF merupakan jumlah kemunculan suatu kata pada suatu dokume

$$TF = \frac{jumlah\ kemunculan\ suatu\ kata\ (x)}{jumlah\ kata\ dalam\ dokumen} \qquad (2)$$

Sedangkan, IDF merupakan perhitungan untuk mengetahui kemunculan suatu kata pada semua dokumen yang digunakan pada penelitian. Hal ini dapat menandakan pentingnya suatu kata bagi suatu dokumen karena sedikitnya kemunculan kata tersebut pada dokumen lainnya. Semakin besar nilai IDF, maka kata tersebut merupakan kata yang sangat penting bagi suatu dokumen.

$$IDF = \log \frac{jumlah\ dokumen}{jumlah\ dokumen\ dengan\ kata\ (x)} \quad (3)$$

Ilustrasi dari proses TF-IDF:



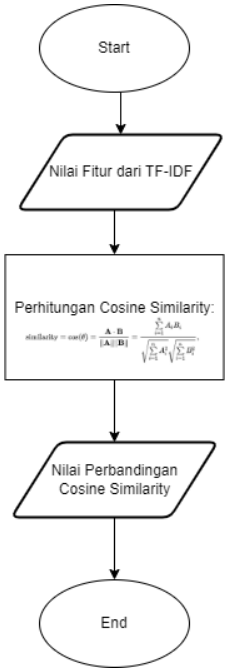
Gambar 2. Flowchart TF-IDF

2.4 *Cosine Similarity*

Cosine Similarity adalah salah satu metode pengukuran nilai kemiripan antar dua dokumen yang berbeda dengan menghitung kosinus sudut yang terbentuk oleh vektor yang merepresentasikan masing-masing dokumen [6].

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}, \quad (4)$$

Ilustrasi dari proses Cosine Similarity:



Gambar 3. Flowchart Cosine Similarity

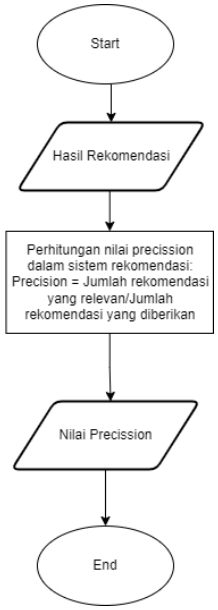
2.5 Pengujian Model

Penerapan metrik precision dilakukan setelah model content-based filtering memberikan hasil rekomendasi dan kemudian menghitung nilai presisi rekomendasi dengan rumus:

$$\text{recommender system precision:} \quad P = \frac{\text{\# of our recommendations that are relevant}}{\text{\# of items we recommended}} \tag{5}$$

Precision adalah proporsi jumlah dokumen yang ditemukan dan dianggap relevan untuk kebutuhan si pencari informasi [7].

Ilustrasi Pengujian Model:



Ganbar 4. Flowchart Evaluasi Model

3. Result and Discussion

3.1. Pembuatan Model Sistem Rekomendasi

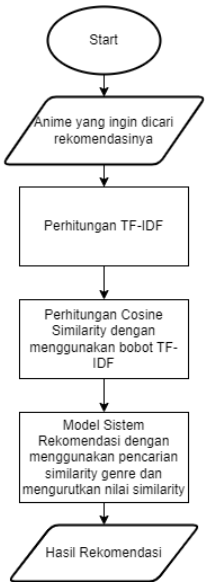
3.1.1. Pengumpulan Data Training

Pada penelitian ini, digunakan dataset yang berisi informasi anime (anime.csv). Dataset ini berdasarkan data dari website myanimelist.net. Jumlah data dalam dataset anime terdiri 12294 data dengan kondisi dataset terdapat missing value yaitu pada kolom genre sebanyak 62 data, kolom *type* sebanyak 25 data, dan kolom *rating* sebanyak 230 data. Setelah dilakukan *cleaning* data, jumlah data yang digunakan dari dataset anime sebanyak 12015. Berikut merupakan tautan pengunduhan dari data yang digunakan pada proyek machine learning ini yang terdapat di website *Kaggle* (https://www.kaggle.com/CooperUnion/anime-recommendations-database) yang diunduh pada hari Jum'at, 23 September 2022 pada pukul 14:21:30.

3.1.2. Data Preperation

a. Data *Cleaning* dilakukan pada data yang bernilai null dalam dataset anime di kolom genre sebanyak 62 data, kolom type sebanyak 25 data, dan kolom rating sebanyak 230 data. Data *cleaning* diperlukan agar dataset memiliki nilai yang valid dan tidak terdapat nilai kosong atau null dalam dataset yang digunakan.
b. *Train-Test-Split* digunakan untuk membagi dataset menjadi data latih (*train*) dan data uji (*test*). Pada proyek ini, data latih (*train*) dibagi menjadi 80% dari dataset dan data uji (*test*) dibagi menjadi 20% dari dataset. Tahapan ini diperlukan karena pembagian dataset diperlukan untuk mempermudah proses evaluasi model, dimana data data train digunakan selama pelatihan model, selanjutnya pada bagian evaluasi, data uji digunakan untuk mengukur kinerja model dengan menggunakan data baru.

3.1.3. Modeling



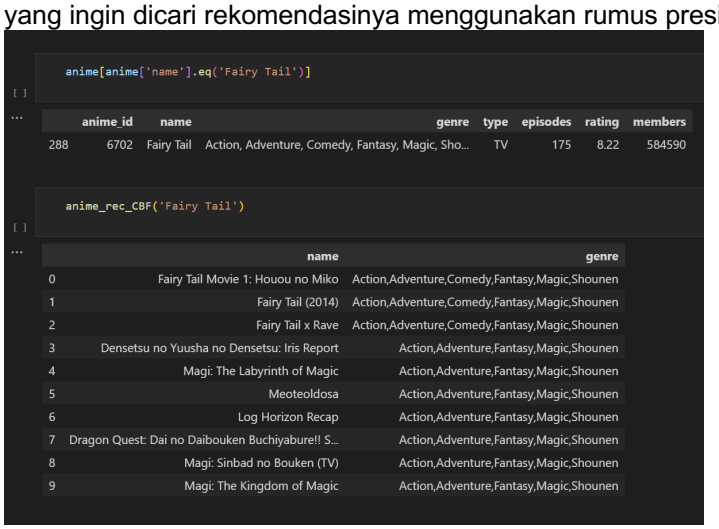
Gambar 5. Flowchart Model Sistem Rekomendasi

Model dari sistem rekomendasi yaitu diawali dengan perhitungan bobot fitur genre dengan menggunakan TF-IDF dengan menggunakan *library* TfidfVectorizer dari *module* sklearn. Setelah itu, bobot tersebut akan dibandingkan dengan *cosine similarity* untuk mencari rekomendasi dengan persamaan nilai bobot dari anime yang dicari dengan anime-anime yang akan direkomendasi. Selanjutnya pembuatan model sistem rekomendasi dirancang dengan melakukan pengurutan nilai similarity yang telah dihitung

sebelumnya. Setelah selesai, hasil rekomendasi berupa list anime dengan nilai similarity tertinggi.

3.1.4. Evaluasi Model

Evaluasi model dilakukan dengan cara menghitung nilai presisi antara hasil rekomendasi dengan anime yang ingin dicari rekomendasinya menggunakan rumus presisi (5).



```
anime[anime['name'].eq('Fairy Tail')]
```

	anime_id	name	genre	type	episodes	rating	members
288	6702	Fairy Tail	Action, Adventure, Comedy, Fantasy, Magic, Sho...	TV	175	8.22	584590

```
anime_rec_CBF('Fairy Tail')
```

	name	genre
0	Fairy Tail Movie 1: Houou no Miko	Action,Adventure,Comedy,Fantasy,Magic,Shounen
1	Fairy Tail (2014)	Action,Adventure,Comedy,Fantasy,Magic,Shounen
2	Fairy Tail x Rave	Action,Adventure,Comedy,Fantasy,Magic,Shounen
3	Densetsu no Yuusha no Densetsu: Iris Report	Action,Adventure,Fantasy,Magic,Shounen
4	Magi: The Labyrinth of Magic	Action,Adventure,Fantasy,Magic,Shounen
5	Meoteoldosa	Action,Adventure,Fantasy,Magic,Shounen
6	Log Horizon Recap	Action,Adventure,Fantasy,Magic,Shounen
7	Dragon Quest: Dai no Daibouken Buchiyabure!! S...	Action,Adventure,Fantasy,Magic,Shounen
8	Magi: Sinbad no Bouken (TV)	Action,Adventure,Fantasy,Magic,Shounen
9	Magi: The Kingdom of Magic	Action,Adventure,Fantasy,Magic,Shounen

Gambar 6. Evaluasi model sistem rekomendasi

Pada hasil sistem rekomendasi diatas yang berjumlah 10 hasil rekomendasi, dapat dilihat terdapat 3 anime yang memiliki genre yang sama persis dengan anime "Fairy Tail", dan 7 anime yang memiliki 5 kesamaan genre dari 6 genre yang terdapat pada anime "Fairy Tail" yang diinputkan ke dalam sistem rekomendasi, sehingga nilai yang diinputkan menjadi 5/6 = 0.83.

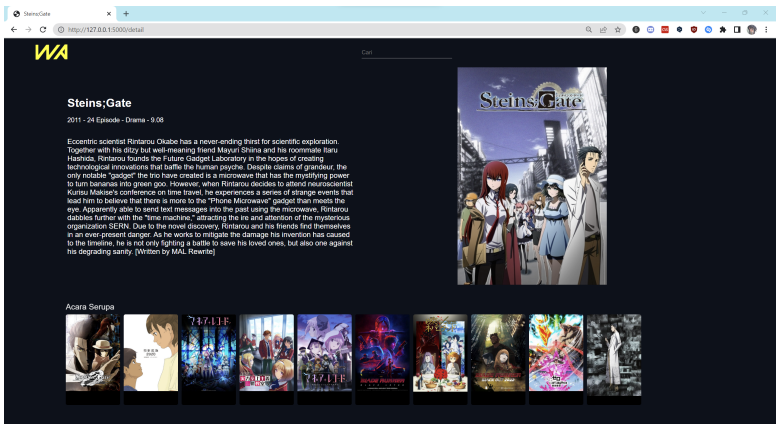
Hasil *Precision*:

* Precision = (1 * 3) + (0.83 * 7)/10
* Precision = 0.881

Jadi, nilai precision dari sistem rekomendasi yang dibuat yaitu 88.1%.

3.2. Model Deployment

Model Deployment dari model sistem rekomendasi yang diintegrasikan ke dalam website. Berikut merupakan hasil integrasi model sistem rekomendasi:



Gambar 7. *Model Deployment* sistem rekomendasi

Deployment sistem rekomendasi ke dalam website menggunakan bantuan *library* dalam bahasa pemrograman python yaitu *pickle* yang berfungsi untuk menyimpan luaran model sistem rekomendasi yang digunakan dalam *deployment* ke dalam website dan *flask* yang berfungsi sebagai *web framework* yang dapat membuat tampilan web lebih terstruktur dan dapat mengatur kinerja web menjadi lebih mudah. Untuk pengujian sistem rekomendasi dalam website menggunakan API yaitu Jikan API yang akan mengambil data anime dari website (myanimelist.net) yang bersifat *open source*. Pengujian model sistem rekomendasi yang sudah diintegrasikan ke dalam website, dalam suatu skenario dengan menghitung nilai precision, sebagai berikut:

anime_id	name	genre
9253	Steins;Gate	Drama,Sci_Fi,Suspense

Gambar 8. Judul Anime yang ingin dicari rekomendasi

Anime “Steins;Gate” yang ditampilkan pada website memiliki genre “Drama, Sci-fi, dan Suspense” dan hasil rekomendasi yang terdapat dalam website sesuai dengan list rekomendasi anime dibawah ini:

name	genre
Steins;Gate 0	Drama,Sci_Fi,Suspense
Nihon Chinbotsu 2020 Gekijou Henshuuban Shizum...	Adventure,Drama,Sci_Fi,Suspense
Magia Record: Mahou Shoujo Madoka☆Magica Gaide...	Drama,Suspense
Youkoso Jitsuryoku Shijou Shugi no Kyoushitsu ...	Drama,Suspense
Magia Record: Mahou Shoujo Madoka☆Magica Gaide...	Drama,Suspense
Blade Runner: Black Lotus	Sci_Fi,Suspense
Yakusoku no Neverland 2nd Season	Sci_Fi,Suspense
Blade Runner: Black Out 2022	Sci_Fi,Suspense
Re:Zero kara Hajimeru Isekai Seikatsu - Hyouke...	Drama,Fantasy,Suspense
Steins;Gate: Kyoukaimenjou no Missing Link - D...	Sci_Fi,Suspense

Gambar 9. Hasil Rekomendasi

Pada hasil sistem rekomendasi diatas yang berjumlah 10 hasil rekomendasi, dapat dilihat terdapat 2 anime yang memiliki ketiga genre yang sama persis dengan anime "Steins;Gate", dan 8 anime yang memiliki 2 kesamaan genre dari 3 genre yang terdapat pada anime "Steins;Gate" yang diinputkan ke dalam sistem rekomendasi, sehingga nilai yang diinputkan menjadi 2/3 = 0.66.

Hasil *Precision*:

* Precision = (2 * 3) + (0.66 * 8)/10
* Precision = 0.728

Jadi, nilai precision dari sistem rekomendasi yang dibuat yaitu 72.8%.

4. Conclusion

Pembuatan sistem rekomendasi anime dengan pendekatan *content based filtering* yang memanfaatkan TF-IDF dan *cosine similarity*. Fitur “*genre”* digunakan sebagai parameter sistem rekomendasi yang akan diproses oleh TF-IDF dan *cosine similarity*. Data *training* menggunakan data yang diunduh dari *Kaggle.* Pembuatan model diawali dengan menghitung bobot nilai fitur *genre* dengan menggunakan TF-IDF dan mencari kesamaan nilai dengan menggunakan *cosine similarity*. Setelah itu, proses yang dilakukan yaitu mengurutkan nilai *similarity* pada sistem rekomendasi yang akan menampilkan hasil rekomendasi anime. Terdapat evaluasi model, yang menghasilkan nilai presisi 88.1%.

Pengujian nilai presisi dilakukan lagi ketika model diintegrasikan ke dalam website dan mendapatkan nilai 72.8%.

References

[1] R. E. Brenner. Understanding manga and anime. Greenwood Publishing Group, 2007

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

[3] Isinkaye, F. O., Folajimi, Y. O. and Ojokoh, B. A. (2015) 'Recommendation systems: Principles, methods and evaluation', Egyptian Informatics Journal. Ministry of Higher Education and Scientific Research, 16(3), pp. 261–273. doi: 10.1016/j.eij.2015.06.005.

[4] Mondi, R. H., Wijayanto, A. and Winarno. (2019) 'Recommendation System With Content-Based Filtering Method For Culinary Tourism in Mangan Application', ITSMART: Jurnal Ilmiah Teknologi dan Informasi. 8(2), pp. 65–72.

[5] Kim, S. and Gil J. (2019). Research Paper Classification Systems Based On TF- IDF and LDA Schemes. Human-centric Computing and Information Sciences. 9(30), pp. 1-21.

[6] Fauzi, M. A., Arifin, A. Z. and Yuniarti, A. (2017) 'Arabic book retrieval using class and book index based term weighting', International Journal of Electrical and Computer Engineering, 7(6), pp. 3705–3710. doi: 10.11591/ijece.v7i6.pp3705-3711.

[7] Lestari, N. P. (2016) 'Uji *Recall* dan *Precision* Sistem Temu Kembali Informasi OPAC Perpustakaan ITS Surabaya'. Universitas Airlangga.