

映画コンテンツ推奨システムに関する研究

S20534 RUMON MD

1. はじめに

推薦システムは、データを使用して、指数関数的に増加する選択肢の中からユーザが探しているものを予測、絞り込み、検索するシステムである。つまり、推薦システムは、データセットに基づいて、本、映画、音楽などの、特定のリソースをユーザに提案するシステムである。

映画レコメンデーションシステムは、何らかのデータセットに基づいてユーザに映画の提案を提供するシステムである。ユーザの好みや興味など、以前に高く評価した映画の属性に基づいて映画を提案し、ユーザがどの映画を好むかを予測する。このような推奨システムは、大量のデータを収集し、可能な限り最良の提案を効果的に提供するサービスにとって有益である。

本研究ではコンテンツベースのフィルタリングシステム手法を用いて、5000本の映画のデータセットを利用した映画コンテンツ推奨システムを試作した。さらに推薦システムに関する関連研究の調査を実施した。

2. データの前処理段階

本研究では図1に示したウェブサイトからデータを収集した。

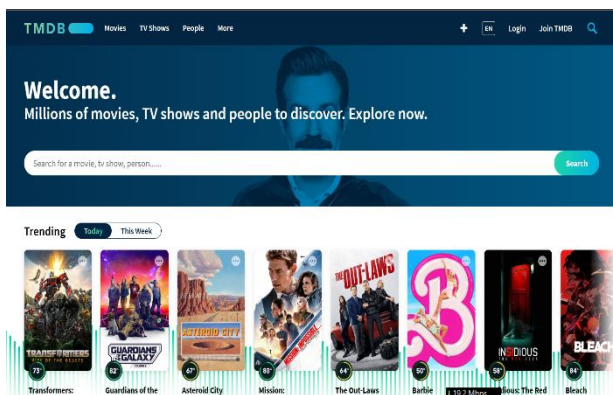


図1: TMDB (The Movie Database)
<https://www.themoviedb.org/>

この段階では収集したメタデータセットから必要なデータを分析し、データの前処理を行う。図2は収集した元データを示しており、23の属性がある。そこから前処理として、映画やジャンル、キーワード、タイトル、監督、出演者など、必要なデータとして7つの属性を選んでデータの前処理を行った(図3)。

```
movies.info()

<class 'pandas.core.frame.DataFrame'>
Int64Index: 4809 entries, 0 to 4808
Data columns (total 23 columns):
#   Column                                Non-Null Count  Dtype
---  -
0   budget                                4809 non-null   int64
1   genres                                4809 non-null   object
2   homepage                               1713 non-null   object
3   id                                     4809 non-null   int64
4   keywords                              4809 non-null   object
5   original_language                    4809 non-null   object
6   original_title                        4809 non-null   object
7   overview                              4806 non-null   object
8   popularity                            4809 non-null   float64
9   production_companies                 4809 non-null   object
10  production_countries                 4809 non-null   object
11  release_date                         4808 non-null   object
12  revenue                              4809 non-null   int64
13  runtime                              4807 non-null   float64
14  spoken_languages                    4809 non-null   object
15  status                               4809 non-null   object
16  tagline                              3965 non-null   object
17  title                                4809 non-null   object
18  vote_average                         4809 non-null   float64
19  vote_count                           4809 non-null   int64
20  movie_id                             4809 non-null   int64
21  cast                                  4809 non-null   object
22  crew                                  4809 non-null   object
dtypes: float64(3), int64(5), object(15)
memory usage: 901.7+ KB
```

図2: 収集した元データ

TMDB (The Movie Database) の5000映画のデータセットは、映画に関する情報が含まれたデータの集合である。このデータセットには属性として、映画の評価、予算、人気度、言語、タイトル、予算、あらすじ、収益、ポスター、監督、出演者、ジャンルなどさまざまな映画の情報が含まれており、このデータセットはCSVフォーマットである。全体としてデータのサイズは50MBである。

また、5000本映画のデータセットから映画のID、タイトル、あらすじ、ジャンル、キーワード、キャスト、クルーの7つの属性を選んだ。映画のIDの属性を使用することで、特定の映画を一意に特定することができる。また、ジャンル情報を使用することで、

特定のジャンルの映画を抽出したり、映画の特徴や傾向を分析したりすることができる。キーワード情報を使用することで、特定のテーマや要素を持つ映画を検索したり、関連する映画を推薦したりすることができる。これらの属性は、映画の基本的な情報や内容、関連要素、制作に関わる人々に焦点を当てており、映画の分析や検索、推薦などの目的に適している。

```
movies.info()

<class 'pandas.core.frame.DataFrame'>
Int64Index: 4809 entries, 0 to 4808
Data columns (total 7 columns):
 #   Column      Non-Null Count  Dtype
---  ---
 0   movie_id    4809 non-null   int64
 1   title       4809 non-null   object
 2   overview    4806 non-null   object
 3   genres      4809 non-null   object
 4   keywords    4809 non-null   object
 5   cast        4809 non-null   object
 6   crew        4809 non-null   object
dtypes: int64(1), object(6)
memory usage: 300.6+ KB
```

図 3 : 前処理後のデータ

3. 関連研究

ミュージックレコメンダーシステムに関する研究がある¹⁾。ミュージックレコメンダーシステムとは、ユーザの最近の視聴履歴から学習し、ユーザが将来的に聞きたいであろう音楽を推奨するシステムである。システムを使用することにより、音楽配信者は、ユーザにとって、かつて聞いたことのある音楽に関連した音楽を配信して提供することができる。タイトル、アーティスト、ジャンルなどの属性間のユークリッド距離を考慮した類似度スコアによって、ユーザの興味のある音楽の推奨を行なった。

機械学習アルゴリズムを用いた映画推薦システムに関する研究ある²⁾。この映画推薦システムでは、協調フィルタリングアプローチに基づいてユーザから提供された情報を使用して分析し、次のユーザに最適な映画を推奨する。そのために、以前のユーザの評価をもとに、機械学習アルゴリズムを利用した。ユーザまたはクライアントに映画をレコメンデーションし、それに対応する評価は、データセット内の他のユーザと共有される。使用したデータセットから類似のユーザを見つけた後、協調フィルタリングアプローチに基づいて、特定のユーザに対してレコメンデーションがお行われる。

機械学習を利用した図書レコメンデーションシ

ステムに関する研究がある³⁾。図書レコメンデーションシステムを使用して、ユーザはウェブ上で利用可能な多数のオプションから書籍を検索して選択できる。システムは、ユーザのニーズに関する説明が与えられた場合に、説明に適合する製品の選択をユーザに提供する。このシステムは、推奨事項を提供するものであり、購入、習慣、レビューなど、以前のユーザーアクティビティに基づいている。このシステムでは、書籍を借りる図書館利用者が評価（5つ星評価）を送信すると、プログラムで高評価の書籍がユーザに表示される。このプログラムでは、書籍をフィルタリングする共有フィルタリングアルゴリズムを使用してユーザの評価と推奨が行われる。

4. 終わりに

本研究では、映画コンテンツ推奨システムに関する研究を行なった。本研究では5000個の映画データを収集して、データの前処理を実施した。データの前処理では映画データベースやメタデータから必要なデータの収集を行なった。今後、前処理ステージから次のデータモデリングステージを実施し、ウェブサイトを作成する。

関連研究として、ミュージックレコメンダーシステム、協調フィルタリングアプローチに基づいた映画推薦、さらに、ユーザのアクティビティに基づいた書籍レコメンデーションシステムの3つについて調査を行なった。

参考文献

- 1) Varsha Verma, Ninad Marathe, Parth Sanghavi, Prashant Nitnaware, "Music Recommendation System Using Machine Learning", International Journal of Scientific Research in Computer Science, Engineering and Information Technology, pp.4-7, Nov 2021.11.
- 2) F. Furtado, A. Singh, "Movie Recommendation System Using Machine Learning", International Journal of Research in Industrial Engineering, pp.8-12, 2020.3.
- 3) S. R. Hiray, Atish Bhosale, Komal Patil, Amruta Gaikwad, Riddhesh Deshmukh, "Book Recommendation System Using Machine Learning" International Journal for Research in Applied Science & Engineering Technology (IJRASET), pp.1981-1983, 2021.12.