Sistema de Recomendação de Filmes

Patrícia Cordeiro Pereira Pampanelli

São Paulo, Brasil

Abstract

Keywords: recommendation, movies, data science, content-based recommender

1. Introdução

Sistemas de recomendação são amplamentes conhecidos e utilizados em diversas aplicações. Estes sistemas tem como objetivo principal indicar produtos/serviços/conteúdos com base em perfis ou características de produtos já conhecidos. Neste sentido, os sistemas de recomendação podem ser divididos em dois tipos: - filtragem colaborativa; - filtragem baseada em conteúdo.

A filtragem colaborativa consiste na utilização de informações provenientes de comportamentos, atividades e preferências e, consequentemente, a análise de similaridade com outros usuários. Tendo como base esta análise comportamental são indicados novos produtos e serviços. A base para estas recomendações é que usuários que têm gostos similares tem grande probabilidade de concordar em recomendações futuras. Em outras palavras, os sistemas de recomendação baseados em filtragem colaborativa são construídos com base no comportamento dos usuários e suas similaridades. Um exemplo clássico deste tipo de recomendação é a sugestão de filmes tendo como base o perfil de cada usuário, os filmes já asssistidos e os ratings fornecidos pelos mesmos.

Os sistemas de recomendação baseados em conteúdo buscam conhecer os produtos e serviços que serão objeto de recomendações. Por exemplo, para recomendação de filmes com base no conteúdo são analisadas as categorias de cada filme, o estilo, os atores e diretores e etc. Neste tipo de recomendação o foco é o conteúdo e suas características. Estes sistemas tem como premissa que o usuário irá gostar de novos produtos e serviços similares ao filmes que receberam altos ratings anteriormente.

Além disso, existem os sistemas híbridos que combinam características das recomendações colaborativas e dos sistemas baseados em conteúdo para obter melhores resultados.

2. Definição do Problema

O problema que será abordado neste projeto é a recomendação de filmes, ou seja, tendo como base um conjunto de características de usuários e/ou filmes recomendar novos itens que ainda não foram assistidos pelos usuários.

Existem duas abordagens principais para a construção de sistemas de recomendação, como descrito na seção anterior. O primeiro tipo, chamado de filtragem colaborativa, demanda o uso de informações provenientes dos usuários. Desta forma, espera-se analisar informações comportamentais dos usuários como os filmes já assistidos, ratings dados pelo usuários e outras informações que permitam analisar similaridade entre os usuários. Já os sistemas de recomendação com base no conteúdo tem como objetivo analisar as caractarísticas dos filmes como, por exemplo, categorias, tags, diretores, elenco, relação entre as categorias e etc. Desta forma, como descrito anteriormente, o foco deste segundo tipo de abordagem está no conhecimento do filme que será recomendado. Uma terceira abordagem pode ainda combinar características das duas abordagens anteriores tendo uma solução mista.

O sistema de recomendação, como apresentado na imagem abaixo (Fig. 1), tem como entrada um identificador único do usuário para o qual serão feitas recomendações de novos filmes. Na segunda etapa é construída a solução de recomendação tendo como base as informações disponíveis no dataset. Estas informações podem ser sobre os filmes e/ou sobre os usuários. A solução pode seguir um dos três tipos de sistema de recomendação: filtragem colaborativa, sistemas baseados em conteúdo ou solução mista.

3. ML-Latest-Small - Dataset

O dataset [ml-latest-small] (http://files.grouplens.org/datasets/movielens/ml-latest-small.zip) que será utilizado neste projeto foi obtido no site [Grouplens] (https://grouplens.org/) mantido pelo Departamento de Ciência da Computação e Engenharia da Universidade de Minnesota, EUA.

Este dataset é composto de informações obtidas no site [Movielens] (https://movielens.org/) que tem como objetivo recomendar filmes para os usuários. Existem duas versões do dataset. A primeira delas é destinada a pesquisas acadêmicas e

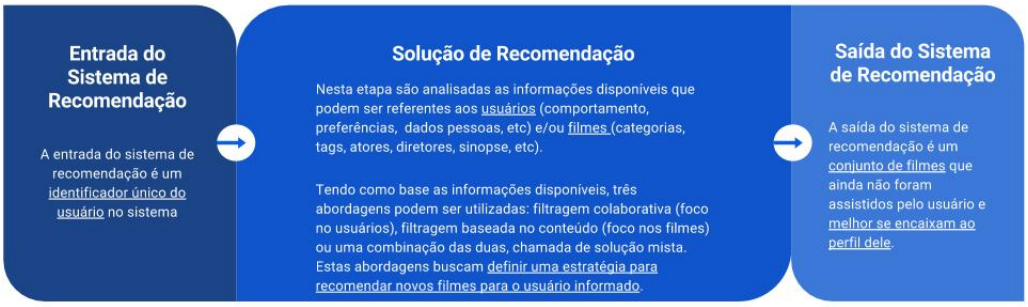

Figura 1: Visão Geral da Solução Proposta

a segunda para desenvolvimento e ensino. Por uma questão de poder computacional, será utilizada a versão reduzida do dataset destinado à ensino. Este é composto de aproximadamente 100 mil ratings, 1300 tags, 9 mil filmes avaliados por 700 usuários.

3.1. Dados Disponíveis

Para formular uma solução para o problema devemos analisar as infoormações disponíveis no dataset a partir do qual o sistema de recomendação será desenvolvido. O conjunto de dados disponíveis neste dataset estão organizados em 4 tabelas (.csv): links, movies, ratings, e tags. As features presentes em cada uma das tabelas são:

3.1.1. Filmes

- Features: movieId, title, genres
- Total de 9125 filmes
- Amostra dos dados:

movieId	title	genres
1	Toy Story (1995)	Adventure
2	Jumanji (1995)	Adventure
3	Grumpier Old Men (1995)	Comedy
4	Waiting to Exhale (1995)	Comedy
5	Father of the Bride Part II (1995)	Comedy

3.1.2. Ratings

- Features: userId, movieId, rating, timestamp;
- Total de 100.004 ratings
- Amostra dos dados:

userId	movieId	rating	timestamp
1	31	2.5	1260759144
1	1029	3.0	1260759179
1	1061	3.0	1260759182
1	1129	2.0	1260759185
1	1172	4.0	1260759205

3.1.3. Tags

- Features: userId, movieId, tag, timestamp;
- Total de 1296 tags
- Amostras dos dados:

userId	movieId	tag	timestamp
15	339	sandra 'boring' bullock	1138537770
1	1029	dentist	1193435061
15	7478	Cambodia	1170560997
15	32892	Russian	1170626366
15	34162	forgettable	1141391765

3.1.4. Links

- Features: movieId, imdbId, tmdbId;
- Total de 9125 links
- Amostra dos dados:

movieId	imdbId	tmdbId
1	114709	862.0
2	113497	8844.0
3	113228	15602.0
4	114885	31357.0
5	113041	11862.0

3.2. Visualização Exploratória

Nesta seção são apresentadas informações extraídas durante a fase exploratória dos dados. Nesta etapa informações como a investigação da correlação entre variáveis e a predominância de alguma característca são buscadas. A fase exploratória é fundamental para o conhecimento aprofundado dos dados e a aplicação da técnica mais apropriada para resolver o problema proposta de recomendação de filmes.

3.2.1. Número de ratings por filme

A primeira avaliação feita diz respeito a distribuição de ratings por filmes. Isso se deve ao fato de que é importante verificar quantos filmes possuem avaliações suficientes para que o sistema de recomendação possa ser construído. Filmes com poucas avaliações são removidos do dataset, pois não apresentam informações suficientes para construção do modelo. Na Figura 2 é possível observar que a maioria dos filmes possuem entre 0 e 1 avaliação.

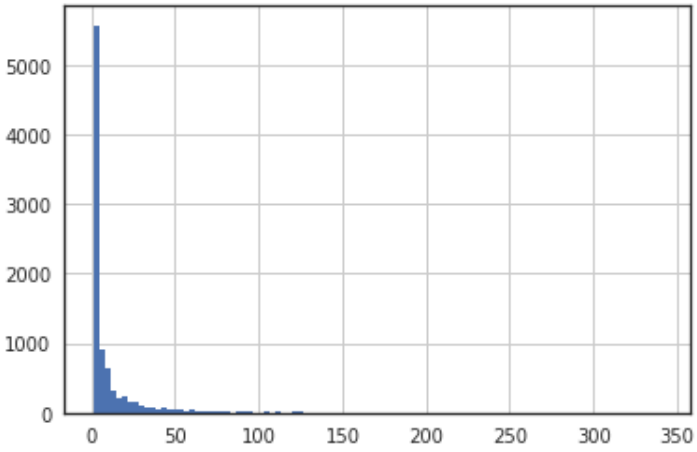

Figura 2: Histograma com a distribuição do número de ratings por filme.

A presença de dados suficientes para construção do modelo é determinante para a decisão de não trabalhar com todos os dados existentes no dataset. Desta forma, na Figura 3 é apresentado o gráfico após a remoção dos filmes que apresentam menos de 4 ratings. É possível observar que muitos dados foram filtrados do dataset. Mais precisamente, foram removidos do dataset 5020 filmes dos 9066 iniciais (equivalente a 55% do dataset). A seguir são apresentadas estatísticas do dataset inicial e do dataset filtrado, respectivamente:

métrica	dataset original	filtrado
número de ratings	9066	4046
média de ratings por filme	11	22
desvio padrão	24	32
número mínimo de ratings	1	4
25%	1	6
50%	3	11
75%	9	25
número máximo de ratings	341	341

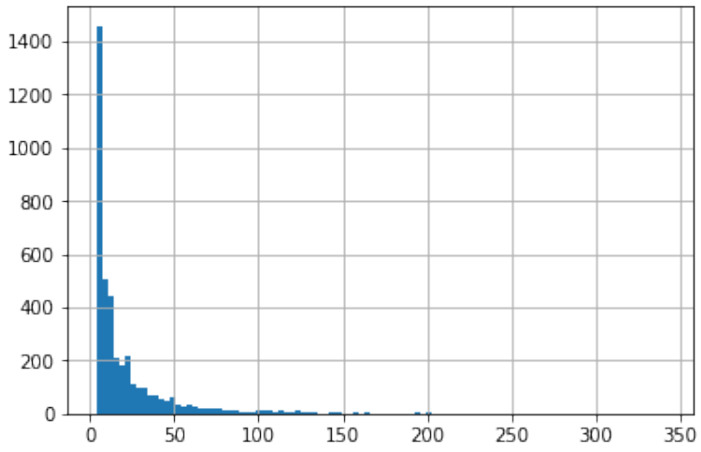

Figura 3: Histograma da distribuição de ratings por filme removendo aqueles que apresentam menos de 4 ratings.

3.2.2. Correlação entre ratings e categorias

Conhecer a correlação entre as variáveis do modelo é fundamental para a construção da solução de recomendação. Desta forma, no Gráfico 4 é

possível observar a correlação entre o rating dado pelo usuário e as categorias dos filmes presentes no dataset. Algumas categorias, como drama e crime possuem uma correlação maior com os ratings que as demais (destacados em vermelho).

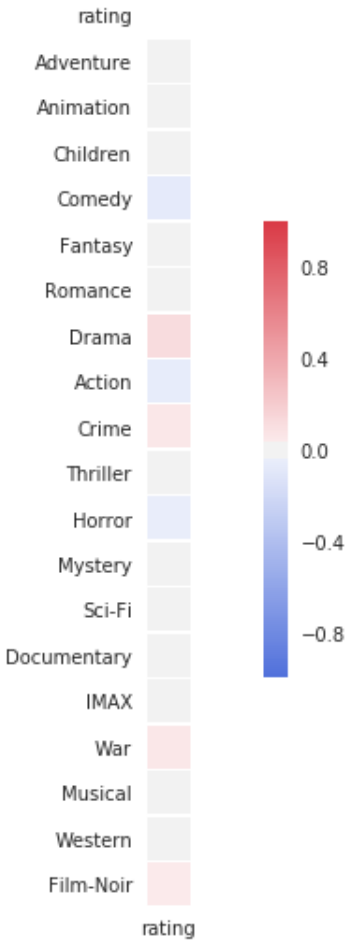

Figura 4: Correlação entre categorias e rating fornecido pelos usuários.

3.2.3. Correlação entre categorias

A correlação entre as categorias também é fundamental para a construção da solução. É espero que exista correlação entre categorias como "Animação"e "Infantil"ou "Ação"e "Aventura". O Gráfico 5 apresenta estas correlações em uma escala de cor entre vermelho (correlação positiva) e azul

(correlação negativa). A maior correlação é apresentada entre filmes infantis e animações e fantasia. Estas categorias claramente estão muito próximas, sendo bastante comum que filmes compartilhem mais de uma delas. Por outro lado, existem filmes com correlação negativa, como drama e ação, que são categorias bem diferentes. Estas correlações entre as categorias são fundamentais para a construção da solução proposta, apresentada na Seção 4.

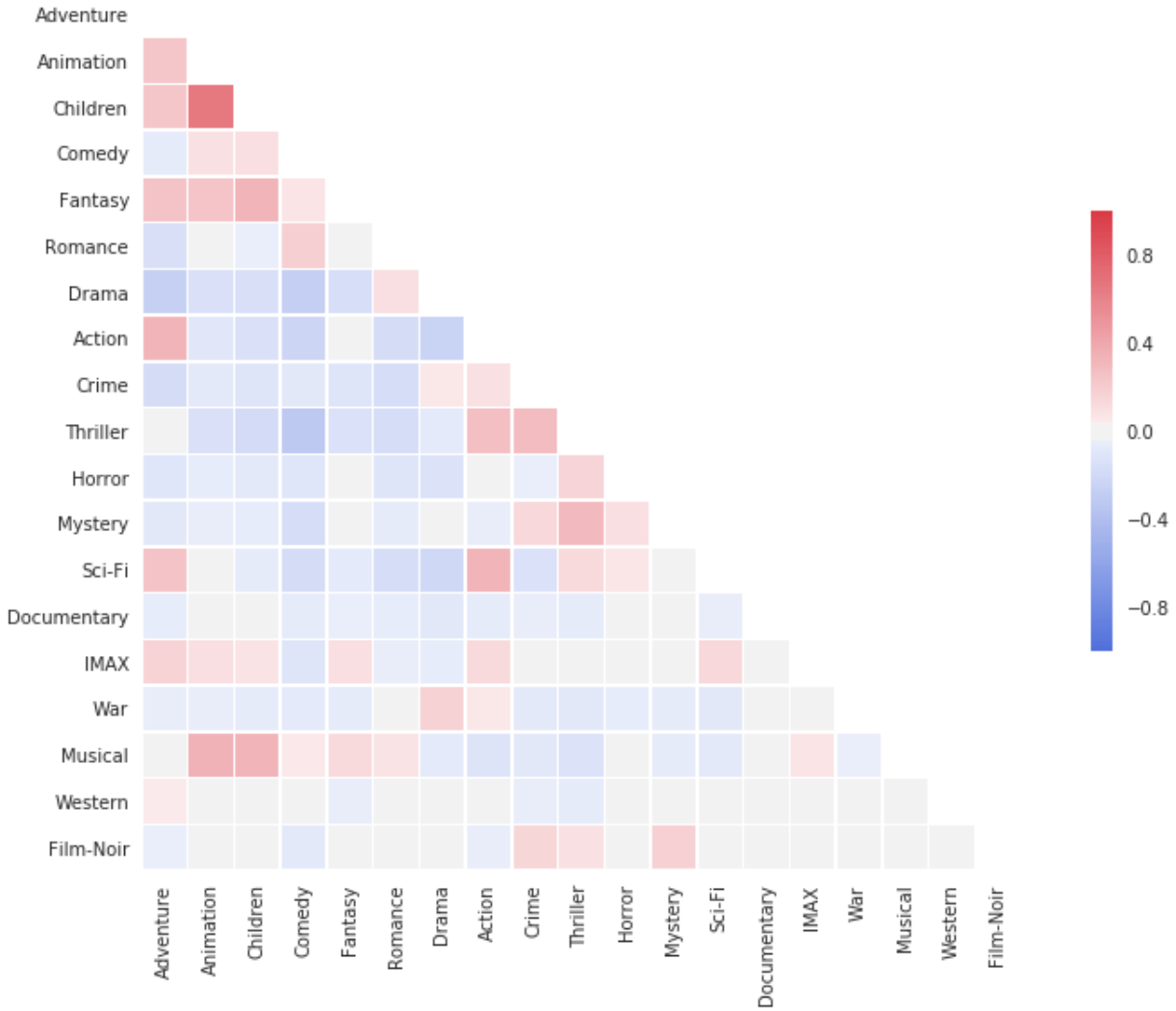

Figura 5: Correlação entre categorias.

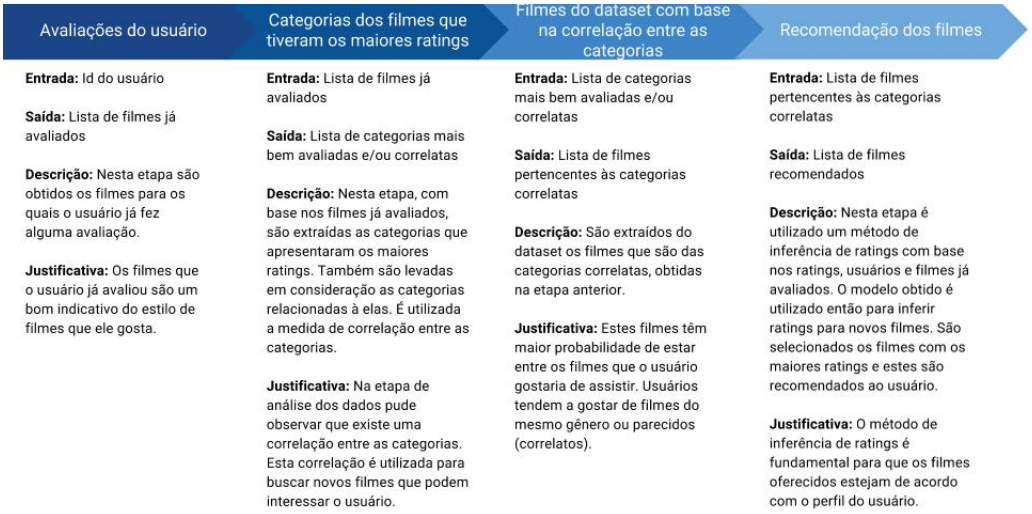

Figura 6: Visão Geral da Solução Proposta

4. Solução Proposta

A solução proposta para este projeto leva em consideração as informações disponíveis no dataset [ml-latest-small] (http://files.grouplens.org/datasets/movielens/ml-latest-small.zip), como apresentado na seção anterior.

O método proposto para o sistema de recomendação de filmes consiste em 4 etapas que compreendem desde a análise dos filmes assistidos pelo usuário até a saída com uma lista de filmes recomendados. A Figura 6 apresenta uma visão geral da solução.

A primeira etapa da solução, chamada de "Avaliação do usuário", consiste em processar a lista de filmes que o usuário já assistiu e avaliou. Esta lista representa um bom indicativo do perfil de filmes que estão dentro das preferências do usuário. Dentre os filmes avaliados, podemos inferir também aqueles estilo que o usuário não gosta, representados pelos ratings mais baixos dados por este usuário.

A segunda etapa consiste em avaliar a lista de filmes obtida na primeira etapa, extraindo as categorias que foram mais bem avaliadas. Além disto, é medida a correlação entre as categorias. As categorias com alta correlação são levadas em consideração também. Esta solução tem como foco as categorias dos filmes e suas correlações. A saída desta segunda etapa corresponde a

uma lista de categorias que repreesentam o perfil do usuário. Vale ressaltar que as correlações entre as categorias devem ser medidas e validadas.

Na terceira etapa são obtidos os filmes que pertencem a estas categorias. São removidos os filmes que já foram assistidos pelo usuário. A justificativa principal para conduzir a busca deste filmes com base nas categorias é que os usuários tendem a gostar de filmes de uma mesmo estilo. Em outras palavras, os usuários tendem a gostar de filmes correlatos.

Na última etapa, para cada um dos filmes obtidos na etapa anterior, é utilizado um método de inferência de ratings. O modelo utilizado para a inferência é treinado utilizando os ratings, filmes e usuários pertencentes ao dataset. Tendo como base os ratings inferidos para o usuário, são selecionados aqueles filmes com a maior pontuação. Por fim, estes filmes são então sugeridos para o usuário de entrada.

5. Resultados

5.1. Métricas

O método proposto será avaliado com base no modelo treinado para inferência de ratings, como apresentado na seção anterior. As métricas utilizadas serão:

5.1.1. Root-mean-square error (RMSE)

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - {y^*}_j)^2}$$

onde: número de elementos:

$$n$$

valor observado:

$$y_j$$

valor estimado:

$${y^*}_j$$

5.1.2. Mean Absolute Error (MAE)

$$MAE = \frac{1}{n} \sum_{j=1}^{n} \left| (y_j - y^*{}_j \right|$$

onde: - número de elementos:

$$n$$

- valor observado:

$$y_j$$

- valor estimado:

$$y^*{}_j$$

5.2. Benchmarks

A avaliação do método proposto é fundamental para verificar quão boas estão sendo feitas as recomendações dos filmes. Existem diversos benchmarks que podem ser utilizados para essa finalidade:

- Surprise Scikit](https://pypi.python.org/pypi/scikit-surprise) - [My Media Lite](http://mymedialite.net/examples/datasets.html)

5.3. Conclusões

O trabalho desenvolvido tem como objetivo a construção, elaboração crítica e avaliação de um sistema de recomendação de filmes. Este tema bastante explorado na literatura foi abordado com foco nos sistemas de recomendação baseados em conteúdo.

Os resultados são avaliados principalmente levando em consideração as métricas: Root-mean-square error (RMSE) e Mean Absolute Error (MAE), apresentadas nas seções anteriores. Estas métricas avaliam a qualidade do modelo construído na quarta etapa do método proposto (Seção 4) para a inferência de ratings.

Existem diversos métodos para inferência de ratings. Neste trabalho foi utilizado o Single-value decomposition (SVD) implementado no pacote python para sistemas de recomendação Surprise ([1]). O dataset foi particionado da seguinte forma: 75% das amostras para treinamento e 25% para teste. Utlizando estas configurações as métricas obtidas foram de: RMSE (1,76) e MAE (1,34).

De forma qualitativa, os filmes recomendados foram considerados coerentes com o dataset de treinamento. É possível observar que as categorias tem

um peso grande na definição dos filmes, representando um ponto que pode ser melhorado em trabalhos futuros. Além disso, outras abordagens podem ser utilizadas para busca de semelhança entre os filmes como, por exemplo, o uso das palavras-chave.

[1] N. Hug, Surprise, a Python library for recommender systems, 2017.