

Modelagem do Sistema RBC para Recomendação de Filmes

1. Introdução

Este documento descreve a modelagem do sistema de Raciocínio Baseado em Casos (RBC) desenvolvido para a recomendação de filmes. O objetivo do sistema é encontrar, numa base de dados de filmes, aqueles que mais se assemelham às preferências e características especificadas pelo utilizador. A modelagem abrange a definição dos atributos que caracterizam um filme, as métricas para comparar esses atributos, os pesos que definem a importância de cada atributo e a forma como a similaridade global é calculada.

2. Definição de Atributos (Representação do Caso)

No sistema, cada "caso" representa um filme. Os filmes são descritos por um conjunto de atributos, que são carregados a partir de um ficheiro CSV (através da função `carregar_base_de_casos_csv`) e armazenados internamente como dicionários Python.

Os atributos definidos para cada filme são:

Atributo	Tipo de Dado (no código)	Valores/Exemplos (no código)	Justificativa da Escolha para o Domínio
titulo	String	Textual (Ex: "Matrix", "O Poderoso Chefão")	Identificador único do filme; usado para exibição, não no cálculo de similaridade.
genero	Categórico (String)	Lista GENEROS_POSSIVEIS (Ex: "Ação", "Comédia", "Ficção Científica")	Principal fator de preferência do utilizador; agrupa filmes com temas semelhantes.
ano_lancamento	Numérico (Inteiro)	Inteiros (Ex: 1999, 2014). Normalização baseada em MIN_ANO (1940) e MAX_ANO (2025).	Influência estilo de produção, tecnologia e relevância temática.
classificacao_etaria	Ordinal (String)	Lista CLASSIFICACOES_PO	Restrição importante para adequação ao

		SSIVEIS (Ex: "Livre", "10", "16"). Ordem gerida por CLASSIFICACAO_MAPA_ORDINAL.	público; a ordem é significativa.
duracao_minutos	Numérico (Inteiro)	Inteiros (Ex: 136, 175). Normalização baseada em MIN_DURACAO (60) e MAX_DURACAO (240).	Reflete preferências por filmes curtos ou longos.
avaliacao_critica	Numérico (Float)	Decimais (Ex: 8.7, 9.2). Normalização baseada em MIN_AVALIACAO (1.0) e MAX_AVALIACAO (10.0).	Indicador comum da qualidade percebida ou aclamação do filme.
tem_sequencia	Booleano (String)	Lista BOOLEANOS_POSSIVEIS ("Sim", "Não")	Informação adicional relevante para utilizadores que procuram franquias ou obras únicas.

Estes atributos foram seleccionados por serem informações comumente utilizadas na descrição, busca e escolha de filmes, e correspondem diretamente às colunas esperadas no ficheiro CSV filmes_base.csv.

3. Métricas de Similaridade Local

Para comparar um novo filme (novo caso) com os filmes na base de dados, são calculadas similaridades para cada atributo individualmente, utilizando as seguintes métricas:

- **Para os atributos genero e tem_sequencia:**
 - **Função no Código:** similaridade_categoria(val1, val2)
 - **Métrica Implementada:** Coincidência Exata.
 - **Cálculo:** Retorna 1.0 se os valores dos atributos forem idênticos; 0.0 caso contrário.
 - **Justificativa:** Para estes atributos, a correspondência exata é o critério mais direto e eficaz.
- **Para os atributos ano_lancamento, duracao_minutos, e avaliacao_critica:**

- **Função no Código:** `similaridade_numerica_normalizada(val1, val2, max_diff)`
- **Métrica Implementada:** Diferença Normalizada Inversa.
- **Cálculo:** $1.0 - (\text{maxdiff} | \text{val1} - \text{val2}|)$
 - `max_diff` é a diferença máxima possível para o atributo (ex: `MAX_ANO - MIN_ANO`).
- **Justificativa:** Permite uma comparação gradual; diferenças numéricas menores resultam em maior similaridade. A normalização assegura que o resultado esteja entre 0 e 1, tornando os atributos comparáveis apesar das suas diferentes escalas.
- **Para o atributo `classificacao_etaria`:**
 - **Função no Código:** `similaridade_ordinal_classificacao(val1_str, val2_str)`
 - **Métrica Implementada:** Similaridade Ordinal (baseada em Diferença Normalizada).
 - **Cálculo:**
 1. As strings de classificação são convertidas para valores numéricos através do `CLASSIFICACAO_MAPA_ORDINAL`.
 2. Aplica-se a fórmula: $1.0 - (\text{MAXDIFFCLASSIFICACAO} | \text{val1num} - \text{val2num}|)$.
 - **Justificativa:** Respeita a ordem intrínseca das classificações (ex: "12 anos" é mais similar a "14 anos" do que a "18 anos").

4. Pesos dos Atributos

O sistema reconhece que diferentes atributos podem ter diferentes graus de importância. Para tal, são utilizados pesos, definidos inicialmente no dicionário `PESOS_PADRAO`:

- `"genero"`: 0.25
- `"ano_lancamento"`: 0.15
- `"classificacao_etaria"`: 0.15
- `"duracao_minutos"`: 0.15
- `"avaliacao_critica"`: 0.20
- `"tem_sequencia"`: 0.10

Justificativa dos Pesos Padrão: Os valores iniciais refletem uma percepção geral da importância de cada atributo. Por exemplo, `genero` (0.25) e `avaliacao_critica` (0.20) são considerados mais influentes do que `tem_sequencia` (0.10).

Flexibilidade: Uma característica fundamental do sistema, implementada na função `obter_caso_entrada_do_usuario`, é a capacidade do utilizador **alterar estes pesos** a cada nova busca, personalizando a relevância de cada atributo de acordo com as suas preferências momentâneas.

5. Métrica de Similaridade Global

A similaridade global entre o filme desejado pelo utilizador e cada filme na base de casos é calculada combinando as similaridades locais ponderadas. Esta lógica está implementada na função `calcular_similaridade_global(caso_novo, caso_base, pesos)`.

Método: Média Ponderada Normalizada.

Funcionamento no Código:

1. Para cada atributo comum ao `caso_novo` e ao `caso_base` (e com `peso > 0`):
 - a. Calcula-se a similaridade local com a função apropriada.
 - b. Esta similaridade é multiplicada pelo peso do atributo.
2. Somam-se todas estas "similaridades ponderadas".
3. O resultado é dividido pela "soma dos pesos que foram efetivamente aplicados" (i.e., pesos dos atributos que existiam em ambos os casos e tinham `peso > 0`).

A fórmula é:

$$\text{Sim}(\text{Novo}, \text{Base}) = \frac{\sum_{i \in \text{AtributosEfetivos}} \text{peso}_i \sum_{i \in \text{Atributos}} (\text{peso}_i \times \text{similaridade}_{\text{local}_i})}{\sum_{i \in \text{AtributosEfetivos}} \text{peso}_i}$$

Isto resulta numa pontuação de similaridade geral entre 0 e 1 (0% a 100%), utilizada para ordenar os filmes da base, do mais similar ao menos similar, em relação à entrada do utilizador. Os resultados são então apresentados pela função `exibir_resultados`.