Uma Análise Comparativa de Modelos Colaborativos, Baseados em Conteúdo e Híbridos em Sistemas de Recomendação de Filmes

Daniel Estevam[1], Fernando Couto[2], Lucas Lima[3], Tito Chen[4], Vinícius Lima[5]
Pontifícia Universidade Católica de Minas Gerais (PUC Minas)
Belo Horizonte, Brasil
daniel.estevam@sga.pucminas.br, fcouto@sga.pucminas.br, lucas.lima.635127@sga.pucminas.br,
tchen@pucminas.br, vinicius.lima.1394935@sga.pucminas.br

***Abstract*—This paper presents the development and comparison of different recommender system approaches applied to the movie domain using the *MovieLens 100k* dataset. User-based collaborative filtering, matrix factorization (SVD), content-based filtering, and a simple hybrid model were implemented. Collaborative models achieved the best performance, especially SVD, which obtained the lowest RMSE and MAE values. On the other hand, the content-based model showed unsatisfactory performance, with *Precision@10* and *Recall@10* equal to zero, highlighting the limitations of the available content features for this approach. The hybrid model proved promising by combining collaborative and content-based information. The results reinforce the effectiveness of collaborative models and emphasize the importance of rich content data for the success of content-based recommendation systems.**

***Index Terms*—Recommender Systems, Collaborative Filtering, Matrix Factorization, Content-Based Filtering, Machine Learning.**

I. Introdução

Com o aumento exponencial da disponibilidade de informações na internet, usuários enfrentam dificuldades crescentes para selecionar conteúdos relevantes em meio a uma vasta quantidade de opções. Nesse contexto, os sistemas de recomendação surgem como ferramentas fundamentais para filtrar informações e personalizar experiências, sendo amplamente utilizados em plataformas de *streaming*, comércio eletrônico e redes sociais [12, 1] .

No domínio do entretenimento, plataformas como Netflix, Amazon Prime e Disney+ dependem fortemente de sistemas de recomendação para sugerir filmes e séries que se alinhem aos gostos individuais dos usuários. Esses sistemas não apenas melhoram a experiência do usuário, como também são cruciais para retenção, engajamento e geração de receita [9, 6].

Os sistemas de recomendação podem ser classificados, de forma geral, em três categorias principais: *filtragem colaborativa*, *filtragem baseada em conteúdo* e *modelos híbridos* [13]. A filtragem colaborativa fundamenta-se no princípio de que usuários que compartilharam preferências no passado tendem a concordar no futuro. Já a filtragem baseada em conteúdo utiliza as características dos itens, como gênero, diretor e elenco (no caso de filmes), para identificar itens similares aos que o usuário avaliou positivamente. Modelos híbridos buscam combinar as vantagens dessas duas abordagens, mitigando limitações como o problema de inicialização (*cold start*) e a esparsidade dos dados [13, 10, 4].

O presente trabalho tem como objetivo desenvolver e comparar diferentes modelos de sistemas de recomendação aplicados ao domínio de filmes. Especificamente, serão implementados e avaliados modelos de filtragem colaborativa baseada em usuários, fatoração de matrizes (SVD), recomendação baseada em conteúdo e um modelo híbrido que combina as abordagens anteriores. A comparação será realizada por meio de métricas quantitativas, visando avaliar a capacidade dos modelos em prever as preferências dos usuários.

A relevância deste estudo reside tanto na compreensão dos fundamentos dos sistemas de recomendação quanto na análise comparativa de suas performances, contribuindo para o desenvolvimento de soluções mais eficientes e aplicáveis em contextos reais. Além disso, o trabalho proporciona uma oportunidade prática de aplicar conceitos de aprendizado de máquina, alinhando-se aos objetivos pedagógicos da disciplina.

II. Método

O desenvolvimento deste trabalho foi estruturado em quatro etapas principais: seleção e preparação dos dados, definição dos modelos de recomendação, implementação dos algoritmos e avaliação dos resultados. Esta seção descreve cada uma dessas etapas, bem como justifica as escolhas metodológicas realizadas.

A. Base de Dados

Para a realização dos experimentos, foi utilizada a base de dados pública *MovieLens 100k*, amplamente utilizada na literatura para estudos relacionados a sistemas de recomendação. Esta base contém 100.000 avaliações realizadas por 943 usuários em 1.682 filmes, com notas variando de 1 a 5. Além das avaliações, estão disponíveis informações sobre os gêneros dos filmes, permitindo a implementação de modelos baseados em conteúdo.

B. Pré-processamento dos Dados

O processo de pré-processamento consistiu na construção de uma matriz de utilidade no formato usuário $\times$ filme, utilizada nos métodos baseados em filtragem colaborativa. Para o modelo baseado em conteúdo, foram extraídas as informações dos gêneros dos filmes, representadas como vetores binários indicando a presença ou ausência de cada gênero.

Os dados foram divididos em conjuntos de treino (80%) e teste (20%), de forma estratificada, garantindo que cada usuário possuísse interações em ambos os conjuntos. Esta abordagem visa simular um ambiente de recomendação realista, onde o sistema precisa prever preferências futuras com base no histórico de interações anteriores.

C. Modelos de Recomendação

Quatro modelos distintos foram implementados e avaliados:

1) Filtragem Colaborativa Baseada em Usuário: Este modelo utiliza o algoritmo de *K-Nearest Neighbors* (KNN) para calcular a similaridade entre usuários, empregando a métrica de similaridade do cosseno. A recomendação é feita com base nas preferências de usuários mais semelhantes, sob a premissa de que usuários que concordaram no passado tendem a concordar no futuro.

2) Fatoração de Matrizes (SVD): O modelo de fatoração de matrizes, especificamente o *Singular Value Decomposition* (SVD), decompõe a matriz de utilidade em fatores latentes que representam características ocultas de usuários e itens. Este modelo é capaz de lidar com dados esparsos de forma eficiente, capturando padrões subjacentes nas interações usuário-filme.

3) Recomendação Baseada em Conteúdo: O modelo baseado em conteúdo utiliza os gêneros dos filmes para calcular a similaridade entre itens, aplicando a métrica de similaridade do cosseno sobre os vetores binários de características. A recomendação é realizada sugerindo filmes com características semelhantes àqueles que o usuário avaliou positivamente.

4) Modelo Híbrido: Foi implementado um modelo híbrido simples, combinando os resultados do modelo de fatoração de matrizes (SVD) e do modelo baseado em conteúdo. A combinação é feita por meio de uma média ponderada dos escores de cada modelo, de acordo com a seguinte equação:

$$\text{Score}_{\text{final}} = \alpha \cdot \text{SVD} + (1 - \alpha) \cdot \text{Conteúdo} \tag{1}$$

Esse valor representa uma pontuação de relevância, que reflete tanto a nota prevista pelo modelo SVD quanto o grau de similaridade do filme em relação aos que o usuário avaliou positivamente no passado. O parâmetro α foi ajustado empiricamente, sendo inicialmente considerado $\alpha = 0.5$ para dar pesos iguais às duas abordagens.

D. Avaliação dos Modelos

Os modelos foram avaliados utilizando métricas quantitativas comumente aplicadas em sistemas de recomendação. Foram utilizadas as métricas de *Root Mean Squared Error* (RMSE) e *Mean Absolute Error* (MAE), que mensuram a diferença entre as avaliações previstas e as avaliações reais dos usuários.

As métricas são calculadas conforme as Equações 2 e 3:

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{r_i} - r_i)^2} \tag{2}$$

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|\hat{r_i} - r_i| \tag{3}$$

onde $\hat{r_i}$ representa a nota prevista, r_i representa a nota real e N é o número total de avaliações no conjunto de teste.

O modelo baseado em conteúdo não prevê notas, então é adequado avaliá-lo utilizando as métricas *Precision@K* e *Recall@K*, que indicam se as recomendações estão realmente alinhadas com as preferências do usuário.

E. Ferramentas e Implementação

A implementação dos modelos foi realizada na linguagem Python, utilizando as bibliotecas `pandas`, `numpy`, `scikit-learn` e `surprise`. As visualizações dos resultados foram produzidas com as bibliotecas `matplotlib` e `seaborn`.

III. Resultados

A seguir, são descritos os resultados do experimento.

A. Descrição do Conjunto de Dados

Os experimentos foram conduzidos utilizando a base de dados *MovieLens 100k*, que contém 100.000 avaliações realizadas por 943 usuários em 1.682 filmes. As avaliações são feitas em uma escala de 1 a 5. Além das avaliações, a base disponibiliza metadados dos filmes, compostos exclusivamente pelos gêneros, representados como variáveis binárias indicando a presença ou ausência de cada gênero para cada filme.

B. Configuração dos Experimentos

O conjunto de dados foi dividido em 80% para treinamento e 20% para teste, de forma estratificada, garantindo que todos os usuários possuam avaliações em ambos os conjuntos. Foram implementados quatro modelos distintos para fins de comparação: filtragem colaborativa baseada em usuários (KNN User-Based), fatoração de matrizes (SVD), recomendação baseada em conteúdo e um modelo híbrido simples que combina os modelos SVD e baseado em conteúdo.

Os modelos colaborativos (KNN e SVD) foram avaliados utilizando as métricas de erro *Root Mean Squared Error* (RMSE) e *Mean Absolute Error* (MAE), enquanto o modelo baseado em conteúdo foi avaliado por meio de métricas de recomendação (*Precision@10* e *Recall@10*).

C. Resultados Obtidos

O modelo de filtragem colaborativa baseada em usuários (KNN User-Based) apresentou um RMSE de 1,0194 e um MAE de 0,8038, conforme as Figuras 1 e 2. Esse desempenho reflete uma capacidade moderada de prever as avaliações dos usuários, embora limitada pela simplicidade do método, que depende exclusivamente da similaridade entre usuários.

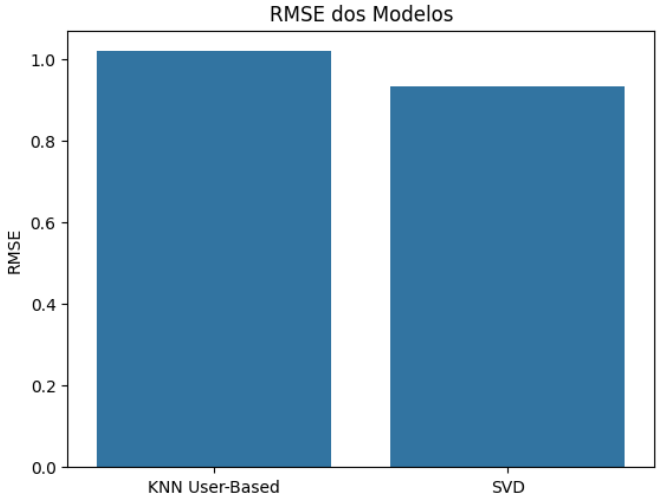

Fig. 1. RMSE dos modelos KNN e SVD

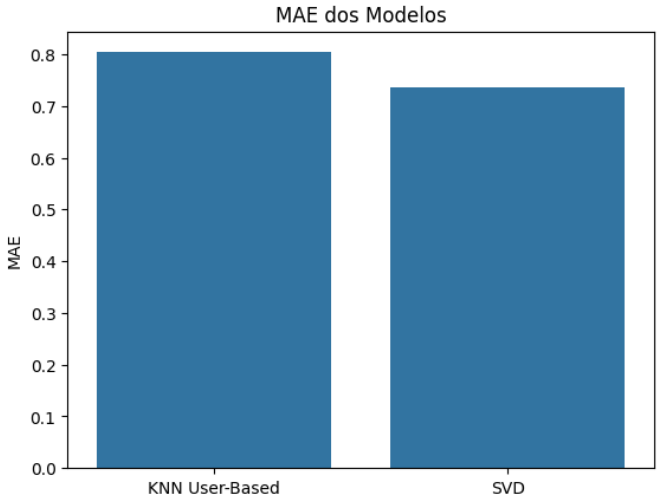

Fig. 2. MAE dos modelos KNN e SVD

O modelo baseado em fatoração de matrizes (SVD) obteve resultados superiores, com RMSE de 0,9341 e MAE de 0,7364, indicando uma melhora consistente na capacidade preditiva em relação ao modelo KNN. Esse desempenho reforça a eficácia da fatoração de matrizes em capturar padrões latentes nas interações entre usuários e filmes, sendo mais robusto frente à esparsidade dos dados.

O modelo baseado em conteúdo, por sua vez, apresentou desempenho insatisfatório nas métricas *Precision@10* e *Recall@10*, ambas com valor médio igual a 0. Esse resultado evidencia uma limitação significativa deste modelo quando aplicado ao conjunto de dados utilizado, cuja representação dos filmes é restrita apenas aos gêneros. A falta de informações mais ricas, como sinopse, elenco ou diretor, compromete diretamente a eficácia deste tipo de abordagem.

Por fim, o modelo híbrido, que combina as predições do modelo SVD com a similaridade baseada em conteúdo, gerou, para um exemplo específico (usuário 100 e filme 50), um score de 1,8576. O resultado relativamente baixo sugere que, para esse usuário, o filme em questão não apresenta uma forte indicação de interesse, seja porque o modelo colaborativo não prevê uma nota alta, seja porque o filme não é suficientemente similar aos filmes previamente apreciados por esse usuário.

De forma geral, os resultados demonstram que métodos baseados em fatoração de matrizes superam abordagens baseadas em vizinhança (KNN), enquanto modelos baseados exclusivamente em conteúdo mostraram-se ineficazes no contexto da base *MovieLens 100k*. Este comportamento evidencia a importância de dados ricos em características para o sucesso de modelos baseados em conteúdo e a robustez dos modelos colaborativos em cenários de dados esparsos.

IV. Conclusões

Os resultados obtidos demonstram que a abordagem baseada em fatoração de matrizes (SVD) apresentou o melhor desempenho nas métricas de erro, superando a filtragem colaborativa baseada em usuários (KNN). Isso evidencia a capacidade dos modelos baseados em fatoração de capturar padrões latentes nas interações usuário-filme, sendo mais eficientes em cenários com alta esparsidade de dados.

Por outro lado, o modelo baseado em conteúdo apresentou desempenho insatisfatório, com valores de *Precision@10* e *Recall@10* igual a zero. Esse resultado reflete uma limitação inerente à escassez de informações descritivas na base de dados utilizada, que possui apenas os gêneros dos filmes como características. Este cenário reforça a importância de dados ricos e diversificados para que modelos baseados em conteúdo possam gerar recomendações efetivas.

O modelo híbrido, embora tenha sido implementado de forma simples e avaliado de maneira pontual, demonstra ser uma abordagem promissora ao combinar os pontos fortes dos modelos colaborativos e de conteúdo. Trabalhos futuros podem explorar o aprimoramento dessa abordagem, utilizando técnicas mais avançadas de combinação, além do enriquecimento dos dados por meio da integração com fontes externas, como APIs do IMDb ou TMDb.

Em suma, os experimentos realizados evidenciam que, no contexto do conjunto de dados *MovieLens 100k*, os modelos colaborativos, especialmente os baseados em fatoração de matrizes, são mais eficazes para o problema de recomendação de filmes. Este trabalho também destaca a relevância da escolha adequada dos dados e dos métodos para o desenvolvimento de sistemas de recomendação eficientes.

References

[1] G. Adomavicius and A. Tuzhilin. "Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions". In: *IEEE Transactions on Knowledge and Data Engineering* 17.6 (2005), pp. 734–749. DOI: 10.1109/TKDE.2005.99.

[2] Gilbert Badaro et al. "A hybrid approach with collaborative filtering for recommender systems". In: *2013 9th International Wireless Communications and Mobile Computing Conference (IWCMC)*. 2013, pp. 349–354. DOI: 10.1109/IWCMC.2013.6583584.

[3] Muhammet cakir, sule gunduz oguducu, and resul tugay. *A Deep Hybrid Model for Recommendation Systems*. 2020. arXiv: 2009.09748 [cs.LG]. URL: https://arxiv.org/abs/2009.09748.

[4] Erion Çano and Maurizio Morisio. "Hybrid recommender systems: A systematic literature review". In: *Intelligent Data Analysis* 21.6 (Nov. 2017), pp. 1487–1524. ISSN: 1571-4128. DOI: 10.3233/ida-163209. URL: http://dx.doi.org/10.3233/IDA-163209.

[5] Rui Chen et al. "A Survey of Collaborative Filtering-Based Recommender Systems: From Traditional Methods to Hybrid Methods Based on Social Networks". In: *IEEE Access* 6 (2018), pp. 64301–64320. DOI: 10.1109/ACCESS.2018.2877208.

[6] Asela Gunawardana and Guy Shani. "A Survey of Accuracy Evaluation Metrics of Recommendation Tasks". In: *Journal of Machine Learning Research* 10.100 (2009), pp. 2935–2962. URL: http://jmlr.org/papers/v10/gunawardana09a.html.

[7] Xiangnan He et al. *Neural Collaborative Filtering*. 2017. arXiv: 1708.05031 [cs.IR]. URL: https://arxiv.org/abs/1708.05031.

[8] Yehuda Koren, Robert Bell, and Chris Volinsky. "Matrix Factorization Techniques for Recommender Systems". In: *Computer* 42.8 (2009), pp. 30–37. DOI: 10.1109/MC.2009.263.

[9] Francesco Ricci, Lior Rokach, and Bracha Shapira. "Recommender Systems Handbook". In: vol. 1-35. Oct. 2010, pp. 1–35. ISBN: 978-0-387-85819-7. DOI: 10.1007/978-0-387-85820-3_1.

[10] Andrew Schein et al. "Methods and Metrics for Cold-Start Recommendations". In: *SIGIR Forum (ACM Special Interest Group on Information Retrieval)*. Aug. 2002, pp. 253–260. DOI: 10.1145/564376.564421.

[11] Ziyuan Xia et al. *Contemporary Recommendation Systems on Big Data and Their Applications: A Survey*. 2024. DOI: https://doi.org/10.1109/ACCESS.2024.3517492. arXiv: 2206.02631 [cs.IR]. URL: https://arxiv.org/abs/2206.02631.

[12] Ziyuan Xia et al. *Contemporary Recommendation Systems on Big Data and Their Applications: A Survey*. 2024. DOI: https://doi.org/10.1109/ACCESS.2024.3517492. arXiv: 2206.02631 [cs.IR]. URL: https://arxiv.org/abs/2206.02631.

[13] Shuai Zhang et al. "Deep Learning Based Recommender System: A Survey and New Perspectives". In: *ACM Computing Surveys* (July 2017). DOI: 10.1145/3285029.