Sistema de Recomendação de Filmes

Edilson Alves de Andrade Junior[1], Marcelino Francisco Gomes das Chagas[1]

[1]Departamento de Computação
Universidade Federal Rural de Pernambuco (UFRPE) – Recife - PE – Brazil
{edilsonalvesjnr, marcelino.francisco.chagas}@gmail.com

Abstract. *The concept of recommendation system has become common in Web applications in the late twentieth century. Due to the expansion of the digital market, companies seek solutions through the IA, to conquer and help users in the process of choice and decision making, offering services and options of choice with a high acceptability by users. In this work we will cover the recommendation system and online streaming of videos.*

Resumo. *O conceito de sistema de recomendação, se tornou comum em aplicações Web no final do século XX. Devido a expansão do mercado digital, empresas buscam soluções através da IA, para conquistar e ajudar usuários no processo de escolha e tomada de decisão, oferecendo serviços e opções de escolha com alta taxa de aceitabilidade pelos usuários. Neste trabalho abordaremos o sistema de recomendação e streaming online de vídeos.*

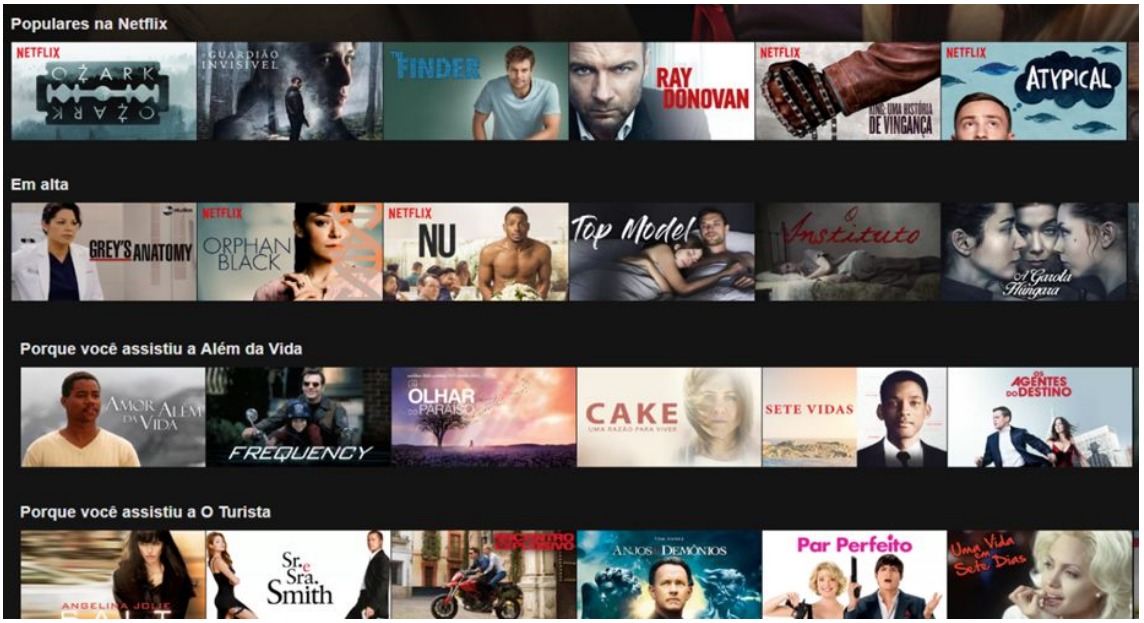
1. Introdução

A crescente variedade de informações disponíveis na *web* e o rápido surgimento de novos serviços de *e-business* proporcionou uma sobrecarga de opções aos usuários. Ter opções é algo bom, mas "infinitas" opções nem sempre é melhor.

Surgem então, os Sistemas de Recomendação (SR), a partir da necessidade de filtrar a quantidade de opções disponíveis para o usuário, automatizando a geração de recomendações baseadas na análise dos dados [0].

O objetivo dos SR é gerar recomendações válidas de itens que possam interessar aos usuários. Sugestões de livros e produtos na Amazon, filmes e séries na Netflix, amigos no Facebook, vídeos no Youtube, lugares no Foursquare e outras infinidades de recomendações. Neste sentido, "item" é um termo geral utilizado para designar o que o sistema recomenda ao usuário, podendo ser, filme, música, vídeo, roupa e até pessoas.

Atualmente, os SR têm um papel fundamental em sites como Amazon, YouTube, Netflix, Yahoo, TripAdvisor, Walmart, Spotify e nas principais Redes Sociais (Facebook, Twitter e Instagram). Neste projeto é abordada a técnica de filtragem colaborativa.



Fonte: Netflix.com - *Screenshot* de Recomendação de Filmes e Séries

2. Materiais

Neste projeto é utilizada a base de dados do MovieLens, um sistema de recomendação baseado na *web* e uma comunidade virtual que recomenda filmes para seus usuários assistirem, com base em suas preferências de filme, usando a filtragem colaborativa das avaliações de filmes e críticas de filmes dos membros. Ele contém cerca de 11 milhões de avaliações para cerca de 8500 filmes [1]. O MovieLens foi criado em 1997 pelo GroupLens Research, um laboratório de pesquisa no Departamento de Ciência da Computação e Engenharia da Universidade de Minnesota [2], com o objetivo de reunir dados de pesquisa sobre recomendações personalizadas [3].

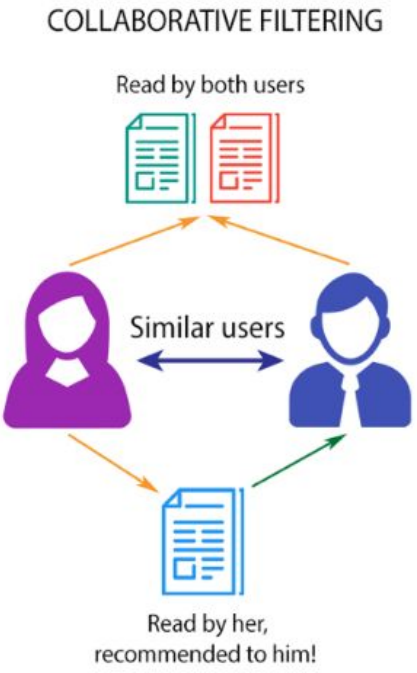
O MovieLens baseia suas recomendações em informações fornecidas pelos usuários do site, como classificações de filmes [2]. O site usa uma variedade de algoritmos de recomendação, incluindo algoritmos de filtragem colaborativa, como item-item [3], usuário-usuário e SVD regularizado [4]. Além disso, para abordar o problema do *cold-start* para novos usuários, o MovieLens usa métodos de elicitação de preferências [5]. O sistema pede que novos usuários avaliem o quanto gostam de assistir a vários grupos de filmes (por exemplo, filmes com humor negro versus comédias românticas). As preferências registradas por esta pesquisa permitem que o sistema faça recomendações iniciais, mesmo antes de o usuário avaliar um grande número de filmes no site.

3. Métodos

3.1 Filtragem Colaborativa (Collaborative Filter)

A abordagem da filtragem colaborativa ignora as características do conteúdo e foca na interação entre o *Usuário X Conteúdo*. Parte do princípio que o sistema não precisa saber as características do conteúdo, mas sim quais conteúdos o usuário consumiu para identificar quais outros usuários tiveram o mesmo comportamento de consumo. Dessa forma é possível "trocar" recomendações entre usuários semelhantes ao processar o coletivo.

Este tipo de abordagem, tira o usuário da bolha de preferência e não necessita da definição de similaridade de conteúdo já que as características são ignoradas.



Este método é com relação à modelagem da "força de interação" do **Usuário** X **Conteúdo** para representar o "quanto o usuário gostou do conteúdo".

Os usuários dariam pontuações pelos filmes já assistidos, onde a forma mais comum de resolver o problema da filtragem colaborativa com Machine Learning é

tentar inferir os valores faltantes em uma matriz de interação, onde (i, j) descreve a pontuação que o usuário deu ao conteúdo.

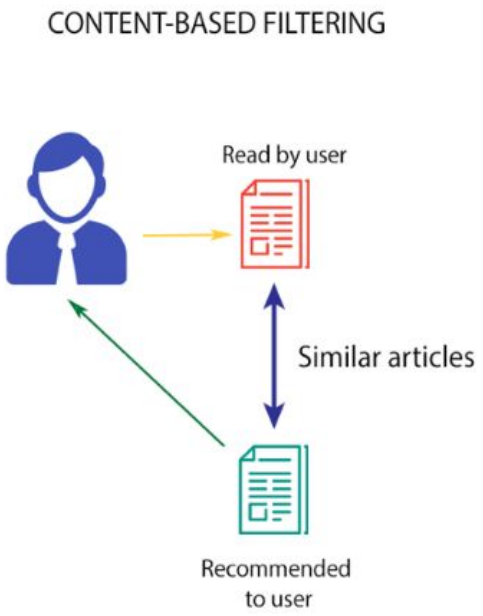
	Freddy x Jason	O Ultimato Bourne	Star Trek	Exterminador do Futuro	Norbit	Star Wars
Ana	2.5	3.5	3.0	3.5	2.5	3.0
Marcos	3.0	3.5	1.5	5.0	3.0	3.5
Pedro	2.5	3.0		3.5		4.0
Claudia		3.5	3.0	4.0	2.5	4.5
Adriano	3.0	4.0	2.0	3.0	2.0	3.0
Janaina	3.0	4.0		5.0	3.5	3.0
Leonardo		4.5		4.0	1.0	

Esse método de filtragem colaborativa também possui suas desvantagens:

- Umas das principais é a própria escala da solução, visto que processar uma matriz (coletivo) de todos os usuários com todos os conteúdos é um desafio computacional. Exemplo disso é pensarmos no catálogo da Netflix e a quantidade de usuários.
- Um detalhe que prejudica ainda mais é que essa matriz é extremamente esparsa, tem muitos valores faltando (poucas avaliações) do que preenchidos, geralmente menos de 3% da matriz é de fato preenchida.
- Necessidade de uma quantidade considerável de registros e *feedback* dos usuários para começar a gerar recomendação.
- Até um usuário novo começar a ter recomendações, do que consumir, ele tem que interagir com muito mais itens do que a filtragem baseada em conteúdo, o que é um problema para plataformas com pouco ou nenhum histórico dos usuários.

3.2 Filtragem Baseada em Conteúdo (Content-Based)

A abordagem da filtragem baseada em conteúdo, depende da similaridade dos itens que estão sendo recomendados. A idéia básica é que, se você gosta de um item, também gostará de um item "*semelhante*".



Geralmente funciona bem quando é fácil determinar o contexto/propriedades de cada item.

Gera as recomendações com base na similaridade do conteúdo já consumido pelo usuário. Ou seja, utiliza os conteúdos que o usuário já consumiu na plataforma (*leu, comprou, assistiu, ouviu, clicou*) para gerar um perfil, então o sistema busca conteúdos semelhantes que ainda não foram vistos pelo usuário e recomenda em seguida.

A principal vantagem dessa categoria é que não requerer muito feedback do usuário para começar a recomendar algo “útil”.

As desvantagens desse método de filtragem:

- O sistema coloca o usuário em uma “*bolha de preferência”*, onde tudo que é recomendado é semelhante ao que já foi consumido. Essa “*bolha de*

preferência" a longo prazo pode levar ao desinteresse do usuário nas recomendações, pois, gera pouca diversidade no conteúdo apresentado. E dependendo do contexto do negócio, pode ser prejudicial, pois não apresentará novos produtos ao usuário e o perfil pode mudar mais lentamente que a própria preferência do usuário.

- Filtragem baseada em conteúdo é realizada ao filtrar conteúdo similar, agora definir as características de similaridade de conteúdo, que pode se tornar um problema, que varia de acordo com ramo de atividade, onde cada contexto tem suas particularidades.

4. Experimentos e Resultados

Os dados analisados foram divididos em dois grupos para realização dos experimentos: `train_data_matrix` e `test_data_matrix`. Um responsável pelo agrupamento de dados usados no treinamento e o outro utilizado na execução dos testes, respectivamente. Por motivos de limitação no processamento dos dados na plataforma Google Colab, apenas 20% dos dados iniciais foram utilizados (`small_data`).

O cálculo para obtenção da matriz de distâncias entre os pares de dados foi realizado através da biblioteca scikit-learn (`sklearn.metrics.pairwise_distances`). Esse método usa uma matriz de vetor ou uma matriz de distância e retorna uma matriz de distância. Se a entrada for um vetor, as distâncias serão calculadas. Se a entrada for uma matriz de distâncias, ela será retornada. Esse método fornece uma maneira segura de obter uma matriz de distância como entrada, preservando a compatibilidade com muitos outros algoritmos que usam um vetor.

Os valores válidos para a métrica são:

- cityblock
- cosine
- euclidiano
- l1
- l2
- manhattan.

Essas métricas suportam entradas de matriz esparsas.

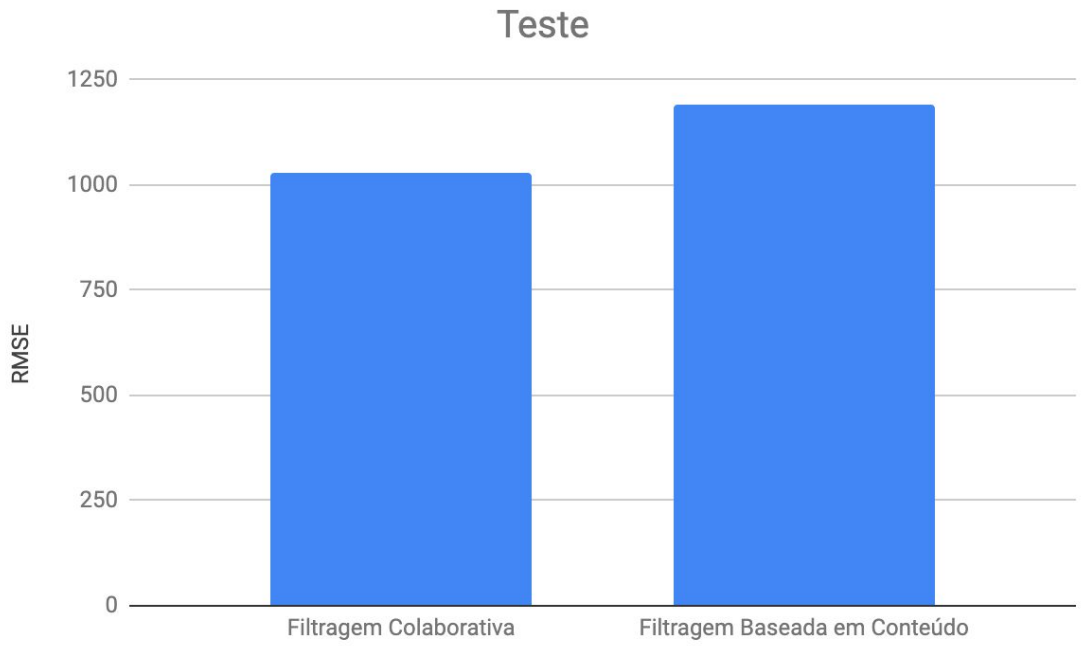
Para cálculo de similaridade entre as amostras foi utilizado a semelhança do cosseno. A semelhança de cosseno, ou o núcleo de cosseno, calcula a similaridade como o produto de ponto normalizado de X e Y:

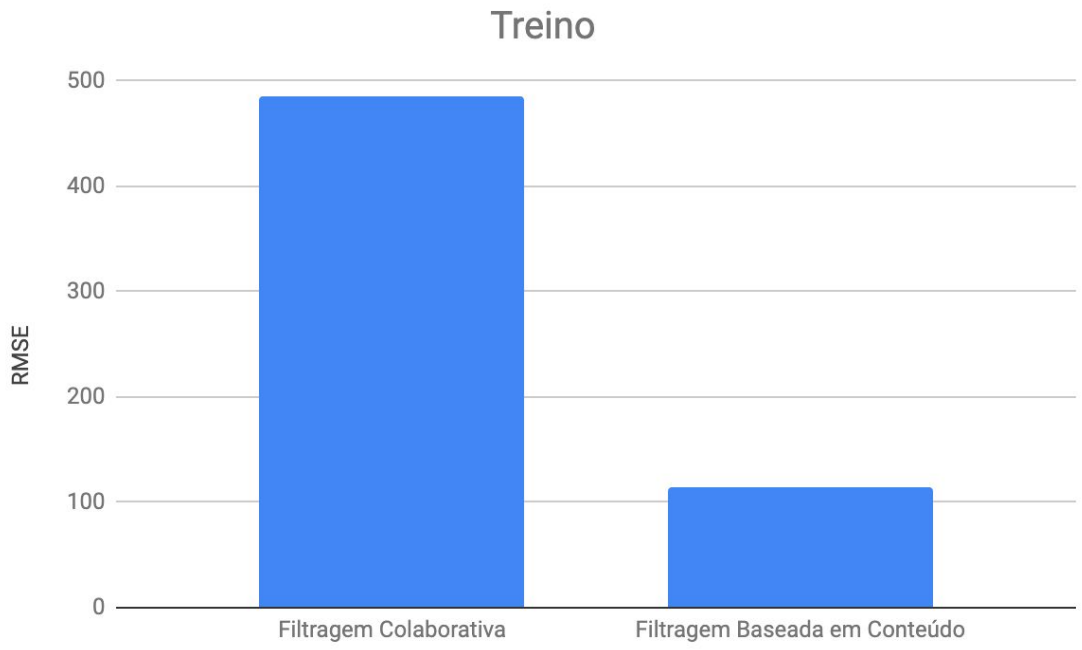
$$K(X, Y) = \langle X, Y \rangle / (\| X \| * \| Y \|)$$

A fórmula de avaliação utilizada para avaliar a precisão dos ratings previstos foi a *Root Mean Squared Error (RMSE)*.

$$RMSE = \sqrt{\frac{1}{N}\sum(x_i - \hat{x}_i)^2}$$

A RMSE foi obtida a partir da raiz quadrada da função *Mean Square Error (MSE)*. Os resultados foram divididos em dois pequenos grupos (Treino e teste) para cada método de recomendação utilizado (Filtragem Colaborativa e Filtragem Baseada em Conteúdo). Ao comparar as duas abordagens pode-se observar que a filtragem colaborativa baseada no usuário oferece um melhor resultado.


Teste
1250
1000
750
500
250
0
RMSE
Filtragem Colaborativa
Filtragem Baseada em Conteúdo


Treino
500
400
300
200
100
0
RMSE
Filtragem Colaborativa
Filtragem Baseada em Conteúdo

5. Conclusão

Sistema de Recomendação é uma tecnologia poderosa cada vez mais presente em soluções empresariais que desejam potencializar seus negócios, promovendo experiências de compra e marketing, por exemplo, diferenciadas para seus clientes. Além disso, sistemas de recomendações ajudam a solucionar o problema de excesso de informação entregue ao cliente final, filtrando proativamente apenas o conteúdo que seja relevante para as pessoas que interagem com essa tecnologia.

Há uma enorme gama de aplicações possíveis para utilização da tecnologia citada, que varia de recomendação de produtos em sites de e-commerce à de pessoas em redes sociais, tornando os sistemas de recomendação algo que deve ser visto de forma especial.

6. Referências Bibliográficas

[0] MELVILLE, P.; SINDHWANI, V. Encyclopedia of machine learning.[s.l.] Springer-Verlag, chapter Recommender systems, 2010.

[1] http://license.umn.edu/technologies/z05173_movielens-database

[2] Schofield, Jack (2003-05-22). "Land of Gnod" . The Guardian . London.

[3] Sarwar, Badrul, et al. "Item-based collaborative filtering recommendation algorithms." Proceedings of the 10th international conference on World Wide Web. ACM, 2001.

[4] Ekstrand, Michael D. Towards Recommender Engineering Tools and Experiments for Identifying Recommender Differences. Diss. UNIVERSITY OF MINNESOTA, 2014.

[5] Chang, Shuo, F. Maxwell Harper, and Loren Terveen. "Using Groups of Items to Bootstrap New Users in Recommender Systems." Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing. ACM, 2015.