UNIVERSIDADE DE SÃO PAULO
INSTITUTO DE CIÊNCIAS MATEMÁTICAS E DE COMPUTAÇÃO

Vinicius de Oliveira Araujo

Um estudo sobre Sistemas de Recomendação

São Carlos

2021

Vinicius de Oliveira Araujo

Um estudo sobre Sistemas de Recomendação

Trabalho de conclusão de curso apresentado ao Centro de Ciências Matemáticas Aplicadas à Indústria do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, como parte dos requisitos para conclusão do MBA em Ciências de Dados.

Área de concentração: Ciências de Dados

Orientador: Prof. Dr. Francisco Aparecido Rodrigues

Versão original

São Carlos

2021

Folha de aprovação em conformidade
com o padrão definido
pela Unidade.
No presente modelo consta como
folhadeaprovacao.pdf

"O grande inimigo do conhecimento não é a ignorância, é a ilusão de ter conhecimento."

Stephen Hawking

RESUMO

ARAUJO, V. **Um estudo sobre Sistemas de Recomendação**. 2021. 49p. Monografia (MBA em Ciências de Dados) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2021.

Os Sistemas de Recomendação - conjunto de técnicas computacionais que auxiliam no processo de escolha dos indivíduos - têm ganhado cada vez mais destaque no contexto moderno e conectado da atualidade. Essas técnicas auxiliam tanto as empresas, que podem vender mais e ofertar melhores serviços, como os próprios clientes, que com a cada vez maior quantidade de opções, conseguem realizar escolhas individuais melhores. Este trabalho objetivou o desenvolvimento de um estudo da teoria por trás das técnicas de construção e avaliação de Sistemas de Recomendação e realizou uma comparação entre cinco abordagens de recomendação diferentes a partir da utilização de um conjunto de dados obtido do site MovieLens durante os anos de 1997 e 1998. Após a criação de um índice de ordenação com o Popularidade, a realização de agrupamento hierárquico com informações demográficas, fatoração de matrizes com Single Value Decomposition e a utilização do índice Adamic Adar na construção de redes complexas, foi possível identificar uma combinação de técnicas que resultassem nos melhores valores para as métricas de avaliação propostas. Por conta dos bons resultados relativos e do equilíbrio apresentado durante o desenvolvimento do trabalho, a melhor abordagem neste cenário foi a híbrida de Adamic Adar Link Prediction com a ordenação proposta por Single Value Decomposition que obteve 72.91% de personalização, 12.68% de cobertura e 35.83% de precisão.

Palavras-chave: Sistema de Recomendação. Netflix. Filmes. Algorítmos de Associação. Redes Complexas. Adamic Adar. Análise Exploratória. Popularidade. Single Value Decomposition.

ABSTRACT

ARAUJO, V. **Machine Learning: Clustering Model in financial sector**. 2021. 49p. Monografia (MBA em Ciências de Dados) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2021.

The Recommender Systems - a set of computational techniques that help in the choosing process by individuals - have been gaining more prominence in today's modern and connected context. These techniques help both companies, which can sell more and offer better services, and customers themselves, who, with the increasing number of options, are able to make better individual choices. This work aimed to develop a study of the theory behind the techniques of construction and evaluation of Recommender Systems and carried out a comparison between five different recommendation approaches using a dataset obtained from the MovieLens website during 1997 and 1998. After creating an ordering index with Popularity, performing hierarchical clustering with demographic information, factoring matrices with Single Value Decomposition and using the Adamic Adar index in the construction of complex networks, it was possible to identify a combination of techniques that resulted in the best values for the proposed evaluation metrics. Due to the good relative results and the balance presented during the development of the work, the best approach in this scenario was the hybrid of Adamic Adar Link Prediction with the ordering proposed by Single Value Decomposition which obtained 72.91% personalization, 12.68% coverage and 35.83% precision.

Keywords: Recommendation System. Netflix. Films. Association Algorithms. Complex networks. Adamic Adar. Exploratory Analysis. Popularity. Single Value Decomposition.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Em um mundo cada vez mais online e expansivo, os Sistemas de Recomendação têm sido amplamente adotados para sugerir serviços, produtos e conteúdos a indivíduos e empresas. Esses sitemas têm a intenção de auxiliar no processo social natural e cotidiano de tomada de decisão dada uma grande e, cada vez maior, variedade de alternativas.

Esses sistemas são uma combinação de várias técnicas computacionais para automatizar e racionalizar a dificuldade das pessoas que frequentemente precisam de mais do que suas próprias experiências para decidirem simplesmente, por exemplo, qual livro escolher, qual filme assistir ou em qual restaurante comer (RESNICK; VARIAN, 1997).

Os Sistemas de Recomendação são também um importante aliado para as companhias, as quais podem entender melhor o comportamento dos seus clientes e ofertar o conteúdo certo para aumentar os indicadores chave que guiam seu negócio. Essa evolução, segundo Tan (2019), de fato só é possível graças a capacidade delas de acumular grandes quantidades de dados de suas operações diárias, realizando, então, recomendações mais emergentes (não triviais) e proporcionando maior credibilidade.

A ideia geral de um Sistema de Recomendação é identificar um item realmente útil ou de interesse para um dado usuário. Para fazer isso, o sistema deve ser capaz de prever essa utilidade para alguns desses diversos itens existentes, ou pelo menos saber comparar a utilidade entre eles e, em seguida, decidir quais itens valem mais a pena serem recomendados com base nessa comparação, maximizando, enfim, a utilidade geral de um conjunto de itens recomendados (RICCI; ROKACH; SHAPIRA, 2011).

Uma das dificuldades iniciais para a evolução das técnicas de recomendação se devia aos conjuntos de dados. A companhia Netflix ajudou a por fim nessa dificuldade ao lançar em 2006 um dos mais significativos concursos sobre Sistemas de Recomendação da história. Com um prêmio de 1 milhão de dólares, este continha um conjunto de dados que possuía 100 milhões de avaliações anônimas de filmes e desafiava a comunidade para desenvolver sistemas que pudessem superar a precisão de seu sistema de recomendação, o Cinematch (BENNETT; LANNING, 2007).

Neste trabalho, tem-se o objetivo de explorar as diferenças e comparar os resultados de algumas das principais técnicas por trás dos Sistemas de Recomendação. Essas técnicas serão aplicadas a um conjunto de dados captados do site MovieLens, que trata também de avaliações de filmes realizadas por usuários anônimos dos Estados Unidos da América entre os anos de 1997 e 1998.

2 REVISÃO BIBLIOGRÁFICA

Este capítulo objetiva apresentar o detalhamento da revisão bibliográfica realizada acerca do desenvolvimento de Sistemas de Recomendação, bem como discutir abordagens envolvidas e métricas necessárias à avaliação do desempenho dos algorítmos que serão propostos.

2.1 Ideia Geral dos Sistemas de Recomendação

Sistemas de recomendação foram inicialmente propostos para lidar com o problema do número excessivo de informação disponível (BOBADILLA et al., 2013). A ideia geral e simplificada é misturar um conjunto de técnicas desenvolvidas com a capacidade de recomendar a um usuário específico os itens mais relevantes entre um conjunto de opções existentes. O sucesso e a importância desses sistemas é claro tanto na academia quanto na indústria (Lü et al., 2012).

Balabanovic e Shoham (1997) propuseram que os Sistemas de Recomendação usualmente podem ser classificados em três grandes categorias: Recomendações Baseadas em Conteúdo, onde um usuário recebe recomendações de itens similares a itens preferidos por ele mesmo no passado; Recomendações Colaborativas, onde um usuário recebe recomendações de itens que pessoas com gostos similares aos dele preferiram no passado; e a terceira categoria que combina as duas primeiras em um algoritmo único de recomendação.

Aggarwal (2016) extende essa definição acrescentando outras duas categorias para Sistemas de Recomendação. Segundo o artigo, os Sistemas de Recomendação baseados em Conhecimento, pautados na obtenção explícita da preferência dos usuários, e o Sistemas de Recomendação Deomográficos, que utilizam as informações descritivas dos usuários para criar grupos de interesse, também têm um papel importânte na construção da teoria.

2.2 Recomendações Baseadas em Conteúdo

Para as Recomendações Baseadas em Conteúdo, segundo Ricci, Rokach e Shapira (2011), o sistema aprende a recomendar itens que são similares àqueles que o usuário gostou no passado. Eles ainda explicam que a similaridade dos itens é calculada com base na boa avaliação e nas características explicativas associadas a eles. Por exemplo, se um usuário classifica positivamente um filme que pertence ao gênero de comédia, o sistema tende a aprender a recomendar outros filmes desse mesmo gênero.

Aggarwal (2016) cita que nesse método, os atributos descritivos dos itens são rotulados pela avaliação recebida pelo usuário e o problema de recomendação se transforma em um problema de classificação ou regressão. Esse modelo específico ao usuário prevê,

de acordo com os parâmetros aprendidos na fase de treinamento, se o correspondente indivíduo irá gostar de algum item ainda não observado por ele.

Basear-se em conteúdo traz algumas vantagens e desvantagens para essa abordagem. Ela se comporta muito bem contra itens que ainda não têm muitas interações no conjunto de dados, uma vez que utiliza apenas a interação do próprio usuário com o item e seus atributos descritivos. Em contrapartida, sofre quando o próprio usuário ainda não possui muitas interações no conjunto de dados, pois isso dificulta o treino do algorítmo e uma consequente má generalização do modelo.

A abordagem por Sistemas de Recomendação Baseados em Conteúdo também tende a realizar recomendações não tão diversificadas, uma vez que utiliza estritamente o histório de um usuário e não leva em consideração as preferências do grupo e do contexto que o usuário está inserido.

2.3 Recomendações Baseadas em Conhecimento

De forma muito similar à apresentada na seção anterior, os Sistemas de Recomendação Baseados em Conhecimento também utilizam informações descritivas de itens no seu processo de recomendação. Entretanto, diferente da outra abordagem, a Baseada em Conhecimento não utiliza o histórico do usuário, mas capta informações externas sobre as suas preferências como, por exemplo, qual o gênero do filme que ele está procurando e quais seriam as palavras-chave que descreveriam sua preferência. Então, o sistema utiliza essas informações para encontrar itens que se aproximem do proposto pelo usuário (AGGARWAL, 2016).

Essa abordagem é particularmente muito útil para contextos em que as interações acontecem de maneira mais rara como, por exemplo, no mercado de compra e venda de imóveis ou automóveis. Nesses casos, geralmente não existem interações suficientes para o treino de modelos de aprendizado de máquina e nem para a construção de vizinhanças de preferências compartilhadas, como se verá nas seções posteriores. Para o bom funcionamento desta abordagem, é comum que na implementação exista conhecimento do domínio ou contexto inserido e uma boa infraestrutura com regras de associação entre itens e bases de conhecimento.

Pela sua proximidade com a abordagem de Recomendação Baseada em Conteúdo, ela também sofre do problema de recomendar itens que podem parecer óbvios para o usuário ou, em outras palavras, exatamente o que foi procurado por ele, existindo pouco ou nenhum grau de surpresa ou diversidade.

2.4 Sistemas de Recomendação Demográficos

Os Sistemas de Recomendação Demográficos recomendam itens baseando-se no perfil demográfico dos usuários. A premissa considerada é a de que diferentes recomendações deveriam ser geradas para diferentes nichos demográficos (RICCI; ROKACH; SHAPIRA, 2011). Esta é uma abordagem bastante utilizadas em marketing, por exemplo, na sugestão de produtos e serviços com base na idade, gênero e localidade dos clientes. Apesar de não entregar os melhores resultados quando implementados de forma sozinha, esses sistemas têm um impacto signficativo em abordagens híbridas, combinando-se com outras técnicas existentes.

2.4.1 Agrupamento Hierárquico

Uma importante aliada para as abordagens baseadas em Recomendação Demográfica é a tarefa descritiva de agrupamento de dados, ou *clustering*. O objetivo dessas tarefas é a criação de grupos específicos de objetos similares, muitas vezes chamados de *clusters*, os quais têm características de possuírem aspectos os mais diferentes possíveis uns dos outros. Existem diversos algorítmos para a execução desta tarefa como os particionais, os baseados em densidade, os hierárquicos, entre outros (TAN; STEINBACH; KUMAR, 2006).

Segundo Lorena et al. (2021), os algorítmos hierárquicos são métodos de construção de grupos que respeitam uma hierarquia aninhada, ou seja, os objetos que são caracterizados como pertencentes a um determinado grupo podem também fazer parte de outros grupos. Além disso, os algorítmos hierárquicos podem ser de dois tipos: aglomerativos ou divisivos. Em aglomerativos, a diferença acontece na inicialização da abordagem, onde os grupos formados são grupos com um único membro. Esses objetos vão sendo reagrupados gradualmente, capturando outros objetos mais próximos. Em divisivos, a inicialização é a com maior número possível de elementos e os grupos vão dividindo-se a cada interação.

Para executar a tarefa principal de minimizar a distância *intra-cluster* dos objetos e maximizar a distância *inter-cluster* dos grupos, é necessária a definição de uma métrica de distância entre os objetos. A mais comum dessas medidas é a Distância Euclidiana, mas outras medidas podem ser escolhidas, como a Distância Cosseno ou a Correlação de Pearson.

$$dist(x_i, x_j) = (|x_{i1} - x_{j1}|^h + |x_{i2} - x_{j2}|^h + ... + |x_{ir} - x_{jr}|^h)^{\frac{1}{h}}, \tag{2.1}$$

onde h é um inteiro positivo e para a distância euclidiana é igual a 2.

Uma outra medida importante que precisa ser definida para a execução da abordagem hierárquica é a distância *inter-cluster*. Existem diversos algorítmos que podem ser utilizados, mas os principais são Single Linkage, Complete Linkage, Average Linkage e

o método de Ward. Neste trabalho, mais adiante, utilizou-se o método de Ward para o desenvolvimento de uma abordagem.

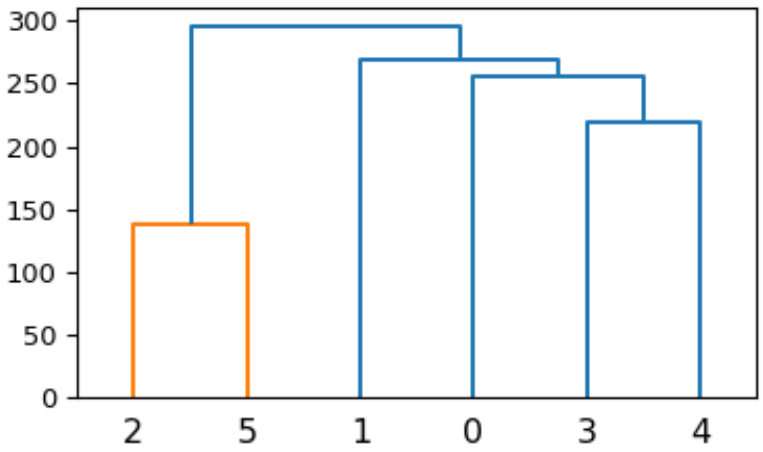
O método de Ward pode ser caracterizado como o aumento no erro quadrático que resulta após a junção de dois grupos. A definição formal do algorítmo segue abaixo:

$$d(C_i \cup C_j, C_k) = \frac{n_i + n_k}{n_i + n_j + n_k} d(C_i, C_k) + \frac{n_j + n_k}{n_i + n_j + n_k} d(C_j, C_k) - \frac{n_k}{n_i + n_j + n_k} d(C_i, C_j)$$

Onde: C_i, C_j e C_k = clusters i, j e k,
n_i, n_j e n_k = tamanho dos clusters i, j e k

Os algorítmos de agrupamento hierárquico geralmente têm uma boa resposta à manipulação de outliers, boa interpretabilidadede e bom desempenho com volumes de dados maiores. A escolha ideal do número de grupos é livre e sua representação pode ser realizada a partir de dendogramas.

Figura 1 – Dendograma representando um agrupamento de dados



Fonte: o autor (2021)

2.5 Recomendações Baseadas em Filtragem Colaborativa

Ao contrário das abordagens Baseadas em Conteúdo, que utilizam o conteúdo de itens previamente avaliados por um usuário u, as abordagens de filtragem colaborativa dependem das avaliações de u e de outros usuários no sistema. A ideia principal é que a avaliação de u para um novo item i é provavelmente semelhante a de outro usuário v, se u e v avaliaram outros itens de maneira semelhante (RICCI; ROKACH; SHAPIRA, 2011).

Seguindo ainda nessa comparação, com a abordagem por Filtragem Colaborativa, alguns itens que receberam menos avaliações, ou ainda, que eram de mais difícil acesso a um usuário em específico, passam a se tornar mais disponíveis, tornando a cobertura maior e o conjunto sugerido mais diverso, isso tudo dado que as recomendações são realizadas com base nos feedbacks de outros usuários.

Os métodos de Sistemas de Recomendação Baseados em Filtragem Colaborativa podem ser divididos em duas grandes categorias: Baseados em Memória e Baseados em Modelo.

2.5.1 Filtragem Colaborativa Baseada em Memória

Em Sistemas de Recomendação por Filtragem Colaborativa Baseada em Memória, as avaliações de itens que foram realizadas pelos usuários e estão armazenadas no sistema são diretamente utilizadas para prever avaliações de novos itens.

Isso pode ser feito de duas maneiras conhecidas como: Recomendação Baseada no Usuário ou Recomendação Baseada no Item. Sistemas Baseados no Usuário avaliam o interesse de um usuário u por um item i utilizando as avaliações para este item por outros usuários, chamados de vizinhos, que têm padrões semelhantes de avaliação. Em outras palavras, assume-se que um usuário deve gostar de itens bem avaliados por usuários com gostos parecidos aos dele.

A abordagem Baseada em Item, por outro lado, prevê a avaliação de um usuário u para um item i com base nas avaliações de u para itens semelhantes a i. Em outras palavras, em vez de medir a semelhança entre os usuários, ela assume que itens semelhantes são avaliados de forma semelhante pelo mesmo usuário.

Em ambas as abordagens, dois itens são semelhantes se vários usuários do sistema classificaram esses itens de maneira semelhante. Os métodos por Filtragem Colaborativa Baseada em Memória têm a vantagem de serem simples e intuitivos e, por serem simples e intuitivos, também têm fácil interpretação. Uma desvantagem dessas aboardagens é o problema da esparsidade dos dados, ou em outras palavras, quando há poucas conexões diferentes e paralelas entre os usuários e itens, pode-se não aproveitar devidamente o conjunto de dados disponível.

2.5.2 Grafos e Redes Complexas

A esparsidade das avaliações observadas causa um grande problema no cálculo de similaridade em métodos Baseados em Memória. Os Grafos são uma abstração poderosa que permite a utilização de muitas técnicas de Redes Complexas para definir vizinhanças e enfrentar os problemas de esparsidade e da busca por similiraridade entre os objetos. Esses Grafos fornecem uma representação estrutural das relações entre vários usuários

e itens e faz uso da transitividade dos nós entre as arestas para inferir recomendações personalizadas (AGGARWAL, 2016).

Nessa abordagem, um Grafo G é definido como sendo não direcionado e bipartido. Seja $G = (N_u \cup N_i, A)$, onde N_u é o conjunto de nós que representam os usuários e N_i é o conjunto de nós que representam os itens. Todas as arestas no Grafo existem apenas entre usuários e itens. Existe uma aresta não direcionada em A entre um usuário i e um item j se, e somente se, o usuário i classificou o item j. Portanto, o número de arestas é igual ao número de entradas observadas na matriz de utilidade.

A ideia de conectividade pode existir, direta ou indiretamente, sempre que exista um caminho possível para a conexão de dois nós. Essa relação de encontrar caminhos para os nós é um problema comum em análise de redes sociais e é relacionada ao desafio do *Link Prediction*, ou predição de conexão. Medidas como *shortest-path* e *Adamic Adar Index* são comumente utilizadas para essas predições.

2.5.2.1 Adamic Adar Link Prediction

Adamic e Adar (2003) definiram a métrica de proximidade de objetos que considera simplesmente o cálculo da quantidade de conexões compartilhadas entre dois nós, dando peso diferente a nós que compartilham de menos conexões. A definição formal segue,

$$Adamic\,Adar(x, y) = \sum_{z \in N(x) \cap N(y)} \frac{1}{log(N(z))}$$

Sejam x e y dois nós pertencentes ao Grafo G definido mais acima. Considere $N(x)$ o conjunto de todos os nós vizinhos ao nó x, analogamente para y. Define-se então que para cada nó z existente entre a vizinhança compartilhada de x e y, se adicione $\frac{1}{log(N(z))}$. Onde $\frac{1}{log(N(z))}$ representa a importância do nó z para a medida:

- Se x e y compartilham um nó z que possui inúmeros nós adjacentes, então esse nó **não** é tão relevante. $N(z)$ é alto, então $\frac{1}{log(N(z))}$ é baixo.

- Se x e y compartilham um nó z que **não** possui inúmeros nós adjacentes, então esse nó é relevante. $N(z)$ **não** é baixo, então $\frac{1}{log(N(z))}$ é alto.

2.5.3 Filtragem Colaborativa Baseada em Modelo

Em contraste com os Sistemas Baseados em Memória, que utilizam as avaliações armazenadas diretamente na previsão, as abordagens Baseadas em Modelo utilizam essas avaliações indiretamente para gerar um modelo preditivo. A ideia geral é modelar as interações usuário-item com fatores que representam características latentes dos usuários e itens no sistema, como a classes de preferências dos usuários e as classes de categorias dos itens.

2.5.3.1 Fatores Latentes e o Single Value Decomposition

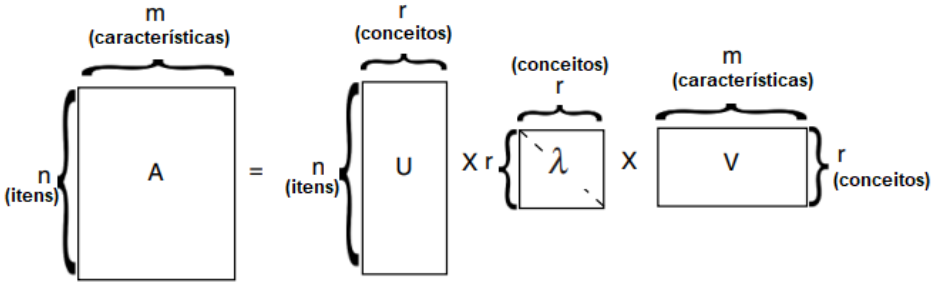
Diversos podem ser os modelos mencionados para a resolução do problema em questão, entretanto, os modelos que têm recebido destaque pela robustês e qualidade para Sistemas de Recomendação são os chamados modelos por Fatores Latentes (KOREN; BELL; VOLINSKY, 2009). Estes envolvem técnicas de fatoração de matrizes para encontrar padrões entre o relacionamento de usuários e os itens, realizando então a sugestão de conteúdo. São importantes também por atuar bem contra os problemas de esparsidade e dimensionalidade usualmente encontrados em modelagens desse tipo.

Uma implementação particular dos Fatores Latentes é a utilização da técnica Single Value Decomposition (SVD). O objetivo chave do SVD é encontrar uma representação mais simplificada do espaço de características representado pela matriz de interação entre os usuários e itens. Cada característica nessa nova representação é uma direção ou um *conceito* e a força de cada um desses conceitos pode ser computada para a realização da predição.

Neste sentido, a ideia da implementação do SVD reside no seguinte teorema: "É sempre possível decompor uma dada matriz A em $A = U\lambda V^T$. Onde A é uma matriz de dados padrão $n\,x\,m$ (n itens, m características) e existe uma matriz U $n\,x\,r$ (n itens, r conceitos), uma matriz diagonal λ $r\,x\,r$ (força de cada conceito), e uma matriz V $m\,x\,r$ (m características, r conceitos)"(RICCI; ROKACH; SHAPIRA, 2011).

A figura abaixo ilustra essa ideia. A matriz diagonal λ contêm os *valores singulares*, que vão ser sempre positivos e ordenados em ordem decrescente. A matriz U é interpretada como a *conversora* dos itens para o seu conceito, enquanto que a matriz V é a conversora da operação *características para conceito*. Até aqui, nesse contexto, pode-se entender que itens e características são análogos a filmes e usuários, e encontrar seus conceitos indicaria encontrar relações não visualmente observáveis entre eles.

Figura 2 – Representação do Teorema Single Value Decomposition



Fonte: Ricci, Rokach e Shapira (adaptado, 2011)

Para computar a SVD em uma matriz retangular A, considere AA^T e A^TA. As colunas de U são autovetores de AA^T, e as colunas de V são os autovetores de A^TA. Os

valores singulares da matriz diagonal λ são as raízes quadradas positivas dos autovalores não nulos de AA^T e A^TA. Portanto, para computar a SVD da matriz A, primeiro calcula-se AA^T, que gera por exemplo a matriz T, depois A^TA que gera por exemplo a matriz D, e por fim computa-se os autovetores e autovalores dessas matrizes recém defindas T e D.

Os r autovalores em λ são ordenados em magnitude decrescente. Portanto, a matriz original A pode ser aproximada simplesmente truncando os autovalores em um determinado k. O SVD truncado cria uma aproximação de ordem k para A de modo que $A_k = U_k\lambda_k V_k^T$. A_k é a matriz de ordem k mais próxima de A. O termo *mais próximo* significa que A_k minimiza a soma dos quadrados das diferenças dos elementos de A e A_k. O SVD truncado é uma representação da estrutura latente subjacente em um espaço *k-dimensional* reduzido, o que geralmente significa que o ruído no novo espaço de características é reduzido.

2.6 Sistemas Híbridos

Os Sistemas de Recomendação Híbridos combinam uma ou mais das abordagens apresentadas acima e tentam, de acordo com Ricci, Rokach e Shapira (2011), utilizar as vantagens de uma para corrigir as desvantagens de outra. Um exemplo clássico pode ser dado pelos métodos de Recomendação Colaborativa, que sofrem de problemas de itens sem histórico, ou seja, eles não conseguem recomendar itens que não têm avaliações suficientes de uma forma muito efetiva. Em contrapartida, isso não limita as abordagens Baseadas em Conteúdo, uma vez que a previsão de novos itens é baseada em sua descrição (características), que normalmente estão mais facilmente disponíveis.

A junção dessas duas abordagens criaria então um modelo mais próximo do ideal, garantindo uma melhora na qualidade das recomendações realizadas aos usuários. Uma possível desvantagem desses sistemas combinados é que a combinação de técnicas pode aumentar a complexidade de implementação consideravelmente.

2.7 Avaliação de Sistemas de Recomendação

Segundo Aggarwal (2016), o desenvolvimento adequado do conjunto de métricas de avaliação de resultado das abordagens é crucial para obter uma compreensão da eficácia dos algoritmos de recomendação. Os Sistemas de Recomendação podem ser avaliados sob dois paradigmas: online e offline.

Os métodos de avaliação do paradigma online contam com uma participação direta dos usuários. Suas reações são medidas de forma instantânea à apresentação dos objetos recomendados e, então, métricas do contexto em que a recomendação se encontra são usualmente utilizadas. Por exemplo, num contexto de recomendação de notícias em uma página da internet, seria possível mensurar a taxa de cliques nas notícias recomendadas e entender a movimentação antes e depois à utilização do novo sistema de recomendação.

Por conta da necessidade da participação online dos usuários, da complexidade de implementação em ambientes produtivos e da dificuldade de escalar com velocidade para diferentes conjuntos de dados, as métricas do paradigma online nem sempre são factíveis de se executar. Uma alternativa são as métricas do paradigma offline.

As métricas offline são as mais utilizadas para a avaliação dos Sistemas de Recomendação. Elas utilizam partes do histórico do conjuntos de dados para simular o ambiente produtivo e, assim, realizar as comparações de eficácia. Frequentemente, uma métrica sozinha não é capaz de captar a eficácia na conversão de Sistemas de Recomendação. O conjunto de métricas utilizadas durante o desenvolvimento deste trabalho foi definido e implementado na Seção 3.5.

3 DESENVOLVIMENTO

O embasamento teórico e as diretrizes de execução técnica deste trabalho já foram discutidas na seção anterior, segue-se agora com fases de diferentes níveis de complexidade para o desenvolvimento do projeto. A saber:

3.1 Metodologia

1. Entendimento da fonte e preparo da base de dados

2. Análise exploratória, para o melhor contextualização da qualidade da informação

3. Aplicação de diferentes abordagens que contribuirão para o Sistema de Recomendação

4. Uso de métricas de avaliação de desempenho para avaliar as diferenças entre as abordagens

3.2 Base de Dados

Como detalhamento da primeira etapa da metodologia, segue-se para a base de dados. O estudo será pautado em cima da base de dados MovieLens 100k Dataset, objeto que representa a coleta de 100,000 avaliações (que variam de 1 a 5) de 943 usuários sobre 1682 filmes do site MovieLens entre 19 de Setembro de 1997 e 22 de April de 1998.

Antes de qualquer processamento, a base era composta por 31 variáveis (colunas) principais, separadas em 3 conjuntos de dados, sendo eles: Itens, Usuários e Interações. O conjunto de Interações armazenava todas as avaliações realizadas entre os usuários e o itens (filmes), continha um total de 100.000 registros (linhas) e 4 variáveis. Já o conjunto de Itens, com um 1682 registros e 22 variáveis, continha todas as informações descritivas relacionadas aos filmes. Por fim, o conjunto de Usuários continha um total de 943 registros e 5 variáveis com informações descritivas relacionadas aos usuários.

As únicas tratativas que foram necessárias inicialmente foram tanto a de identificação de objetos com valores faltantes, que resultou na remoção de uma das linhas do conjunto de dados de Itens, como a de transformação de uma coluna do conjunto de dados de Usuários que continha o equivalente dos Estados Unidos da América ao código CEP brasileiro, identificando o endereço do avaliador do filme.

A transformação em si utilizou uma biblioteca chamada pyzipcode para identificar o estado de origem do endereço, o que seria melhor para se analisar se comparado com a variável original que continha um código com menor granularidade.

A disposição final das colunas por conjunto de dados é dada abaixo:

- Interações

Tabela 1 – Lista de variáveis no conjunto de dados de Interações

Variável	Tipagem	Descrição
id_usuario	string	Identificador do usuário
id_filme	string	Identificador do filme
avaliacao	integer	Avaliação do usuário
data_avaliacao	datetime	Data da avaliação

- Itens

Tabela 2 – Lista de variáveis no conjunto de dados de Itens

Variável	Tipagem	Descrição
id_filme	string	Identificador do filme
nome_filme	integer	Nome do filme
data_lancamento	datetime	Data de lançamento do filme
tema_desconhecido	boolean	Identificador do tema 'desconhecido'
tema_acao	boolean	Identificador do tema 'ação'
tema_aventura	boolean	Identificador do tema 'aventura'
tema_animacao	boolean	Identificador do tema 'animação'
tema_criancas	boolean	Identificador do tema 'crianças'
tema_comedia	boolean	Identificador do tema 'comédia'
tema_crime	boolean	Identificador do tema 'crime'
tema_documentario	boolean	Identificador do tema 'documentário'
tema_drama	boolean	Identificador do tema 'drama'
tema_fantasia	boolean	Identificador do tema 'fantasia'
tema_preto_branco	boolean	Identificador do tema 'preto e branco'
tema_horror	boolean	Identificador do tema 'horror'
tema_musical	boolean	Identificador do tema 'musical'
tema_misterio	boolean	Identificador do tema 'mistério'
tema_romance	boolean	Identificador do tema 'romance'
tema_fic_cientifica	boolean	Identificador do tema 'ficção científica'
tema_suspense	boolean	Identificador do tema 'suspense'
tema_guerra	boolean	Identificador do tema 'guerra'
tema_velho_oeste	boolean	Identificador do tema 'velho oeste'

- Usuários

Tabela 3 – Lista de variáveis no conjunto de dados de Usuários

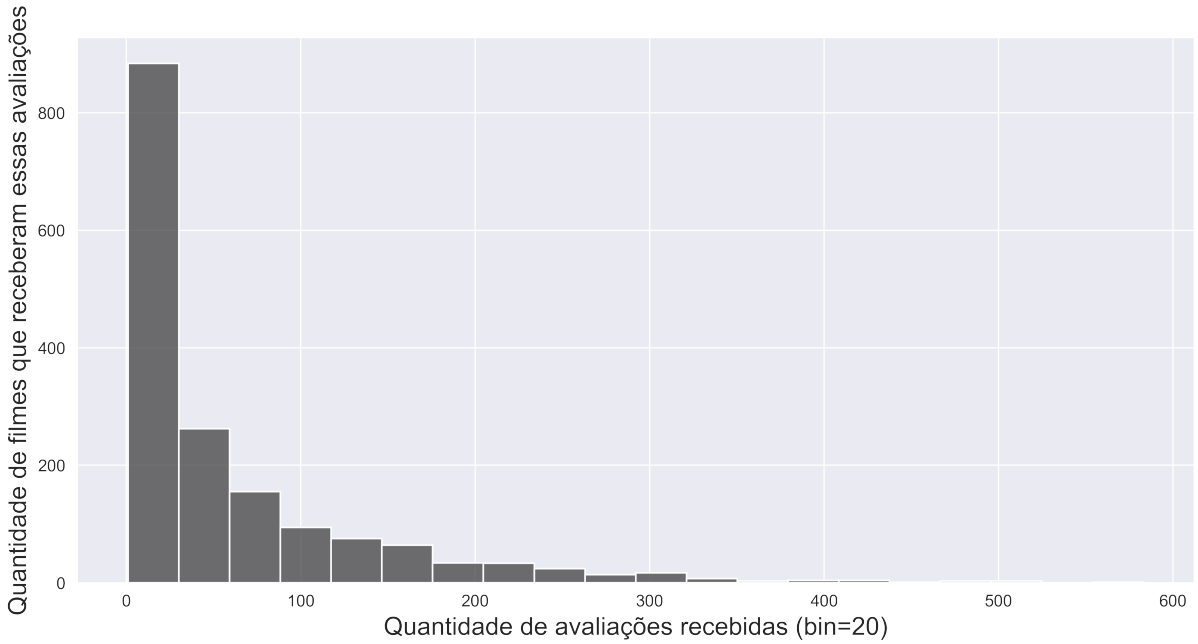
Variável	Tipagem	Descrição
id_usuario	string	Identificador do filme
idade	integer	Idade do usuário
sexo	string	Sexo do usuário
profissao	string	Profissão do usuário
estado_usa	string	Estado de residência do usuário

3.3 Análise Exploratória de Dados

Nesta seção, desenvolveu-se uma breve interpretação para as principais variáveis da base, a fim de que o leitor possa ter uma mínima contextualização do ambiente de dados que se foi trabalhado.

A primeira perspectiva observada foi do conjunto de dados de Interações, mostrando a quantidade de avaliações recebidas pelos filmes. A leitura é simples, no eixo das abscissas tem-se a quantidade de avaliações recebidas pelo filme, separadas por grupos de 20. Já no eixo das ordenadas, tem-se a quantidade de filmes que receberam a quantidade de avaliações referidas no grupo.

Figura 3 – Histograma da distribuição da quantidade de avaliações

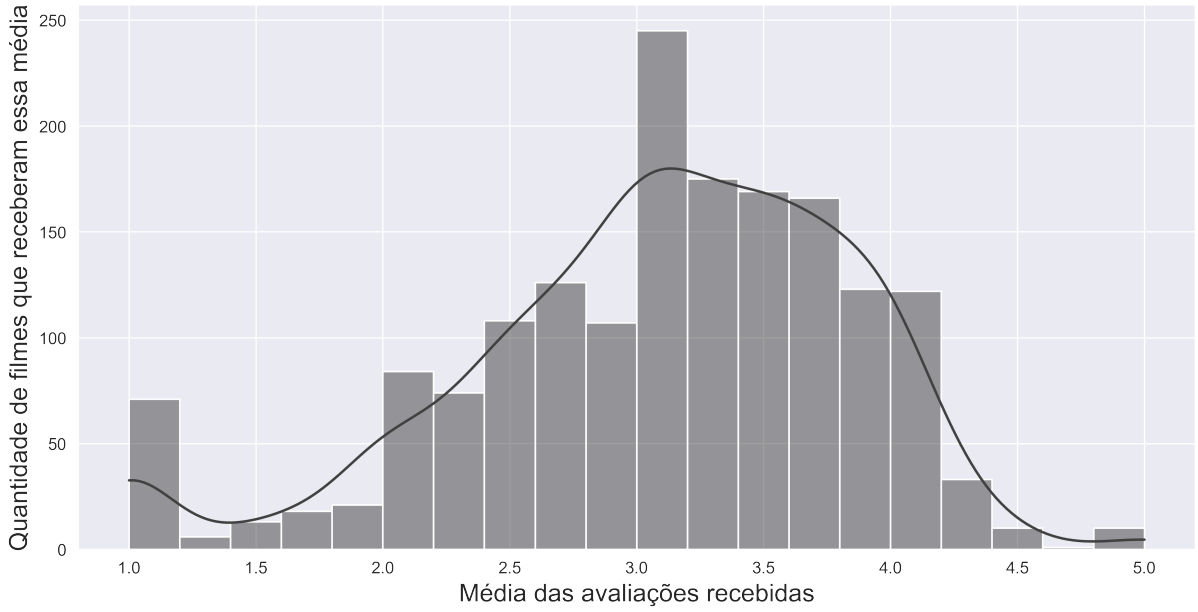


Fonte: o autor (2021)

Ao se interpretar os valores, nota-se um comportamento de cauda mais alongada, demonstrando certa assimetria à direita na distribuição, ou em outras palavras, observa-se que muitos filmes recebem poucas avaliações e poucos filmes recebem muitas avaliações. Em destaque à esquerda, os mais de 800 filmes que receberam entre 0 e 20 avaliações. A quantidade máxima recebida por um filme foi 583 avaliações, mas a mediana da amostra se encontra em 27.

O segundo gráfico gerado faz referência a média das avaliações recebidas pelos filmes. O eixo das abscissas representa à média de avaliações recebidas pelos filme e o eixo das ordenadas representa a quantidade de filmes que receberam essa média de avaliações.

Figura 4 – Histograma da distribuição da média das avaliações por filme



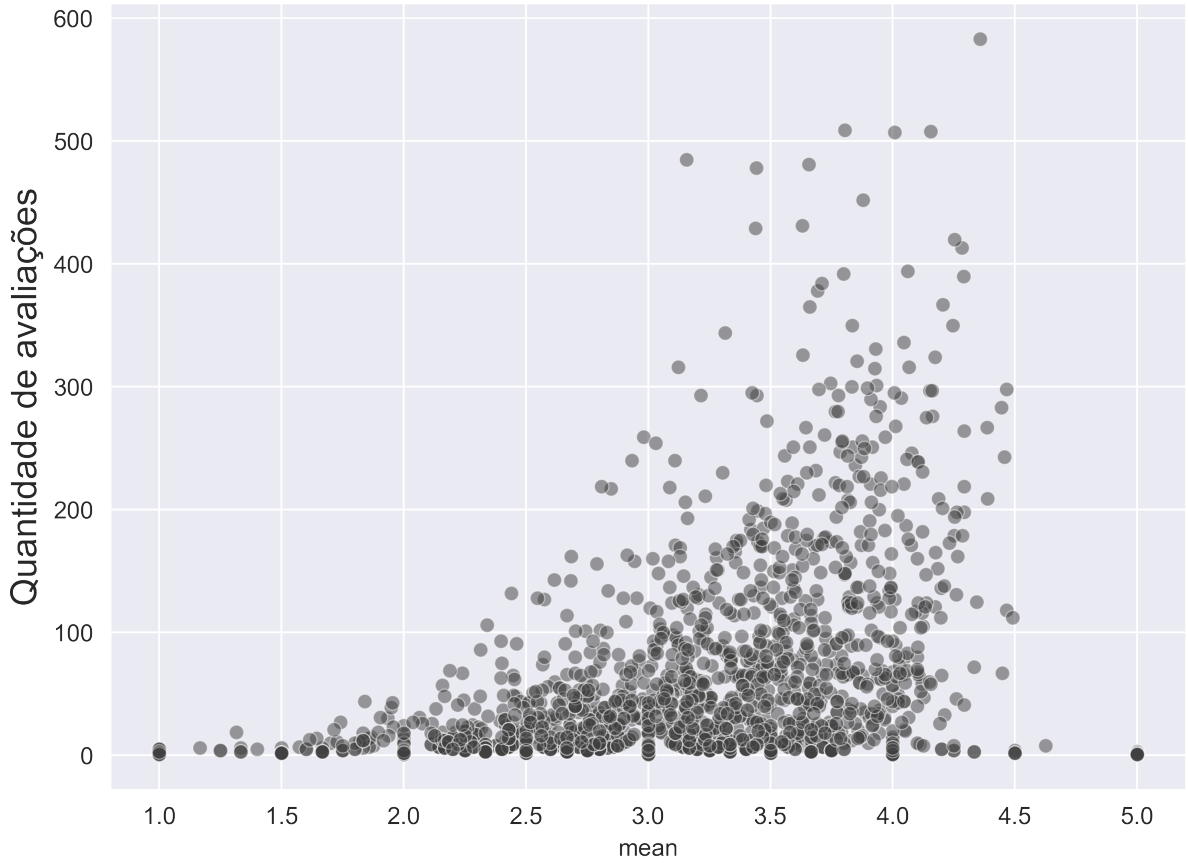
Fonte: o autor (2021)

Nota-se uma distribuição um pouco mais comportada, que se aproxima de uma Gausiana, com uma concentração maior das médias de avaliações dos filmes entre 2,5 e 3,7. Mas a mediana da amostra se encontra por volta dos 3,2. Há uma boa quantidade de filmes que receberam nota média superior a 4, mas menos que tenham recebido nota média inferior a 2.

As duas distribuições acima provocam o pensamento sobre algumas hipóteses com relação ao conjunto de dados. Será que pode-se dizer que os filmes mais avaliados, o que é uma aproximação para talvez os filmes mais vistos, também são os filmes melhor avaliados? De fato, apesar de existir alguma correlação entre essas mensurações, ela não é tão alta. Precisamente de 0,4, e o próximo gráfico ilustra isso.

Mais explicitamente falando, pode-se encontrar filmes com boas notas médias de

Figura 5 – Scatterplot da correlação entre quantidade e a média nas avaliações



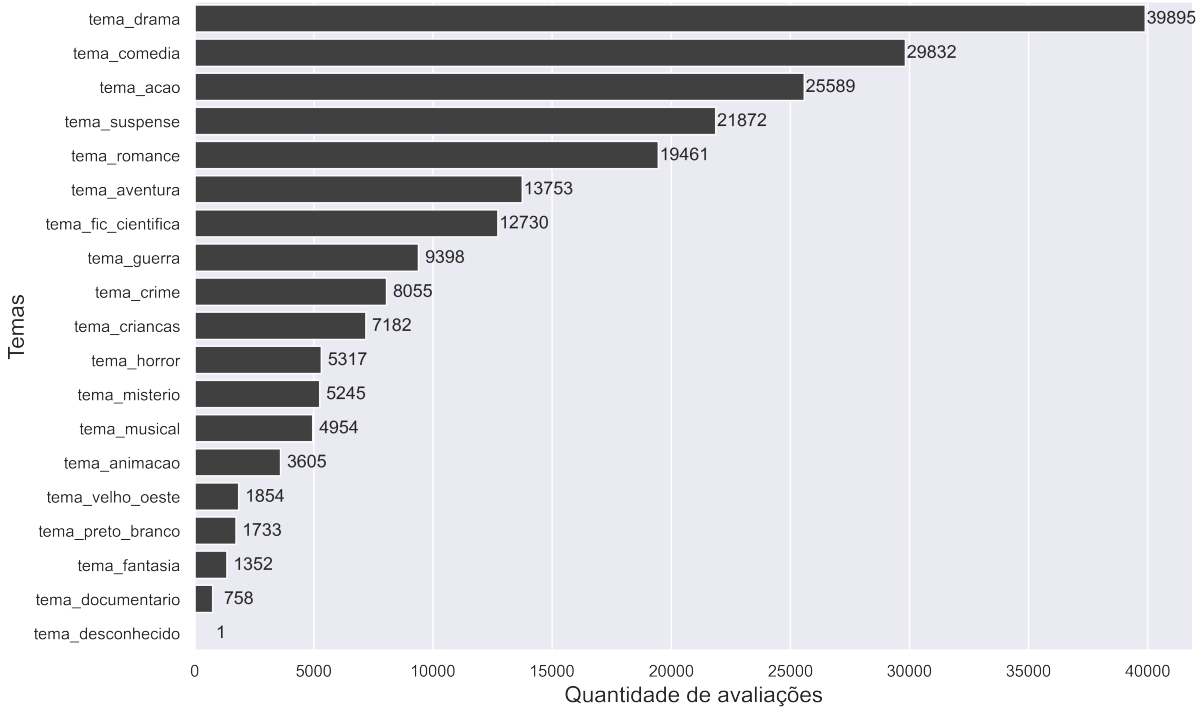
Fonte: o autor (2021)

avaliação, mas que não tenham quantidade relevante de avaliações para sustentar uma recomendação assertiva. Essas implicações são importantes e voltarão a ser utilizadas mais adiante neste trabalho.

Na perspectiva do conjunto de dados de Itens, pode-se observar no próximo gráfico a distribuição das indicações de tema dos filmes. Importante notar que, apesar de pouco frequente, um filme nesse contexto pode estar associado a mais de um assunto.

A interpretação que segue é a de que no eixo das ordenadas, tem-se o conjunto de temas da base de dados e, no eixo das abscissa, tem-se a quantidade de indicações que cada tema recebeu. Pode-se notar que os três temas que mais se destacaram foram, respectivamente, Drama, Comédia e Ação.

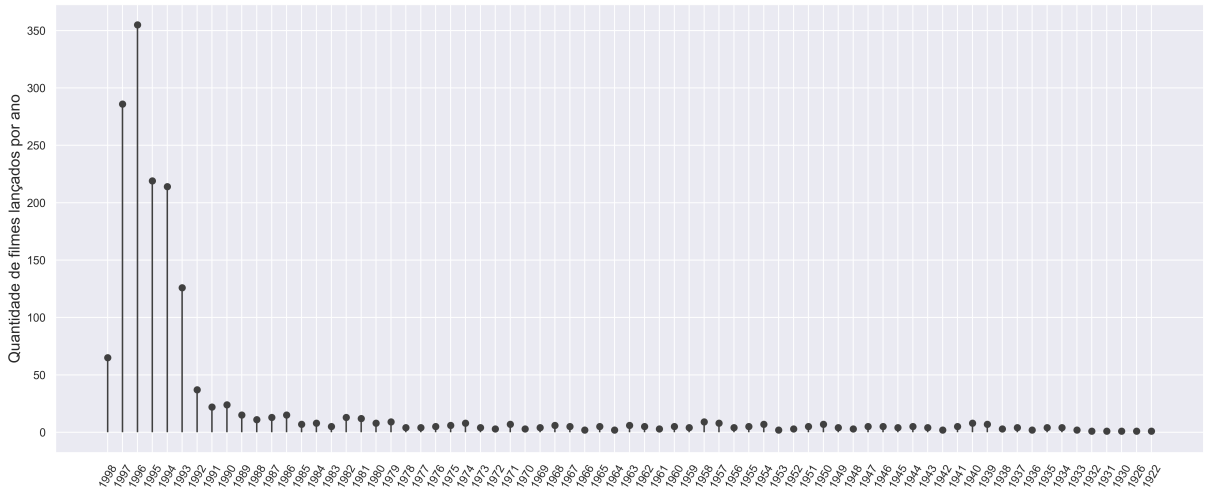
Figura 6 – Temas que mais receberam indicações na base de dados



Fonte: o autor (2021)

Ao se analisar a quantidade de filmes publicados por ano pela Figura 7, pode-se observar que os últimos 6 anos do cojunto de dados retêm as maiores concentrações de lançamentos. Entretanto, o resto da distribuição possui uma quantidade relativamente constante de lançamentos.

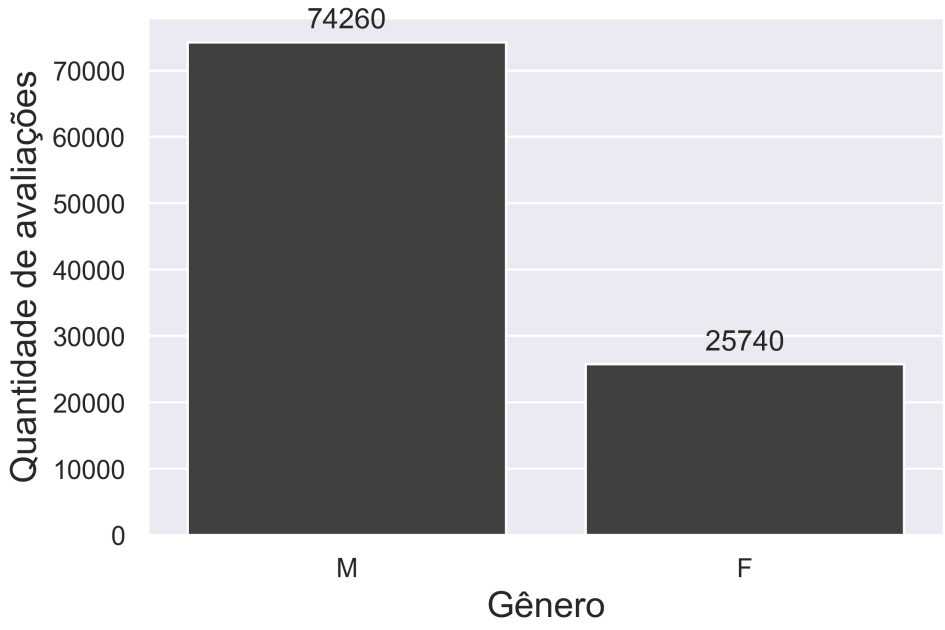
Figura 7 – Lollipop da quantidade de filmes lançados por ano



Fonte: o autor (2021)

Na perspectiva do conjunto de dados de Usuários, pode-se observar a distribuição por gênero das avaliações feitas aos filmes durante o período. No próximo gráfico, o eixo das ordenadas representa a quantidade de avaliações realizadas e, o eixo das abcissas representa o gênero que realizou a avaliação. A maior parte das avaliações foi realizada pelo gênero masculino, cerca de 2.9 vezes mais avaliações do que o gênero feminino.

Figura 8 – Quantidade de avaliações por gênero



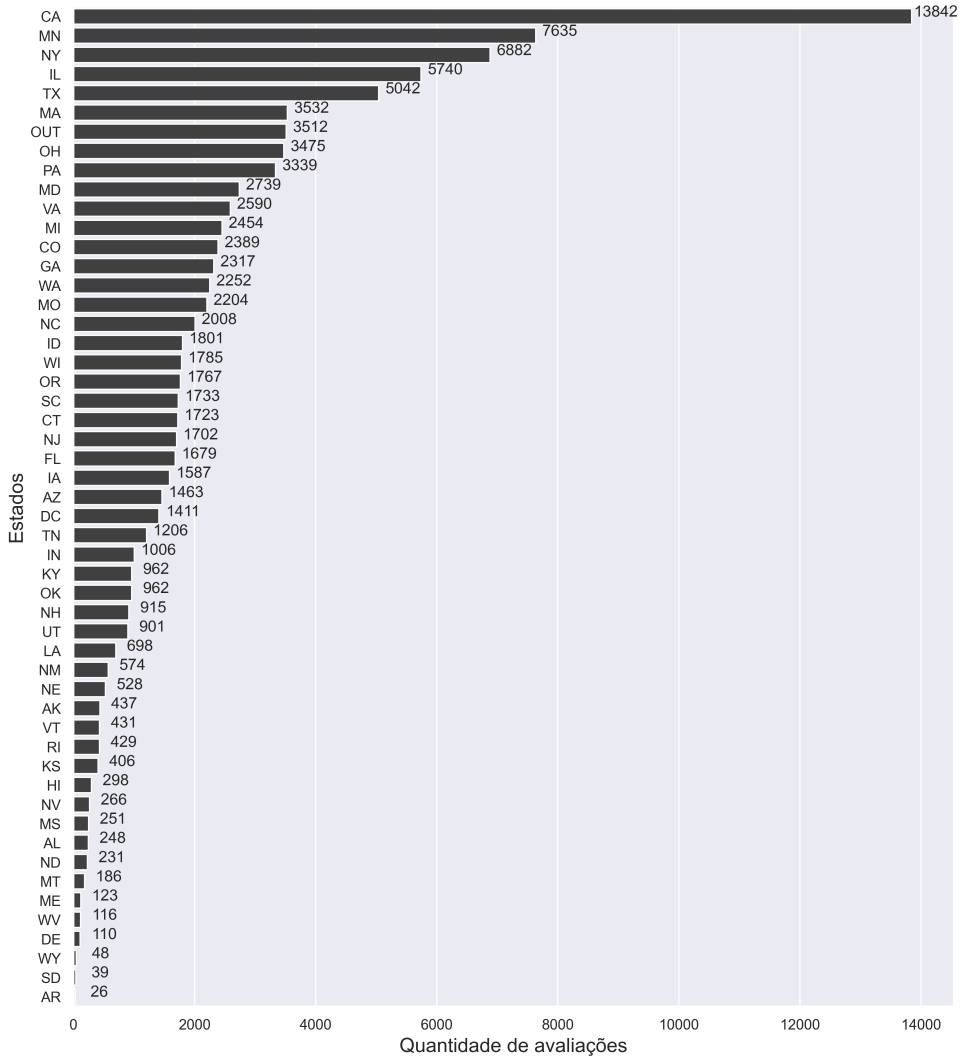
Fonte: o autor (2021)

Entender essas questões demográficas e os padrões encontrados no perfil dos

avaliadores da base é um indicativo futuro da presença de grupos de interesse que podem e serão utilizados para uma filtragem mais personalizada do conteúdo a ser recomendado.

Ao analisar-se a quantidade de avaliações (eixo das ordenadas) feitas por cada estado dos EUA na base (eixo das abscissas), tem-se que Califórnia, Minnesota e Nova Iorque são os estados que têm maior presença na base de dados.

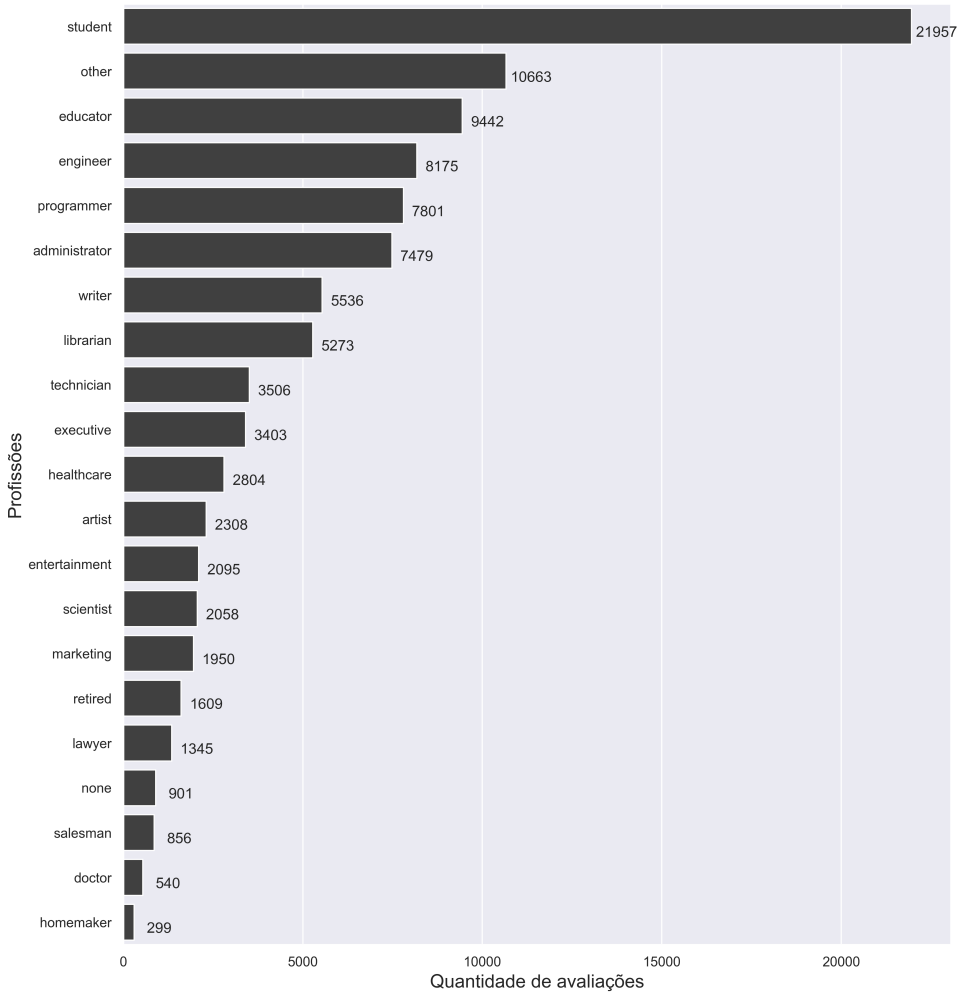
Figura 9 – Estados que mais realizaram avaliações na base de dados



Fonte: o autor (2021)

Ao analisar-se a quantidade de avaliações feitas por cada profissão na base, tem-se que os estudantes são os que têm maior presença na base de dados. Cerca de 2x mais avaliações que a segunda categoria.

Figura 10 – Profissões que mais realizaram avaliações na base de dados



Fonte: o autor (2021)

Agora que algumas das variáveis dos conjuntos de dados foram analisadas, pode-se prosseguir para o desenvolvimento das diferentes abordagens para o Sistema de Recomendação.

3.4 Algorítmos

Para a implementação dos algorítmos responsáveis pelas recomendações, dividiu-se o conjunto de dados em subconjuntos de treino (70%) e teste (30%). A divisão foi realizada de forma a preservar no conjunto de teste uma quantidade mínima de 10 avaliações de filmes por cada usuário. Assim, permitiu-se mais adiante comparar as diversas técnicas utilizadas sobre a ótica das métricas selecionadas na seção 3.5.

Além disso, com a intenção de criar uma referência para as diferentes metodologias, construiu-se também uma abordagem inicial de recomendação pela qual 10 filmes aleatórios foram selecionados para cada usuário do conjunto de teste.

3.4.1 Popularidade

A abordagem baseada em popularidade é a primeira e a mais simples deste trabalho. Não faz uso das informações descritivas dos usuários e nem dos próprios filmes, mas relaciona a quantidade de avaliações com a qualidade das avaliações recebidas por cada filme. Por fim, realiza-se então uma recomendação sem levar em consideração as preferências individuais do usuário.

As etapas para a definição dessa métrica seguem abaixo:

1. Definir uma métrica de popularidade que relacione quantidade e qualidade das avaliações

2. Aplicar esta métrica para todos os filmes e ordenar o conjunto de dados

3. Retirar os filmes já assistidos pelo usuário que receberá a recomendação

4. Recomendar os 10 primeiros filmes

O primeiro ponto é muito importante para o prosseguimento, uma vez que pode-se encontrar filmes muito bem avaliados que ainda não tenham uma quantidade razoável de avaliações recebidas para ponderar essa média. Utilizou-se então uma métrica desenvolvida pelo IMDb (Internet Movie Database, 2021) que visa balancear estes pesos.

Seja MP a Média Ponderada do IMDb, tem-se:

$$MP = (R * v + C * m)/(v + m)$$

onde,

- R –> média das avaliações recebidas por um filme

- v –> quantidade de avaliações recebidas por um filme

- C –> média de todas as avaliações da base de dados

- m –> é o mínimo de avaliações que um filme precisa para entrar no ranking

Para m, definiu-se o 90º percentil como linha de corte. Em outras palavras, por causa do interesse nos filmes mais populares do conjunto de dados, considerou-se então que somente os filmes que possuíssem mais do que 113 avaliações poderiam ser retornados pela função de recomendação.

3.4.2 Recomendação Demográfica: Agrupamento Hierárquico com Popularidade

Com a intenção de tornar a recomendação mais personalizada, considerando aspectos indivuduais dos usuários, desenvolveu-se também a abordagem por Clusterização demográfica com Popularidade. Essa abordagem considera filtros baseados em conteúdo, especificamente os metadados do conjunto de dados de usuários, para criar grupos de mesmo interesse e apartir de cada grupo, priorizar os filmes mais populares entre membros pertencentes ao grupo para a recomendação. Em resumo, as etapas abaixo descrevem a abordagem:

1. Agrupar o conjunto de dados de usuários utilizando os atributos disponíveis (idade, sexo, profissão, estado)

2. Para cada um dos grupos encontrados, selecionar todos os filmes assistidos pelo menos uma vez por um usuário pertencente a esse grupo

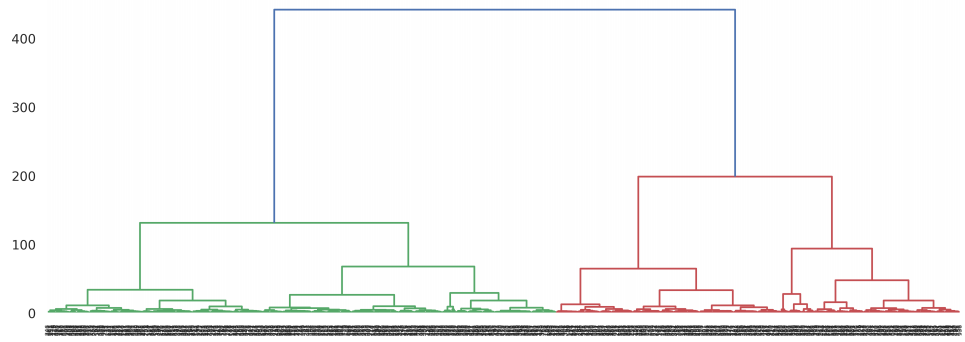
3. Aplicar a métrica de popularidade desenvolvida no item 3.4.1 para ordenar os filmes mais populares da seleção de filmes do ponto anterior

4. Retirar os filmes já assistidos pelo usuário que receberá a recomendação

5. Recomendar os 10 primeiros filmes

Para o primeiro ponto, utilizou-se a técnica de aprendizagem de máquina não supervisionada de Agrupamento Hierárquico Aglomerativo com métrica distância euclidiana e método de Ward. Escolheu-se a quantidade de 10 grupos como ponto de corte da árvore desenvolvida, essa escolha considerou como referência o valor de Silhouette Score de 0.5184.

Figura 11 – Visualização do agrupamento dos usuários do conjunto de dados



Fonte: o autor (2021)

3.4.3 Filtragem Colaborativa Baseada em Modelo: Single Value Decomposition

A abordagem anterior é capaz de oferecer recomendações mais personalizas pois leva em consideração algumas características descritivas dos próprios usuários, os classificando em grupos. Entretanto, ela ainda não faz uso dos vieses e preferências individuais de cada usuário. Desenvolveu-se então a abordagem baseada em Filtragem Colaborativa utilizando o método de Single Value Decomposition.

Sistemas que utilizam a abordagem de Filtragem Colaborativa assumem que um usuário deve gostar de itens bem avaliados por outros usuários com gostos parecidos. A tabela abaixo ajuda a ilustrar o problema:

Tabela 4 – Matriz de avaliações usuário-filme

	f1	f2	f3
u1	4	1	
u2		2	4
u3	4	1	2

Na matriz acima, os termos u (u1, u2 e u3) representam os usuários, os termos f (f1, f2 e f3) representam os filmes e os valores centrais representam uma avaliação de um dado usuário para um dado filme. A ideia geral dessa abordagem é encontrar as correlações existentes entre os eixos dessa matriz.

Pode-se notar, por exemplo, que o usuário u3 tem uma correlação forte com o usuário u1, pois possuem preferências iguais quando observadas para os filmes f1 e f2. De forma simplista, se fosse necessário predizer a avaliação do usuário u1 para o filme f3, poderia ser dito, com base na correlação notada anteriormente, que estaria perto de 2.

A utilização do algorítmo de Single Value Decomposition é uma maneira de lidar com o problema de escalabilidade e esparsidade criada pela Filtragem Colaborativa e utiliza um método fatoração da matriz para otimizar a busca da similaridade entre usuários e itens. Uma vez encontrada essas similaridades, é possível predizer as avaliações que os usuários dariam para filmes ainda não vistos por eles, e utilizando esse conhecimento, uma recomendação poderia ser realizada.

A abordagem geral é descrita nas etapas abaixo:

1. Treinar o modelo de Single Value Decomposition com as avaliações do conjunto de treino

2. Predizer as avaliações que seriam dadas para um dado usuário a cada filme do conjunto de dados

3. Para esse usuário e com os valores previstos anteriormente, ordenar os filmes que ele melhor avaliaria do conjunto de dados

4. Retirar os filmes já assistidos por esse usuário

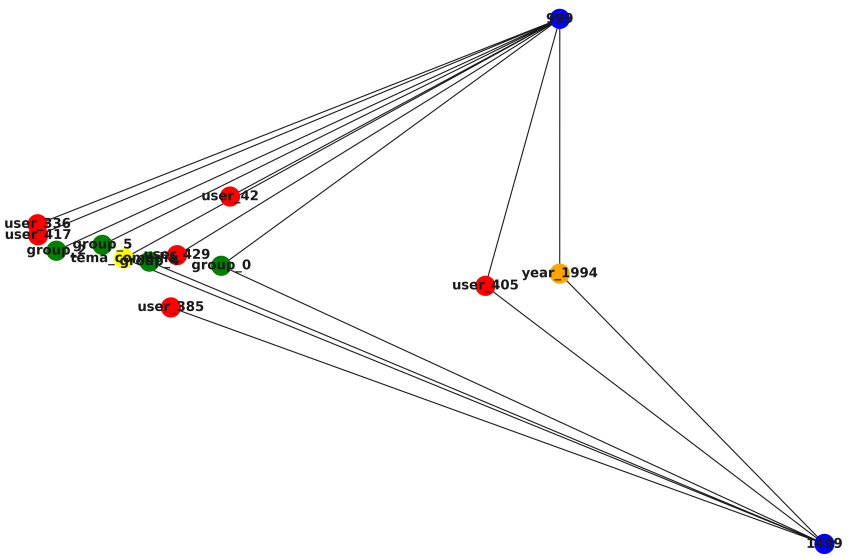
5. Recomendar os 10 primeiros filmes

A partir da utilização dos dados de treino e com validação cruzada, num tempo de treinamento de aproximadamente 6 segundos conseguiu-se predições das avaliações com Erro Absoluto Médio de 0.7520.

3.4.4 Filtragem Colaborativa Baseada em Vizinhança: Adamic Adar Link Prediction

Uma outra possível maneira de se trabalhar com Filtragem Colaborativa para Sistemas de Recomendação é com a utilização das técnicas de Redes Complexas. A ideia por trás da abordagem é utilizar grafos para representar as conexões entre as avaliações dos usuários aos filmes, bem como algumas das características descritivas dos dois conjuntos de dados.

Figura 12 – Grafo com as conexões existentes entre usuários e filmes



Fonte: o autor (2021)

Na ilustração acima, nota-se as conexões entre os filmes ’999’ (Clean Slate) e ’1499’ (Grosse Fatigue). Os nós em azul representam os próprios filmes, enquanto que o nó em laranja representa a primeira característica em comum entre esses filmes: ambos foram lançados em 1944.

Além do ano de lançamento esses filmes também compartilham de outras conexões, por exemplo, em amarelo no grafo acima tem-se o nó que representa o tema ’comédia’, indicando que ambos os filmes são classificados com essa categoria.

As conexões mencionadas acima representavam características descritivas dos próprios filmes, mas também existem nós que representam a característica descritiva dos usuários. Em verde, por exemplo, tem-se os diversos nós que representam os grupos (os mesmos definidos em 3.4.2), dentro dos quais pelo menos um dos membros assitiu ao filme conectado. O grupo '0' assistiu tanto ao filme '999' como ao filme '1499', entretanto o grupo '5' assistiu apenas ao filme '999'.

Além das conexões por atributos descritivos dos objetos, as interações existentes entre eles também foram contabilizadas. Os nós em vermelho, por exemplo, representam uma conexão indireta entre os filmes efetivada pelos usuários que os assistiram. O usuário '405', por exemplo, assistiu a ambos os filmes, enquanto que o usuário '385' assistiu apenas ao filme '1499'.

Teoricamente, quanto mais conexões compartilhadas possuem os filmes, mais próximos uns dos outros eles são. A métrica denominada Adamic Adar quantifica essa proximidade existente entre os nós e foi utilizada como ferramenta de recomendação neste material. Os passos abaixo resumem a abordagem:

1. Construir o grafo a partir dos atributos descritivos e das interações entre os conjuntos de dados de usuários e filmes

2. Aplicar a métrica Adamic Adar ao grafo e computar as proximidades entre os filmes

3. Para um dado usuário, selecionar os últimos k filmes assitidos por ele

4. Para cada um dos k filmes assistidos por esse usuário, utilizar a métrica do Adamic Adar definida acima e selecionar os 10 filmes mais próximos de cada um

5. Ordenar essa lista final que contém todos os filmes pela métrica Adamic Adar

6. Retirar os filmes repetidos e/ou já assistidos pelo usuário

7. Recomendar os 10 primeiros filmes

A quantidade ideal para k selecionada empiricamente neste trabalho foi k = 20. Utilizar tanto os atributos descritivos dos objetos, como as interações indiretas existentes entre eles caracterizou essa abordagem como híbrida e a escolha ideal de k preservou também a característica dinâmica da evolução das preferências de cada usuário no tempo.

3.4.5 Sistemas de Recomendação Híbridos: Adamic Adar Link Prediction + SVD

Utilizou-se nesta última abordagem uma combinação entre a métrica Adamic Adar, que computa a proximidade para nós (filmes) em Redes Complexas, com a previsão dada pela abordagem Single Value Decomposition (Seção 3.4.3), que estimava o valor da avaliação dada por um usuário a um filme ainda não assistido por ele.

A ideia era verificar se, utilizando critérios diferentes para ordenar os filmes próximos

àqueles que usuário assistiu, poderia-se melhorar as recomendações feitas e, consequentemente, afetar de alguma maneira as métricas que serão definidas mais adiante na Seção 3.5.

A abordagem geral é descrita nas etapas abaixo:

1. Construir o grafo a partir dos atributos descritivos e das interações entre os conjuntos de dados de usuários e filmes

2. Aplicar a métrica Adamic Adar ao grafo e computar as proximidades entre os filmes

3. Para um dado usuário, selecionar os últimos k (k=20, conforme definido anteriormente) filmes assitidos por ele

4. Para cada um dos k filmes assistidos por esse usuário, utilizar a métrica Adamic Adar e selecionar os 10 filmes mais próximos de cada um

5. Ordenar essa lista final que contém todos os filmes pela avaliação predita pela abordagem de Single Value Decomposition (Seção 3.4.3)

6. Retirar os filmes repetidos e/ou já assistidos pelo usuário

7. Recomendar os 10 primeiros filmes

3.5 Métricas de Avaliação de Resultado

Uma vez definidas as abordagens utilizadas durante o desenvolvimento deste trabalho, fez-se necessário comparar esses algorítmos e analisar suas vantagens e desvantagens. Para isso, escolheu-se 5 métricas de avalição de resultados comuns a algorítmos de recomendação e elas serão apresentadas nas subseções que seguem.

3.5.1 Precision

A primeira métrica que será discutida neste trabalho é a Precision. Muito comum para algorítmos de Classificação em Machine Learning, esta métrica avalia a quantidade de acertos dada a quantidade de predições realizadas pelo modelo (Sonya Sawtelle, 2016).

De maneira análoga a mencionada acima, tem-se a definição formal abaixo:

$$Precision = \frac{Verdadeiros\,Positivos}{Verdadeiros\,Positivos + Falsos\,Positivos}$$

A fórmula acima apresenta o cálculo da quantidade de Verdadeiros Positivos (acertos preditos pelo modelo) sob todos os valores preditos pelo modelo, ou seja, Verdadeiros Positivos e Falsos Positivos (acertos preditos pelo modelo que não foram na prática verificados no conjunto de teste).

A tabela abaixo resume a aplicação desta métrica de avaliação para as abordagens apresentadas na Seção 3.4:

Tabela 5 – Métricas de Avaliação de Resultado: Precision

Abordagem	Resultado
Aleatória	06.30%
Popularidade	23.28%
Agrupamento Hierárquico com Popularidade	22.71%
Single Value Decomposition	14.24%
Adamic Adar Link Prediction	34.83%
Adamic Adar Link Prediction + SVD	**35.83%**

Pode-se notar com os valores acima que a abordagem que utiliza sistemas híbridos de recomendação e a abordagem que utiliza a métrica Adamic Adar sozinha apresentaram os melhores resultados para a Precision. Ao passar do simples para o sistema híbrido teve-se um ganho de 1 ponto percentual na métrica, saindo de 34.83% para 35.83%

Apesar dessa ligeira melhora quando combinada, a abordagem Single Value Decomposition sozinha não apresentou bons resultados para a métrica Precision, com aproximadamente 14% de acertos ela só não foi pior que a abordagem Aleatória, que foi criada justamente para a realização dessas comparações de referência e obteve valores aproximados de 6% de acertos.

Ainda, as abordagens por Popularidade também apresentaram valores curiosos para a Precision. A tentativa de utilizar o agrupamento dos usuários para criar os grupos de mesmo interesse não aumentou a métrica quando comparada com a Popularidade sem a utilização do agrupamento. Em números, ficaram aproximadamente em torno de 23%.

3.5.2 MAP@K: Mean Average Precision

A métrica Precision, como mencionada anteriormente, é boa para problemas de classificação. Entretanto, considerando que um usuário nesse contexto de recomendação de filmes tem uma quantidade finita de tempo e atenção, pode-se querer saber, então, não apenas os dez produtos que ele pode gostar, mas também quais são os mais mais prováveis dele assistir, transformando o problema de classificação num problema de ordenação.

Para esse tipo de tarefa, quer-se uma métrica que se recompense por obter muitas recomendações corretas ou relevantes e se recompense por tê-las no início da lista (classificação mais alta) (Sonya Sawtelle, 2016). Uma métrica que reflete essa solução é a Mean Average Precision @ K.

A primeira parte do cálculo envolve a fórmula abaixo e, para isso, considere AP@K como equivalente a Average Precision @ K:

$$AP@K = \frac{1}{m} \sum_{k=1}^{N} P(k).rel(k)$$

onde rel (k) é apenas uma função indicadora que diz se o k-ésimo filme foi (para rel (k) = 1) assistido ou (para rel (k) = 0) não assistido pelo usuário. P(k) na fórmula acima diz respeito a métrica Precision calculada até o recorte dos k-ésimos primeiros filmes.

A Tabela 6 ilustra recomendações de filmes realizadas para 5 usuários diferentes e o resultado do uso para a fórmula para cada um deles. Considere o valor 1 caso o filme da k-ésima posição do conjunto recomendado seja de fato assistido pelo usuário e 0 caso contrário.

Tabela 6 – Aplicação da métrica AP@K para 3 filmes

Conjunto recomendado	AP@3
$[0, 0, 1]$	$(1/3)[0 + 0 + (1/3)] = 0.11$
$[0, 1, 0]$	$(1/3)[0 + (1/2) + 0] = 0.15$
$[1, 0, 0]$	$(1/3)[(1 + 0 + 0)] = 0.33$
$[0, 1, 1]$	$(1/3)[0 + (1/2) + (2/3)] = 0.38$
$[1, 1, 1]$	$(1/3)[(1) + (2/2) + (3/3)] = 1$
Média	0.39

Os três primeiros usuários tiveram a mesma quantidade de filmes recomendados que de fato foram assistidos por eles, mas a ordem em cada um deles é diferente e com o resultado da aplicação da métrica foi possível verificar que um acerto em uma posição mais adiantada torna o valor da métrica maior e o mesmo acontece se há uma maior quantidade de acertos no conjunto de dados recomendado.

A média apresentada na última linha da tabela completa sua definição e é, resumidamente, o cálculo da média dos resultados da aplicação da AP@K para cada um dos usuários do conjunto de dados.

Assim, voltando para o caso de uso, a tabela abaixo ilustra o resultado da métrica para cada uma das abordagens definidas na Seção 3.4:

Tabela 7 – Métricas de Avaliação de Resultado: Média Ponderada da Precisão

Abordagem	Resultado
Aleatória	02.49%
Popularidade	14.43%
Agrupamento Hierárquico com Popularidade	14.33%
Single Value Decomposition	07.05%
Adamic Adar Link Prediction	**23.63**%
Adamic Adar Link Prediction + SVD	23.46%

As comparações possíveis de serem realizadas a partir da utilização desta métrica são bem parecidas com as já comentadas para a métrica Precision, porém, com a utilização da métrica MAP@K tem-se uma avaliação mais adequada ao contexto de algorítmos de recomendação e, neste caso, nota-se uma redução dos valores para todas as abordagens. A melhor abordagem com MAP@K foi Adamic Adar Link Prediction com 23.63% em comparação com a Adamic Adar Link Prediction + SVD, melhor abordagem com a métrica Precision, que obteve valores aproximados de 35.83% dos acertos.

3.5.3 Catalog Coverage

Enquanto as duas primeiras estavam preocupadas com a relevância de uma recomendação ao usuário, as próximas métricas trazem uma visão de quão surpreendentes (positivamente) podem ser as recomendações aos usuários. Uma métrica bastante interessante para algorítmos de recomendação é a Catalog Coverage, isto é, o quão abrangente ou o quanto de alcance tem uma certa função de recomendação dado um conjunto de possíveis filmes a serem recomendados (GE; DELGADO; JANNACH, 2010).

Segue a definiçao formal,

$$Catalog\,Coverage = \frac{|\cup_{j=1...N} I_L^j|}{|I|}$$

onde I_L^j representa o conjunto de itens contidos na lista L retornados na j-ésima recomendação. N pode ser compreendido como a quantidade de recomendações realizadas e I é o conjunto de todos os itens existentes nos dados.

Uma tradução literal à formalidade acima e ainda adaptada ao contexto de recomendação seria a contagem de filmes únicos recomendados após todas as interações acontecerem divido pelo tamanho do catálogo de filmes disponíveis. É uma métrica que diz quão boa é a função no sentido de aproveitar todas as oportunidades existentes no conjunto de dados.

A tabela abaixo resume a aplicação desta métrica para todas as abordagens estudadas:

Tabela 8 – Métricas de Avaliação de Resultado: Cobertura

Abordagem	Resultado
Aleatória	**49.11%**
Popularidade	02.33%
Agrupamento Hierárquico com Popularidade	07.59%
Single Value Decomposition	13.96%
Adamic Adar Link Prediction	07.23%
Adamic Adar Link Prediction + SVD	12.68%

A abordagem Aleatória, utilizada apenas como referência para as métricas, foi a que apresentou os melhores resultados. Justamente por conta da sua natureza aleatória, quase 50% dos filmes do catálogo foram aproveitados.

Para a abordagem de Popularidade, tem-se que a utilização das informações demográficas dos usuários provocou uma melhora considerável na métrica que chegou a aproximadamente 7% de cobertura, o mesmo valor encontrado na abordagem Adamic Adar Link Prediction.

Os melhores resultados nesta métrica ficaram para a utilização da abordagem Single Value Decomposition, ela não só foi capaz de melhorar os valores de Adamic Adar Link Prediction para aproximadamente 13%, como também apresentou sozinha uma cobertura de aproximadamente 14% dos filmes do catálogo.

3.5.4 Personalization

A última métrica estudada neste trabalho foi a Personalization ou personalização. Ela destaca quão diferente são as listas de filmes recomendados aos usuários do conjunto de dados (GE; DELGADO; JANNACH, 2010). Os exemplos abaixo ajudam a entender melhor a definição dessa métrica.

Tabela 9 – Recomendações de filmes a 3 usuários

1	$[A, B, C, D]$
2	$[A, B, C, X]$
3	$[A, B, C, Z]$

Na primeira etapa há uma transformação do conjunto de filmes recomendados para cada um dos usuários para uma matriz binária, construída através de uma função indicadora (1 se o filme foi recomendado para o usuário, 0 caso contrário).

Tabela 10 – Transformação do conjunto de filmes recomendados

	A	C	B	D	X	Z
1	1	1	1	1	0	0
2	1	1	1	0	1	0
3	1	1	1	0	0	1

A partir daí, aplica-se a similaridade cosseno entre as linhas da matriz gerada, isto é, mede-se a similaridade, dois a dois, desses vetores em um espaço vetorial. Formalmente, é medida pelo cosseno do ângulo entre esses vetores e, em outras palavras, determina se esses estão apontando aproximadamente para a mesma direção. Sejam x e y vetores (linhas) da matriz, tem-se:

$$Similaridade(x, y) = cos(\theta) = \frac{x.y}{|x||y|}$$

A matriz de similaridade cosseno é então gerada,

Tabela 11 – Matriz de similaridade cosseno

$$\begin{bmatrix} 1 & 0.75 & 0.75 \\ 0.75 & 1 & 0.75 \\ 0.75 & 0.75 & 1 \end{bmatrix}$$

A definição da métrica de Personalization é dada pelo inverso da similaridade, ou em outras palavras, do valor gerado com a utilização da média dos valores do triângulo superior (acima da diagonal principal) da matriz de similaridade cosseno.

$$Personalization = 1 - similaridade = 1 - 0.75 = 0.25$$

De volta ao contexto de recomendação, a tabela abaixo resume a aplicação desta métrica para cada uma das abordagens estudadas:

Tabela 12 – Métricas de Avaliação de Resultado: Personalização

Abordagem	Resultado
Aleatória	**97.09**%
Popularidade	41.56%
Agrupamento Hierárquico com Popularidade	72.47%
Single Value Decomposition	73.33%
Adamic Adar Link Prediction	65.49%
Adamic Adar Link Prediction + SVD	72.91%

Nota-se, assim como para a Coverage, que a abordagem Aleatória apresentou os melhores resultados. Isso se deve, é claro, ao fato da abordagem gerar listas de recomendação totalmente aleatórias e que, para a métrica em questão, gerou valores perto do máximo.

A abordagem de Popularidade sozinha não gerou bons resultados, foi a pior dentre as abordagens desenvolvidas. Enquanto que Adamic Adar Link Prediction sozinha foi melhor, mas também não passou dos 65.49%.

A segunda melhor abordagem foi o Agrupamento Hierárquico com Popularidade que apresentou valores de aproximadamente 72.47% de personalização. O destaque positivo nessa métrica ficou para Single Value Decomposition que conseguiu aumentar pouco mais de 7 pontos percentuais na métrica Adamic Adar Link Prediction e, quando utilizada sozinha, foi a melhor abordagem testada, com 73.33%.

4 CONCLUSÃO

Neste capítulo são apresentadas as conclusões relativas ao desenvolvimento do estudo sobre sistemas de recomendação.

4.1 Considerações Finais

Este trabalho contribuiu de diversas maneiras para o aprendizado e fixação do conhecimento passado pelo curso de MBA em Ciência de Dados para o presente autor ao explorar diversas técnicas e abordagens de algorítmos de recomendação utilizando como insumo um conjunto de dados público em um contexto de usuários e filmes assistidos por eles.

A princípio, o conjunto de dados obtido foi interpretado e manipulado em linguagem Python para se adequar às condições de aplicação posterior dos algorítmos, recebeu um tratamento de dados simples com ajuste de tipagem e transformação de um campo de CEP em um campo de ESTADO. Posteriormente, aplicou-se uma análise exploratória de dados para contextualição e entendimento dos principais atributos descritivos.

Para o desenvolvimento das técnicas de recomendação, cinco abordagens foram estudadas e quatro métricas de avaliação de desempenho foram aplicadas.

A primeira abordagem foi baseada na criação de um índice de Popularidade para cada um dos filmes assistidos pelos usuários. Simples e sem levar em conta aspectos individuais, a técnica apresentou valores pouco destacáveis em cada uma das métricas de avaliação, mas melhorou resultados de outras abordagens quando utilizada em conjunto com elas.

A segunda abordagem baseou-se na criação de grupos de usuários parecidos considerando aspectos de seus atributos demográficos. Sexo, idade, estado de origem e profissão foram utilizados pela técnica de Agrupamento Hierárquico para a identificação de grupos. A partir desses grupos, recomendações ordenadas pelo índice de Popularidade foram realizadas.

Essa abordagem se mostrou interessante, apresentou excelentes valores para a personalização de recomendações e bons valores para as métricas de precisão e de cobertura. Destaque também pela relativa simplicidade de cálculo e a possibilidade de gerar boas recomendações mesmo que um usuário não tenha um histórico de interações.

A terceira abordagem foi baseada em Single Value Decomposition, uma técnica de fatoração de uma matriz que contém o histórico de avaliação dos usuários aos filmes. Essa técnica requer um pouco mais de complexidade de cálculo e um histórico considerável de

interações pelo usuário para melhorar a assertividade. Apesar de ter apresentado valores baixos para as métricas de precisão, obteve os melhores desempenhos de personalização e cobertura.

A quarta e a quinta abordagem foram baseadas na construção de grafos com as interações entre usuários e filmes, além da utilização da teoria de Redes Complexas, especificamente da métrica de proximidade entre nós Adamic Adar. Apesar de complexas e também necessitarem de histórico de interações pelo usuário, elas apresentaram os melhores índices de precisão e bons índices de personalização e cobertura.

A combinação das abordagens de Single Value Decomposition para ordenar de forma diferente a lista de filmes recomendada para os usuários melhorou, de modo geral, as métricas de avaliação.

O equilíbrio entre as métricas de avaliação apresentadas levou o presente autor a entender que a abordagem híbrida de Adamic Adar Link Prediction com ordenação dada por Single Value Decomposition obteve o melhor desempenho geral entre todas as abordagens estudadas.

Pela simplicidade de cálculo e pelos resultados razoáveis obtidos em cada métrica de avaliação, entende-se também que a abordagem baseada em Agrupamento Hierárquico com Popularidade seria uma boa alternativa aos casos em que não há o histórico de interações do usuário.

Finalmente, entende-se que seria interessante explorar em trabalhos futuros os valores ideais dos parâmetros definidos em cada abordagem deste trabalho e também os impactos nas métricas de avaliação com uma utilização conjunta dessas duas abordagens.

REFERÊNCIAS

ADAMIC, L. A.; ADAR, E. Friends and neighbors on the web. **Soc. Networks**, v. 25, p. 211–230, 2003.

AGGARWAL, C. **Recommender Systems: The Textbook**. Springer International Publishing, 2016. ISBN 9783319296593. Disponível em: <https://books.google.com.br/books?id=GKjWCwAAQBAJ>.

BALABANOVIC, M.; SHOHAM, Y. Fab: Content-based, collaborative recommendation. **Communications of the ACM**, v. 40, p. 66–72, 03 1997.

BENNETT, J.; LANNING, S. The netflix prize. In: . [S.l.: s.n.], 2007.

BOBADILLA, J. et al. Recommender systems survey. **Knowledge-Based Systems**, v. 46, p. 109–132, 2013. ISSN 0950-7051. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0950705113001044>.

GE, M.; DELGADO, C.; JANNACH, D. Beyond accuracy: Evaluating recommender systems by coverage and serendipity. In: . [S.l.: s.n.], 2010. p. 257–260.

Internet Movie Database. **IMDb**. 2021. Disponível em: <https://help.imdb.com/article/imdb/track-movies-tv/ratings-faq/G67Y87TFYYP6TWAV#ratings>. Acesso em: 20 novembro 2021.

KOREN, Y.; BELL, R.; VOLINSKY, C. Matrix factorization techniques for recommender systems. **Computer**, v. 42, n. 8, p. 30–37, 2009.

LORENA, A. et al. **Inteligência Artificial: uma abordagem de Aprendizado de Máquina (2a edição)**. [S.l.: s.n.], 2021. ISBN 9788521637493.

Lü, L. et al. Recommender systems. **Physics Reports**, v. 519, n. 1, p. 1–49, 2012. ISSN 0370-1573. Recommender Systems. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0370157312000828>.

RESNICK, P.; VARIAN, H. R. Recommender systems. **Commun. ACM**, Association for Computing Machinery, New York, NY, USA, v. 40, n. 3, p. 56–58, mar. 1997. ISSN 0001-0782. Disponível em: <https://doi.org/10.1145/245108.245121>.

RICCI, F.; ROKACH, L.; SHAPIRA, B. Introduction to recommender systems handbook. In: **Recommender systems handbook**. [S.l.]: Springer, 2011. p. 1–35.

Sonya Sawtelle. **Mean Average Precision (MAP) For Recommender Systems**. 2016. Disponível em: <http://sdsawtelle.github.io/blog/output/mean-average-precision-MAP-for-recommender-systems.html>. Acesso em: 20 julho 2021.

TAN, P.; STEINBACH, M.; KUMAR, V. **Introduction to Data Mining**. Pearson Addison Wesley, 2006. (Always learning). ISBN 9780321321367. Disponível em: <https://books.google.com.br/books?id=YHsWngEACAAJ>.

TAN, P.-N. **Data mining introduction**. 2019.