UNIVERSIDADE FEDERAL RURAL DE PERNAMBUCO

RODRIGO NATIVO DO BRASIL BROCHARDT

UTILIZAÇÃO DE FILTRAGEM COLABORATIVA NO AUXÍLIO DE RECOMENDAÇÃO PERSONALIZADA PARA LEITORES DE MANGÁ

RECIFE
2024

Rodrigo Nativo do Brasil Brochardt

Utilização de Filtragem Colaborativa no auxílio de recomendação personalizada para leitores de mangá

Monografia apresentada ao Curso de Bacharelado em Sistemas de Informação da Universidade Federal Rural de Pernambuco, como requisito parcial para obtenção do título de Bacharel em Sistemas de Informação.

Universidade Federal Rural de Pernambuco – UFRPE
Departamento de Estatística e Informática
Curso de Bacharelado em Sistemas de Informação

Orientador: Cícero Garrozi

Recife
2024

Dados Internacionais de Catalogação na Publicação
Universidade Federal Rural de Pernambuco
Sistema Integrado de Bibliotecas
Gerada automaticamente, mediante os dados fornecidos pelo(a) autor(a)

B864u Brochardt, Rodrigo Nativo do Brasil
Utilização de filtragem colaborativa no auxílio de recomendação personalizada para leitores de mangá / Rodrigo Nativo do Brasil Brochardt. - 2024.
58 f. : il.

Orientador: Cicero Garrozi.
Inclui referências.

Trabalho de Conclusão de Curso (Graduação) - Universidade Federal Rural de Pernambuco, Bacharelado em Sistemas da Informação, Recife, 2024.

1. SVD. 2. coeficiente de correlação de Pearson. 3. sistema de recomendação. 4. rastreador web. 5. mangá. I. Garrozi, Cicero, orient. II. Título

CDD 004

Rodrigo Nativo do Brasil Brochardt

Utilização de Filtragem Colaborativa no auxílio de recomendação personalizada para leitores de mangá

Monografia apresentada ao Curso de Bacharelado em Sistemas de Informação da Universidade Federal Rural de Pernambuco, como requisito parcial para obtenção do título de Bacharel em Sistemas de Informação.

Aprovado em: 04 de março de 2024.

BANCA EXAMINADORA

Cícero Garrozi (Orientador)
Departamento de Estatísitca e Informática
Universidade Federal Rural de Pernambuco

Rodrigo Gabriel Ferreira Soares
Departamento de Estatísitca e Informática
Universidade Federal Rural de Pernambuco

Recife
2024

Resumo

Este trabalho investigou, elaborou e comparou duas abordagens para a geração de recomendações de mangás: o modelo de Decomposição em Valores Singulares (SVD) e o Coeficiente de Correlação de Pearson. A metodologia envolveu a preparação dos dados a partir do desenvolvimento e execução de um rastreador web para extrair informações de obras de mangá e avaliações de um fórum bastante movimentado na internet. As dificuldades que surgem para a aplicabilidade destes métodos de extração de dados, bem como alternativas para lidar com situações de bloqueio da fonte, treinamento dos modelos de recomendação e avaliação de desempenho, foram abordadas, com foco na filtragem colaborativa e recomendações personalizadas para perfis de usuários e para obras de mangá. Na implementação do SVD, foi possível identificar padrões latentes nos dados de avaliação dos usuários, permitindo recomendações personalizadas com base nas preferências individuais a partir do compartilhamento de experiências com perfis similares. No entanto, métricas como o Mean Absolute Error (MAE) e Root Mean Squared Error (RMSE) revelaram a necessidade de refinamento do modelo para melhorar sua precisão, assim como alternativas de implementações para realização de comparações e métricas relacionadas à massa de dados específica utilizada no trabalho. Por sua vez, a abordagem baseada no Coeficiente de Correlação de Pearson priorizou a similaridade entre as avaliações de mangás para gerar recomendações focadas em itens. Embora dependesse significativamente do número de avaliações disponíveis, essa metodologia ofereceu uma lógica direta e válida para recomendações personalizadas a partir dos relacionamentos advindos das avaliações. A conclusão destacou a possibilidade futura de explorar métodos híbridos que combinem as vantagens do SVD e do Coeficiente de Correlação de Pearson, visando alcançar recomendações mais precisas e abrangentes, bem como a possibilidade de validar técnicas que ofereçam abordagens diferentes de recomendação para obter um comparativo palpável. A utilização de dados adicionais reunidos na massa de dados gerada para enriquecer a qualidade das recomendações, a fim de utilizar parâmetros mais detalhados em sua recomendação, assim como a utilização de abordagens indiretas, como por exemplo, a utilização de LLMs para auxiliar no processo de recomendação. Por fim, o trabalho concluiu a importância dos avanços destas tecnologias de recomendação para facilitar a vida do leitor, auxiliando na filtragem de grandes conteúdos oferecidos pela indústria e internet.

Palavras-chave: SVD; coeficiente de correlação de Pearson; sistema de recomendação; rastreador web; mangá.

Abstract

This study investigated, developed, and compared two approaches for generating manga recommendations: the Singular Value Decomposition (SVD) model and the Pearson Correlation Coefficient. The methodology involved data preparation through the development and execution of a web scraper to extract manga information and reviews from a highly active internet forum. Challenges arising in the applicability of these data extraction methods were addressed, along with alternatives for handling source blocking situations, model training, and performance evaluation, focusing on collaborative filtering and personalized recommendations for user profiles and manga works. In the implementation of SVD, latent patterns in user review data were identified, enabling personalized recommendations based on individual preferences through the sharing of experiences with similar profiles. However, metrics such as Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) revealed the need for model refinement to improve its accuracy, as well as alternative implementations for conducting comparisons and metrics related to the specific data mass used in the study. Conversely, the approach based on the Pearson Correlation Coefficient prioritized similarity between manga reviews to generate item-focused recommendations, although it significantly relied on the number of available reviews. This methodology offered a direct and valid logic for personalized recommendations based on relationships derived from reviews. The conclusion highlighted the future possibility of exploring hybrid methods combining the advantages of SVD and the Pearson Correlation Coefficient to achieve more precise and comprehensive recommendations, as well as validating techniques that bring different recommendation approaches for tangible comparison. The utilization of additional data gathered in the generated data mass to enrich the quality of recommendations was suggested, aiming to use more detailed parameters in recommendations, along with the employment of indirect approaches, such as using LLMs to aid in the recommendation process. Finally, the study emphasizes the importance of advancing these recommendation technologies to facilitate readers' lives by assisting in filtering the vast content offered by the industry and the internet.

Keywords: SVD; Pearson correlation coefficient; recommendation system; web scraper; manga.

Lista de Figuras

Lista de Tabelas

Sumário

1. Introdução

Os mangás são uma forma de arte e entretenimento que nasceu no Japão no século XVII e que faz parte da cultura japonesa até hoje. Com o intuito de contar histórias por meio de desenhos e quadrinhos, muitas vezes em preto e branco, os mangás ganharam milhões de fãs em todo o mundo ao longo do tempo, com uma indústria que constantemente lança novidades. Diante dessa vasta oferta de obras, surge o questionamento: como escolher entre tantas opções disponíveis os mangás que mais agradam a cada leitor?

Com uma receita de cerca de 6,8 bilhões de dólares em 2020 (STATISTA, 2021), o mercado de mangás é um dos mais promissores do mundo, com uma previsão de crescimento de 9% ao ano até 2024 (TECHNAVIO, 2022). Diante dessa grande popularização e desenvolvimento da indústria, aliados à enorme quantidade de dados disponíveis na internet sobre o tema e sua comunidade, surge a necessidade de buscar formas de manipular essas informações de maneira rápida e direta ao ponto. Essa atividade, se realizada manualmente, seria inviável devido à enorme quantidade de dados.

Dessa necessidade surgem os sistemas de recomendação, com o propósito de lidar com grandes volumes de dados, um desafio que simples sistemas de recomendação sociais dificilmente conseguiriam enfrentar, como recomendações em clubes de leitura, recomendações a partir de profissionais como bibliotecários e livreiros que possuíam conhecimento profundo sobre livros ou a partir de críticas literárias e prêmios as quais eram conquistados por obras. Esses modernos sistemas fornecem resultados personalizados baseados nas preferências derivadas do perfil do usuário, a partir de seus interesses ou interações dentro de um sistema, utilizando frequentemente técnicas de agrupamento, predição e correlacionamento de informações dos itens e usuários (F.O. Isinkaye et al, 2015).

Foi realizada uma pesquisa como parte do projeto com o objetivo de compreender o contexto e as necessidades as quais os leitores de mangá possuem em relação aos sistemas de recomendação personalizados, localizado no Capítulo 3, sobre a Pesquisa de Mercado. A partir de um questionário distribuído em comunidades de leitores, foram reunidas 65 respostas para 28 questões que revelaram informações relevantes e usadas como base para o desenvolvimento de um sistema de recomendação eficaz.

Seguindo esta premissa, o objetivo do trabalho é executar as fases descritas por (F.O. Isinkaye et al, 2015), que são as fases de extração de informações, aprendizado e predição, a fim de implementar um algoritmo de recomendação colaborativa de maneira que, a partir de correlacionamentos, entenda o perfil dos usuários, possibilitando a recomendação personalizada de leitura.

Como ponto principal para o desenvolvimento do trabalho, serão utilizadas técnicas de coleta das informações necessárias por meio da criação de um rastreador de informações web (web crawler) com o objetivo de gerar um banco de informações com dados que reflitam as interações sociais e preferências dos usuários comuns deste tipo de comunidade de entretenimento. A partir disso, será possível realizar o tratamento dos dados e ter um ponto de partida na realização da predição de sugestões por meio da filtragem colaborativa, resultando em sugestões mais realistas e que tornem a experiência de leitura mais satisfatória.

Portanto, destaca-se que os desenvolvedores de sistemas de recomendação enfrentam desafios como a qualidade dos dados disponíveis, a utilização de técnicas de recomendação

por filtragem adequadas para cada contexto e a interpretação dos resultados obtidos. Ainda assim, espera-se que este projeto contribua para o avanço dos estudos na área de sistemas de recomendação, fornecendo um estudo que auxilie em futuros trabalhos e forneça uma ferramenta com resultados úteis para a comunidade de leitores de mangás e manhwas.

2. Referencial teórico

Para tornar mais compreensíveis os elementos que serão utilizados na implementação do sistema de recomendação de mangás deste trabalho, serão abordados os principais conceitos entre as técnicas de filtragem colaborativa mais utilizadas. Serão discutidas as justificativas de quais cenários são mais adequados para cada tipo de implementação, a importância da qualidade das informações utilizadas para o aprendizado e predição do sistema de informação, assim como maneiras de limpar e preparar os dados que serão usados como base. Além disso, serão esclarecidos quais benefícios as técnicas de filtragem existentes oferecem e quais são suas desvantagens ao contribuir para a qualidade dos resultados da filtragem de informação destinada à predição das preferências do usuário.

2.1 Sistemas de recomendação

Como parte principal deste trabalho, as primeiras definições dos sistemas de recomendação foram desenvolvidas inicialmente nos anos 90 por Jussi Karlgren como um estudo de “estante digital” que seria capaz de auxiliar pessoas a encontrar livros relevantes em uma biblioteca digital. Foi proposta uma álgebra para representar e definir as preferências e características dos livros que os usuários buscavam.

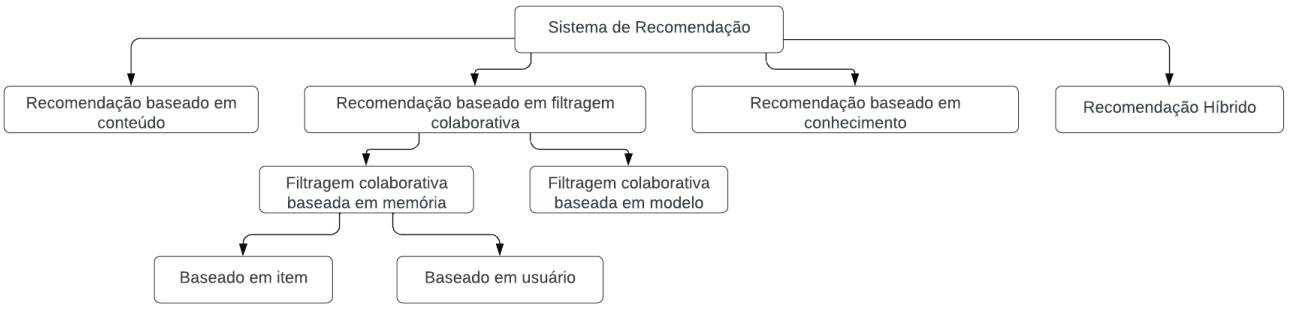
Tal álgebra proposta (An Algebra for Recommendations, Jussi Karlgren, 1990) representa a definição das preferências e características dos livros que os usuários buscavam, sendo uma linguagem matemática que mapeia informações como gêneros, autores e ano de publicação, por exemplo. A partir disso, seria capaz de realizar recomendações personalizadas relacionando a proximidade entre documentos. Karlgren define em seu trabalho que proximidade é a medida de similaridade entre dois documentos, que para o nosso contexto seria a similaridade entre obras de mangá, baseado nos interesses do leitor. Ele buscou definir graus de interesse e relacionamentos de um documento para os outros, permitindo, desta forma, a possível recomendação.

A utilização dos sistemas de recomendação ao longo do tempo foi se espalhando e não ficou apenas restrita a recomendações de livros em bibliotecas digitais, mas também teve bastante importância para recomendações de produtos em e-commerce e mais adiante também teve bastante impacto na área de publicidade, mídias e notícias.

Sendo uma ferramenta tão versátil, a utilização de recomendações e sugestões a partir de massas de dados acumuladas pode ser implementada para auxiliar diversos nichos de conteúdo digital, com a finalidade de filtrar e entregar resultados aos usuários de uma maneira relevante, que possua coerência e, como (F.O. Isinkaye. 2015) descreve em seu trabalho, oferece redução de custo relacionado à busca e seleção de itens.

Com o passar dos anos, foram surgindo mais trabalhos a respeito, resultando no desenvolvimento de diversos tipos de sistemas de recomendação, os quais possuem diferentes implementações e benefícios a serem explorados. Como ilustrado na Figura 1, os sistemas de recomendação são divididos em quatro categorias: Sistema de recomendação baseado em conteúdo, Sistema de recomendação de filtragem colaborativa, sistema de recomendação baseada em conhecimento e sistemas de recomendação híbridos (Khanal et al. 2019).

Figura 1 - Representação dos tipos de sistema de recomendação.

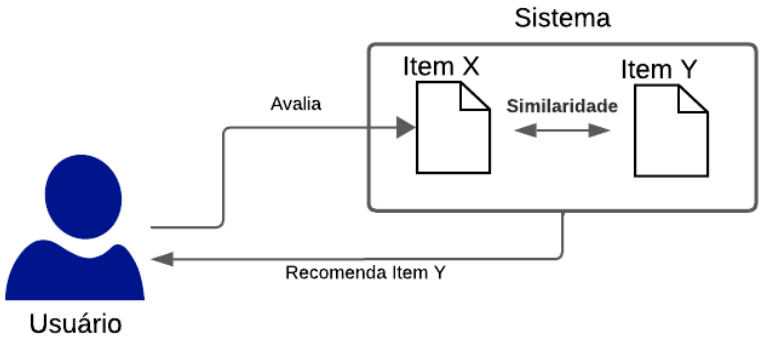


Fonte: autoria própria.

2.1.1 Sistema de recomendação baseado em conteúdo

Os métodos de recomendação por conteúdo são realizados a partir das características dos itens que estão sendo pesquisados dentro de determinado sistema. A partir dos atributos, o sistema é capaz de realizar a predição de possíveis itens desejados pelo usuário utilizando como base o seu histórico de consumo de conteúdo. Ou seja, são analisadas as avaliações feitas pelo usuário em pesquisas anteriores sobre produtos específicos para assimilar outros conteúdos que possuem as mesmas características, como representado na Figura 2. Por natureza, este tipo de sistema é capaz de se adaptar às novas tendências das preferências do usuário durante o decorrer do tempo, já que suas sugestões são feitas como reação às interações do usuário perante os itens.

Figura 2 - Representação de recomendação baseado em conteúdo.



Fonte: autoria própria.

Trazendo para o contexto deste trabalho, um exemplo de utilização da recomendação baseada em conteúdo seria a procura de similaridade entre obras já lidas e avaliadas pelo determinado usuário por meio das características que definem a obra já avaliada. Quanto mais profundas esses mapeamentos, mais precisa a recomendação se torna. Por exemplo, características como gênero, ano de lançamento, nome dos autores, nome dos personagens, demografia à qual a obra pertence, entre muitas outras informações, compõem os atributos da

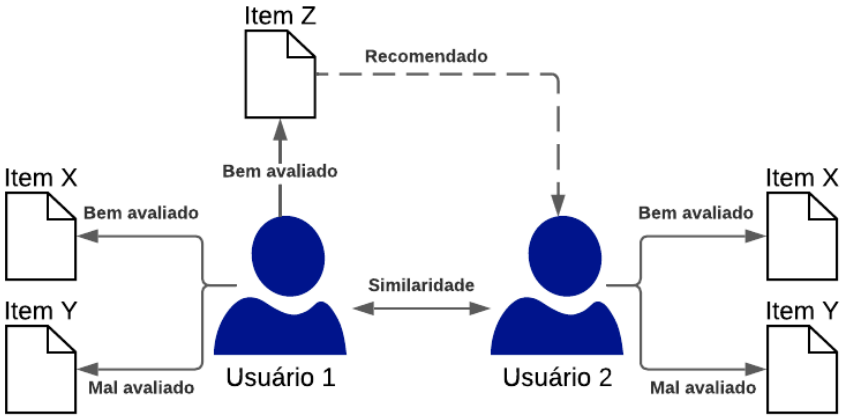
obra e serão utilizadas para a sugestão. A partir disso, o sistema busca prever ao usuário quais obras mais irão agradá-lo.

Porém, segundo o estudo de Michael J. Pazzani & Daniel Billsus em (The Adaptive Web: Methods and Strategies of Web Personalization 2007), tal modelo possui certas desvantagens ao ser implementado. Uma delas é a falta de diversidade ou sugestão de novos itens, ocasionada pela recomendação apenas dos conteúdos que possuem exatamente as mesmas características. Além disso, esse modelo possui muita dependência em relação à disponibilidade de informações de qualidade para facilitar a identificação de tais indicadores. Outro ponto limitante é a incapacidade de utilizar informações da experiência de outros usuários para refinar as sugestões, já que está limitado apenas ao histórico de ações do indivíduo que está utilizando.

2.1.2 Sistema de recomendação de filtragem colaborativa

O modelo de recomendação de filtragem colaborativa tem como premissa basear-se no comportamento de usuários com perfis de preferências similares para compartilhar entre eles os conteúdos que cada um demonstrou interesse (Figura 3). Isso é realizado através da correlação entre os itens bem avaliados de cada usuário e identificando quais itens ainda não foram avaliados por um dos usuários, mas que poderiam ser tomados como sugestões pelo sistema.

Figura 3 - Representação de recomendação por filtragem colaborativa.



Fonte: autoria própria.

Este modelo de recomendação toma um rumo diferente em comparação à recomendação baseada em conteúdo. Em vez de extrair e assimilar características do conteúdo que está sendo avaliado, ele realiza a correlação das avaliações entre usuários. Isso é um ponto forte ao combater a escassez de informação que poderia existir na base de dados. No caso da recomendação de filtragem colaborativa, ao invés do sistema recomendar obras com características semelhantes com base no histórico de avaliações do usuário, as sugestões serão baseadas nas experiências e avaliações de todos os usuários que possuem o perfil com as mesmas preferências. Isso pode tornar as sugestões mais realistas, já que estão tomando como base outros usuários e interações reais dentro do sistema, colaborando para recomendações

mais diversas que não ficariam restritas apenas às predições relacionadas às obras específicas que o usuário avaliou no passado.

Em contrapartida, as sugestões neste tipo de modelo ficam sujeitas à inconsistência em relação à confiabilidade das recomendações, pois podem ser afetadas pela possível manipulação das avaliações de usuários que foram utilizados como ponto de partida para as correlações. Além disso, há a questão do "início frio", como mencionado por (F.O. Isinkaye et al, 2015), que se refere a cenários nos quais não existem avaliações até o momento para serem tomadas como ponto inicial de recomendação aos usuários, resultando em possíveis incoerências.

2.1.3 Sistema de recomendação baseado em conhecimento

O modelo de sistema de recomendação baseado em conhecimento utiliza conceitos sobre um conteúdo para sugerir ao usuário, aplicando relações entre contextos utilizando mineração em sequenciamento (Tarus et al., 2017), com o objetivo de entender padrões dentro dos dados em questão. Ele identifica informações consideradas relevantes e aplica critérios pré-definidos ou toma como verdade casos de recomendação anteriores semelhantes como condição para aplicar a recomendação.

Diferentemente das recomendações baseadas em conteúdo, que buscam apenas a similaridade entre atributos de um item, a recomendação baseada em conhecimento tenta considerar o contexto e o significado dos itens, podendo incorporar conhecimentos específicos como a história de uma obra, enredo, temas complementares e críticas dos leitores.

O diferencial desse tipo de modelo é a possibilidade de recomendar conteúdo que não possui uma alta demanda ou histórico de consumo definido, algo que é controverso em relação aos modelos anteriores, que dependem de informações de qualidade previamente disponíveis para realizar uma boa sugestão.

No entanto, segundo (Khanal et al., 2019), esses sistemas possuem algumas desvantagens que os tornam viáveis apenas em alguns cenários específicos, como por exemplo um sistema de recomendação de viagens, que necessita sugerir destinos de viagem baseado em critérios que definem uma boa viagem, como o clima preferido, atividades interessantes e orçamento, ou seja, com um nível de granularidade alta em quesito de compreensão da situação. Eles necessitam de um entendimento profundo do contexto desses conteúdos para definir os critérios que atendam às preferências dos usuários, além da necessidade de manter e atualizar constantemente esses critérios para se adaptar às mudanças nos dados. Os sistemas de recomendação baseados em conhecimento não utilizam o histórico de avaliações ou ações anteriores do usuário, o que torna a recomendação restrita e a filtragem rigorosa, reduzindo a diversidade das sugestões. Além disso, há a complexidade necessária para a obtenção dos indicadores que poderiam ser utilizados para gerar o modelo de predição.

2.1.4 Sistema de recomendação híbrido

Compreendendo todas as possibilidades, limitações e benefícios de cada categoria de sistema de recomendação descrita até agora, os sistemas de recomendação híbridos têm a característica de unir conceitos de ambos os modelos para mitigar e flexibilizar as

desvantagens e vantagens de cada abordagem, permitindo adaptação ao contexto e domínio em que o sistema está sendo empregado. Por exemplo, eles podem utilizar indicadores como avaliações explícitas do usuário e predição baseada no histórico de ações anteriores. Isso combina as qualidades de identificar conteúdos com características semelhantes aos itens já vistos pelo usuário, além de agregar avaliações de outros usuários com perfis semelhantes para potencializar a diversidade e qualidade das sugestões.

Portanto, é essencial estudar o contexto em que o sistema de recomendação será introduzido, avaliando os indicadores essenciais e os mais impactantes para uma sugestão de qualidade. No caso de um ambiente de e-learning, isso envolve considerar as características do conteúdo disponível e aplicar uma combinação de técnicas, como a indicação de avaliação individual e colaborativa em cada obra, adaptação de sugestões relevantes com base nas características explícitas do perfil do usuário e histórico de ações, e aproveitar ao máximo a natureza dos dados utilizados para a aplicação dos modelos de recomendação.

2.2 Modelos de recomendação por filtragem colaborativa

Diante da análise das diferentes abordagens de sistemas de recomendação e considerando que este projeto utilizará uma massa de dados existente que reflete as interações reais entre leitores em uma comunidade de avaliações de mangá, com informações sobre os mangás, usuários e suas respectivas avaliações, a implementação será realizada com base na abordagem de recomendações baseadas em filtragem colaborativa. Essa escolha é respaldada pelo fato de que, como afirmado por (F.O. Isinkaye et al, 2015), essa é a abordagem mais comum e bem desenvolvida na área de recomendação, justamente por sua aplicação prática ser bastante eficaz em grandes áreas como o e-commerce, devido a sua capacidade de recomendação direcionada e diversificada. No contexto deste trabalho, essa abordagem pode trazer benefícios significativos, pois utiliza uma massa de dados real com interações existentes entre usuários e obras, minimizando problemas como o "início frio" no momento das previsões.

Dentro da abordagem de recomendação por filtragem colaborativa, existem duas grandes categorias: a abordagem baseada em memória e a abordagem baseada em modelo. Ambas compartilham a premissa de utilizar as experiências de usuários com perfis de preferência semelhantes para facilitar a predição de novos conteúdos.

2.2.1 Abordagem de filtragem baseada em memória

As técnicas de filtragem colaborativa baseadas em memória funcionam por meio do cálculo da similaridade entre itens ou usuários, baseando-se no histórico de interações do mesmo, ou seja, no contexto de leitores de mangá, seria utilizado as avaliações realizadas aos itens ao longo do tempo para identificar graus de similaridade, a fim de sugerir itens que ainda não foram avaliados por eles e que possivelmente seriam de sua preferência.

A filtragem baseada em memória é dividida em duas frentes, a filtragem orientada ao usuário e a filtragem orientada ao item. Sobre a filtragem orientada ao usuário, seu comportamento é realizar sugestões a partir da busca por usuários que possuem o perfil semelhante, baseado em suas avaliações, recomendando itens resultantes desse

compartilhamento de preferências entre eles. No geral, este tipo de filtragem por usuário se beneficia nos cenários em que a quantidade de avaliações do usuário é pouca para a quantidade de itens existentes, já que, a partir dos perfis semelhantes, ocorre uma inferência de sugestão. Este tipo de abordagem também é mais versátil em cenários em que as mudanças de preferências ocorrem mais repentinamente, devido ao cálculo de interações passadas entre os usuários semelhantes.

Contudo, essas características, como também citadas por (Miguel Angelo, 2011) em seu trabalho sobre filtragem colaborativa, também trazem visíveis desvantagens ao serem implementadas, uma delas é o problema de custo computacional derivado da necessidade de realizar as correlações entre os pares de usuários, tal custo cresce quanto maior for a massa de dados disponível, resultando no aumento de sua complexidade, a outra desvantagem referente à implementação da filtragem orientada ao usuário está relacionada ao "início frio", já que novos usuários dentro do sistema não terão dificuldade em receber sugestões coerentes de itens, por consequência da pouca correlação entre outros perfis de usuários.

Já a filtragem orientada a itens calcula a similaridade entre itens em relação às avaliações feitas pelos usuários, realizando a recomendação de itens similares aos que o determinado usuário avaliou anteriormente. Suas vantagens, diferente da filtragem orientada ao usuário, resolvem o problema de escalabilidade, já que o cálculo de similaridade entre itens é menos dinâmico, podendo ser pré-calculado, economizando tempo e necessidade de poder computacional e tornando o processo de recomendação mais estável, facilitando também quando surge a necessidade de atualizar a massa de dados com novos itens, podendo incorporar esses novos conteúdos sem influenciar na estrutura já existente das correlações. Porém, como desvantagem, a filtragem orientada em item pode sofrer inconsistências nas suas sugestões em cenários em que existem muitos itens para recomendar, porém poucas avaliações de usuários realizadas, tornando difícil o cálculo de similaridade para itens com poucas avaliações, como também é possível surgir problemas em realizar recomendações de itens novos dentro do sistema, já que, como tal recomendação leva em conta interações de avaliações passadas, é necessário um certo grau de similaridade para que seja integrado como possibilidade de recomendação.

Segundo (Khanal, 2019), como as abordagens de filtragem colaborativa utilizam a massa de dados para a extração da informação (realização da sugestão), e por ser um processo bastante lento, comumente não é preferível a utilização de filtragem baseada em memória, diferente da abordagem baseada em modelo, este tipo de processo acaba sendo mais prolongado, já que eles utilizam a memória completa da massa de dados para fazer as recomendações, quanto maior for a quantidade de relações entre itens ou usuários necessárias, maior será a necessidade de processamento. Porém, a partir deste entendimento, é possível também compreender as vantagens da utilização da filtragem baseada em memória, como este tipo de abordagem realiza os cálculos diretamente dos dados mais recentes, em cenários em que os gostos dos usuários mudam rapidamente, é possível ter uma adaptabilidade maior em relação às mudanças na massa de dados, tornando-se um ponto positivo para a implementação deste tipo de sistema de recomendação.

2.2.2 Abordagem de filtragem baseada em modelo

Os princípios por trás dos modelos de abordagem de filtragem baseada em modelo têm como objetivo a predição de itens pelos quais os usuários possuem preferência por meio da identificação de padrões de outros usuários com perfis semelhantes a partir do histórico de avaliações disponíveis. Utilizando primariamente modelos matemáticos ou estatísticos, sua implementação treina um modelo a partir de uma porção da massa de dados para aprender como funcionam as avaliações de usuários para itens e, a partir disso, aplica a predição em avaliações ainda não realizadas em itens a fim de sugerir ao usuário os itens com possíveis maiores notas.

A partir do trabalho de (Su, Xiaoyuan, 2009), o qual realiza um estudo sobre as técnicas de filtragem colaborativa, é afirmado que a abordagem da filtragem baseada em modelo foi elaborada com o intuito de resolver os problemas emergentes da abordagem baseada em memória, sendo eles o problema de esparsidade dos dados, a dificuldade de escalar o modelo para massas de dados cada vez maiores e os problemas de "início frio" aos itens e usuários que ainda não possuem histórico de avaliações. Porém, estas abordagens baseadas em modelo necessitam de uma grande quantidade de dados para a realização de treinamento de forma eficaz, a obtenção desses dados pode ser difícil em situações nas quais existem poucos usuários que usam o sistema. Além disso, por utilizar modelos matemáticos ou estatísticos mais complexos em sua resolução, já que é comum aplicar técnicas avançadas de entendimento de padrões e multidimensionalidade, este tipo de abordagem acaba sendo mais trabalhoso de entender e adaptar no sistema.

2.2.3 Single Value Decomposition (SVD)

O SVD é uma ferramenta matemática bastante utilizada em diversas áreas por ser capaz de oferecer três características, redução de dimensionalidade, redução de ruídos e recomendação (Jason Browniee, 2019).

A partir da manipulação e fatoração de matrizes, o SVD é capaz de, matematicamente, decompor uma matriz em três componentes fundamentais capazes de representar eficientemente os dados as quais estão sendo relacionados na matriz original.

Sua decomposição dar-se na fórmula:

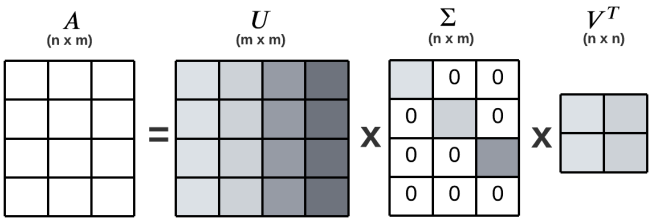
$$A = U \cdot \Sigma \cdot V^T$$, onde

- A representa a matriz $m\ x\ n$ a qual será decomposta,
- U é uma matriz $m\ x\ m$, contendo os vetores singulares à esquerda da matriz original,
- Σ representa uma matriz diagonal $m\ x\ n$, as quais são as dimensões da matriz original,
- V^T representa uma matriz transposta de $n\ x\ n$.

Com estes três componentes resultantes, é possível aplicar suas características em diversas áreas. A matriz U, a matriz diagonal Σ e a matriz transposta VT oferecem essa característica, sua capacidade de representar os dados e capturar informações essenciais para reduzir a complexidade dos dados, indicar a importância relativa de cada dimensão e descrever como as características se relacionam entre si. Juntas, essas partes permitem a

extração de insights como reconhecimento de padrões, processamento e compressão das informações utilizadas. Uma visualização agradável dessas relações pode ser encontrada na Figura 4, que apresenta um diagrama da definição da fórmula do SVD.

Figura 4 - Diagrama da definição matemática do SVD.



Fonte: autoria própria. (baseada em AskPython, 2020).

No caso dos sistemas de recomendação, o papel do SVD é identificar, a partir das relações estabelecidas entre colunas e linhas dentro da matriz, e ser capaz de replicar o comportamento das informações, utilizando o reconhecimento de padrões, para preencher relacionamentos de usuários e itens que ainda não foram feitos. Este preenchimento dos relacionamentos ausentes na matriz então pode ser utilizado como recomendação em que muitas vezes reflete eficientemente a similaridade entre itens de forma linear.

2.2.4 Coeficiente de Correlação de Pearson

O coeficiente de correlação de Pearson, uma medida estatística fundamental, é amplamente empregado em diversas disciplinas devido à sua habilidade em quantificar a relação linear entre duas variáveis (Bruno Oliveira, 2019). Esta medida é representada por um valor que indica o quão similar os elementos comparados são baseados na matriz de relacionamento entre eles. Para calcular o coeficiente de correlação de Pearson entre duas variáveis x e y, a fórmula é a seguinte:

$$r = \frac{\sum_{i=1}^{n}(x_i - X)(y_i - Y)}{\sqrt{[\sum_{i=1}^{n} x_i - X]^2 [\sum_{i=1}^{n} y_i - Y]^2}}$$, onde:

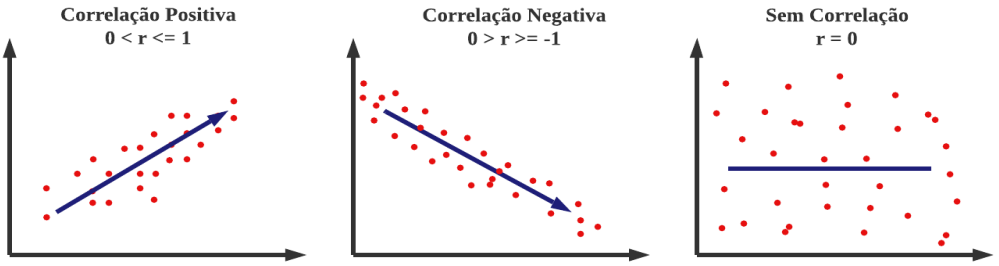
- r é o coeficiente de correlação de Pearson;
- n é o número total de observações;
- x e y são as variáveis analisadas;
- X e Y são as médias das variáveis x e y.

A fórmula do coeficiente compara as diferenças entre os valores observados e as médias das variáveis X e Y, medindo o grau de covariância, ou seja, avalia as interações de dependência entre elas, sendo elas positivas ou negativas. Os resultados são então

normalizados pela raiz quadrada dos produtos das variâncias de x e y , o que permite uma interpretação do grau de correlação.

O coeficiente de correlação de Pearson gera um valor que varia entre -1 e 1, onde -1 indica uma correlação negativa perfeita, 0 indica ausência de correlação e 1 indica uma correlação positiva perfeita, como apresentado na Figura 5.

Figura 5 - Definição dos estados da medida de Correlação de Pearson.



Fonte: autoria própria. (baseada em statisticashowto).

A partir desta medida é possível analisar estatisticamente dados precisamente, resultando em uma fonte comparativa baseada na similaridade dos itens. Sua aplicação é ampla, podendo ser utilizada para interpretar padrões e tendências nos dados.

2.3 Biblioteca Surprise

Para a implementação de modelos de sistema de recomendação, é comum a utilização de bibliotecas Python capazes de fornecer ferramentas que facilitam o desenvolvimento desse tipo de tecnologia. A biblioteca Surprise (disponível online) é uma delas, sendo uma ferramenta poderosa para a construção, treinamento e aplicação de modelos de recomendação de filtragem colaborativa. Nela, é possível utilizar diversos algoritmos do ecossistema de recomendação. A biblioteca oferece vários modelos implementados dos algoritmos mais comuns de recomendação, incluindo algoritmos baseados em memória, algoritmos baseados em filtragem colaborativa e algoritmos baseados em conteúdo. Além disso, oferece métodos de validação da precisão das predições geradas pelos algoritmos e integração com bibliotecas essenciais para esse tipo de desenvolvimento, como a biblioteca NumPy para diversas operações matemáticas e o Pandas para manipulação de massas de dados.

2.4 Tratamento de datasets

Como ponto de partida de todo projeto relacionado a sistemas de recomendação, a massa de dados é crucial para sua realização, sendo um dos parâmetros que deve ser buscado com qualidade a fim de garantir uma recomendação precisa. A qualidade das informações extraídas para o treinamento do modelo de aprendizado de máquina influencia muito no resultado desse tipo de projeto (Kavika Roy, 2023).

Datasets são conjuntos de dados de determinada natureza que representam alguma entidade, como a abstração de características de um leitor ou informações de uma obra

literária. Dependendo da fonte em que estão sendo obtidas tais informações, é possível que haja inconsistências, informações desnecessárias para o tipo de atuação que será realizada e falta de estruturação. Por isso, é necessário, principalmente em casos de realização de cálculos estatísticos e de similaridade, garantir a qualidade e organização das informações armazenadas para não prejudicar o desempenho e a confiabilidade do modelo preditivo.

Uma das alternativas disponíveis, de maneira open-source, como utilitários relacionados à manipulação de massas de dados, é a biblioteca de Python chamada Pandas. Nela, são disponibilizadas inúmeras ferramentas capazes de manipular arquivos CSV, extrair informações, realizar operações complexas e interagir com várias outras bibliotecas (W. McKinney, 2018), principalmente quando se trata de aprendizado de máquina e seus derivados, aumentando a produtividade em atividades de limpeza e estudo exploratório de massas de dados para serem utilizadas em estudos e implementações.

Na implementação de sistemas de recomendação, a manipulação de dados envolve diversos processos para garantir a qualidade e a integridade dos dados utilizados nos modelos de recomendação, esse processo de tratamento de dados geralmente segue quatro passos importantes: carregamento de dados, pré-processamento, separação de dados e manipulação de dados.

Na etapa de carregamento dos dados, utiliza-se a biblioteca Pandas para importar o arquivo CSV contendo as informações, neste caso seria as informações dos mangás ou dos usuários as quais realizaram avaliações em obras, tais como identificadores únicos, títulos e descrições. Essa etapa é fundamental para que seja obtida uma visão completa dos dados e prepará-los para análises.

Já na etapa de pré-processamento, busca-se identificar e corrigir valores ausentes, resolver erros de formatação e eliminar duplicatas quando necessário, garantindo a integridade e a qualidade dos dados, prevenindo possíveis distorções nos resultados das recomendações.

Em seguida, é realizada a separação dos dados em conjuntos de treinamento e teste, reservando uma proporção específica para o treinamento dos modelos e utilizando o restante para avaliação de desempenho, essa divisão permite uma avaliação do desempenho dos modelos, podendo ajudar no refinamento dos algoritmos.

Por fim, na etapa de manipulação de dados, aplica-se técnicas como filtragem, ordenação e agregação para prontificar os dados conforme as exigências dos modelos, como por exemplo a seleção das colunas que serão utilizadas para recomendação, no caso da filtragem colaborativa seriam as notas de avaliação, os identificadores das obras e dos usuários.

2.5 Métricas

Dentro da área de sistemas de recomendação, a avaliação do desempenho dos algoritmos é essencial para medir sua acurácia. Duas métricas comumente utilizadas são o Erro Médio Absoluto (MAE) e o Erro Quadrático Médio (RMSE), que fornecem insights valiosos sobre a qualidade das predições.

O MAE é uma medida que quantifica a média das diferenças absolutas entre os valores previstos e os valores reais. Isso é feito calculando a média das diferenças absolutas

entre cada par correspondente de valores sugeridos e valores reais. Em cenários de regressão, o MAE é particularmente útil e fornece uma interpretação direta: quanto menor o resultado, mais próxima a predição está do valor real.

A fórmula para calcular o erro médio absoluto é dado por:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right|$$, onde:

- n é o número total de pontos de dados,
- y é o valor real para o ponto de dado i,
- e $\hat{y}$ é o valor predito para o ponto de dado i.

A partir disso é realizada a soma de todos os pontos e calculado o valor absoluto da diferença entre os valores previstos e preditos.

Já o RMSE é utilizado em conjunto com o MAE para aprofundar a análise de desempenho. Esta medida representa a dispersão dos erros de previsão e é sensível a desvios outliers. O cálculo do RMSE envolve a raiz quadrada da média dos quadrados das diferenças entre os valores sugeridos pelo algoritmo e os valores reais. Assim como no MAE, um resultado menor indica uma predição mais próxima dos valores reais.

A fórmula para calcular o erro quadrático médio é dado por:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$, onde:

- n é o número total de pontos de dados,
- y é o valor real para o ponto de dado i,
- e $\hat{y}$ é o valor predito para o ponto de dado i.

A partir disso é realizada a soma de todos os pontos e calculado o valor quadrático da diferença entre os valores previstos e preditos. Após isso, é tirada a raiz quadrada da média calculada a partir de todos os pontos de dados.

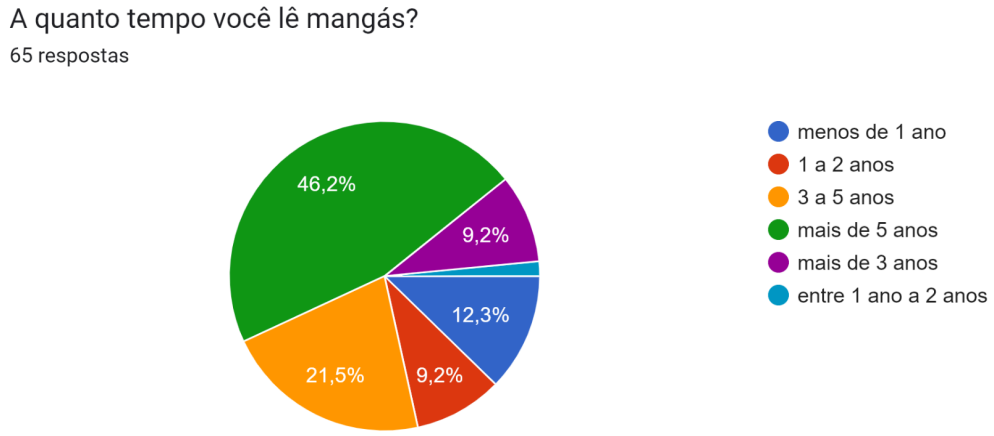
Ambas as métricas, MAE e RMSE, desempenham um papel crucial na avaliação de sistemas de recomendação, oferecendo uma visão abrangente da precisão das predições e da dispersão dos erros. Essas medidas são ferramentas valiosas para comparar e validar o desempenho de diferentes algoritmos, contribuindo para o aprimoramento contínuo desses sistemas. (Referência: Sallam, R. M. et al, 2015).

3. Pesquisa de mercado

Foi realizado um estudo preliminar para entender os interesses dos leitores de mangá relacionados aos sistemas de recomendação de leitura personalizada. Os participantes foram questionados sobre as fontes mais comuns utilizadas para encontrar novas obras para leitura, se essas fontes utilizam, de alguma forma, mecanismos com algoritmos de recomendação e quais são suas opiniões sobre a disseminação dessas ferramentas dentro da comunidade de leitores de mangá. Esse estudo preliminar, que serviu de motivação para o desenvolvimento do trabalho em questão, consistiu na aplicação de um questionário online com 28 perguntas, distribuído por usuários em comunidades online no Discord e no Reddit, onde os leitores de mangás costumam interagir e trocar informações sobre as obras que leem. O estudo recebeu 65 respostas de voluntários, cujas respostas foram analisadas para extrair insights sobre o perfil e as necessidades dos usuários potenciais do sistema de recomendação, com os principais pontos observados a seguir:

Os resultados do questionário mostraram que grande parte da amostra de leitores começa desde cedo. Como observado no gráfico da Figura 6, a maioria dos voluntários já pratica a leitura de mangás há no mínimo 3 anos, demonstrando um hábito de leitura por longos períodos durante suas vidas. Esse padrão sugere um forte vínculo com a cultura dos mangás.

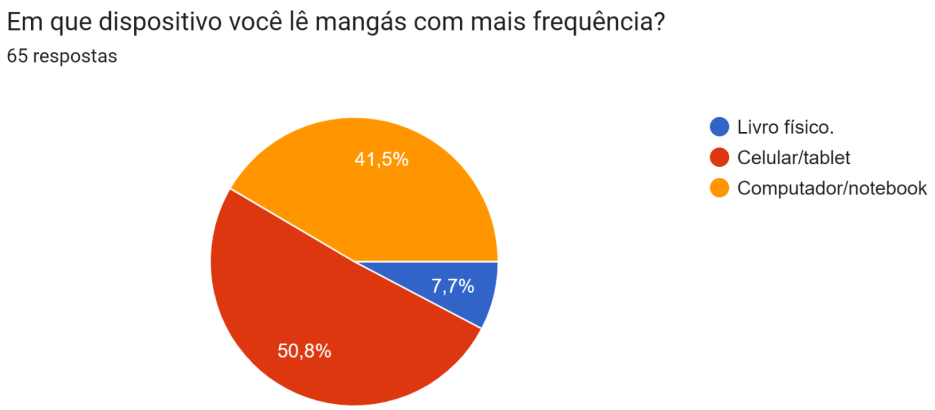
Figura 6 - Distribuição do tempo de leitura.



Fonte: autoria própria.

Algumas das perguntas realizadas também foram feitas para entender o ambiente no qual os leitores estão inseridos e quais ferramentas são mais utilizadas para ler e acompanhar suas séries de mangá favoritas. Essa evidência revela como esse nicho está conectado às tecnologias digitais (Figura 7), uma vez que apenas 7,7% afirmam que preferem ler obras na forma física.

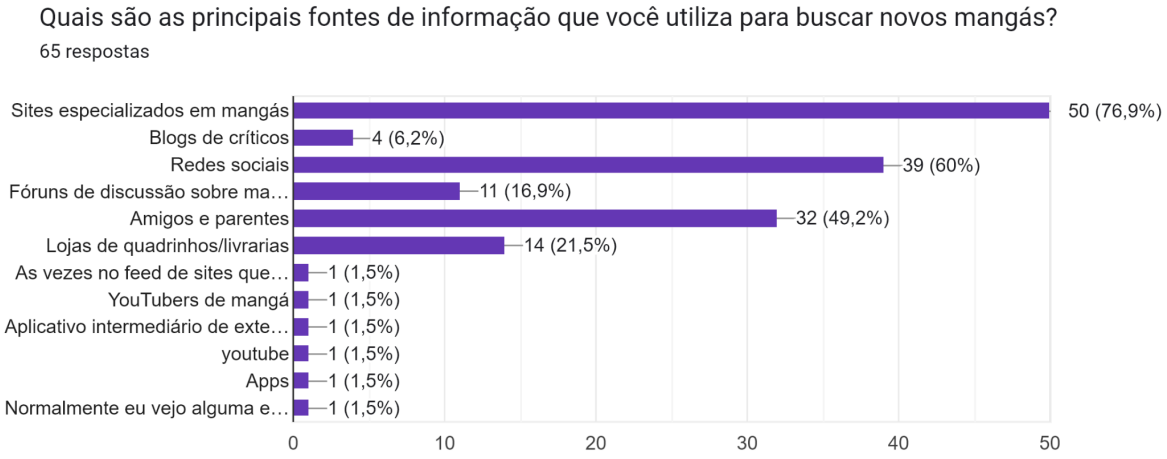
Figura 7 - Distribuição de dispositivos mais usados em leitura.



Fonte: autoria própria.

As fontes mais utilizadas para receber recomendações e ficar atualizado sobre obras populares, de qualidade e lançamentos, conforme as informações coletadas dos leitores na pesquisa (Figura 8), são os sites especializados em mangás e as comunidades em redes sociais que divulgam informações.

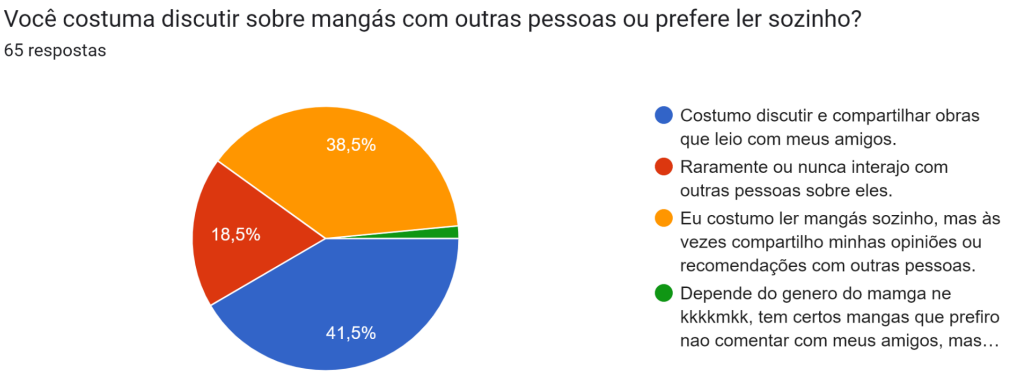
Figura 8 - Distribuição das fontes mais usadas.



Fonte: autoria própria.

O comportamento social predominante para esse tipo de público, como observado na Figura 9, é a tendência de compartilhar e discutir sobre o tema com outras pessoas dentro do círculo social. Isso pode indicar a força do ato de recomendação de leitura por meio da experiência pessoal, uma atividade que sempre foi bastante empregada, especialmente antes do surgimento dos sistemas de recomendação. Anteriormente, essa recomendação era realizada por meio de análise e pesquisa humanas, algo que ainda é praticado hoje em dia, embora com menos dificuldade graças ao auxílio das tecnologias de busca e indexação, como o Google.

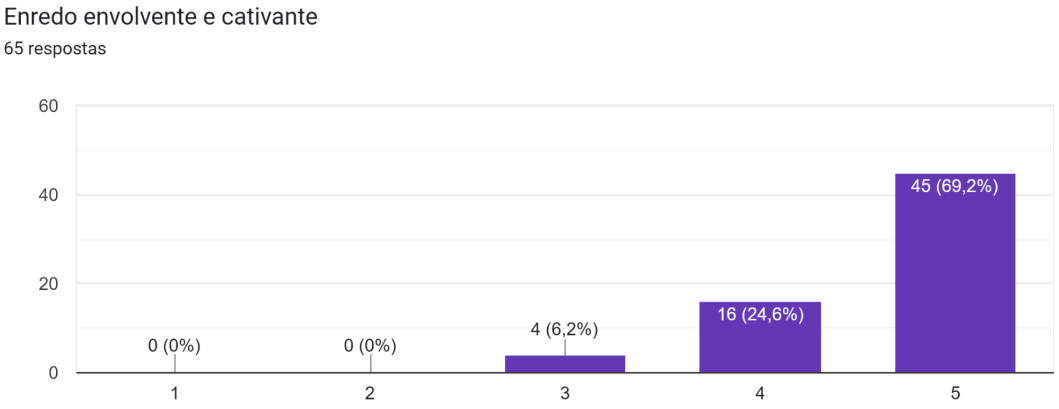
Figura 9 - Distribuição das interações sociais no nicho



Fonte: autoria própria.

Na pesquisa, os voluntários também relataram algumas frustrações ao procurar por novos mangás, como a dificuldade em encontrar obras com temas específicos, a qualidade das traduções e das imagens, e a falta de críticas confiáveis. Durante a compreensão das características mais valiosas, a maioria dos usuários relatou que valoriza principalmente o enredo, como é evidenciado na Figura 10.

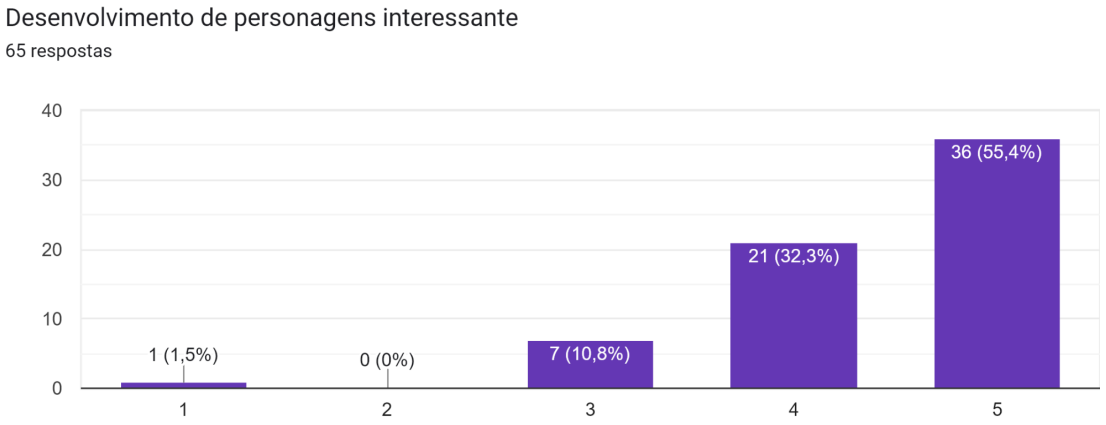
Figura 10 - Distribuição da relevante do enredo.



Fonte: autoria própria.

E também, como demonstrado no gráfico da Figura 11, o desenvolvimento individual dos personagens é considerado um elemento de qualidade em um mangá e um ponto de importância ao recomendar obras para outras pessoas ou buscar novos possíveis interesses para si próprio.

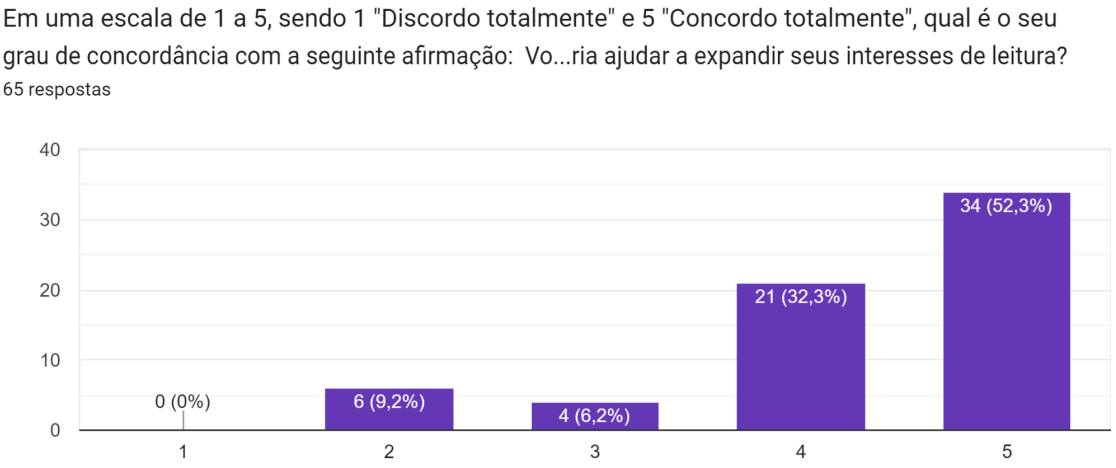
Figura 11 - Distribuição da relevante dos personagens.



Fonte: autoria própria.

Uma das perguntas principais do estudo preliminar está relacionada ao questionamento levantado na frase: “Você acredita que um sistema de recomendação de mangás poderia ajudar a expandir seus interesses de leitura?” Essa pergunta gerou respostas que mostram um favoritismo por parte de mais de 60% da amostra que participou da pesquisa (Figura 12), afirmando o quanto o nicho de leitores de mangás está disposto a experimentar e utilizar novas abordagens para auxiliar nas sugestões personalizadas de obras que se encaixem em seus perfis.

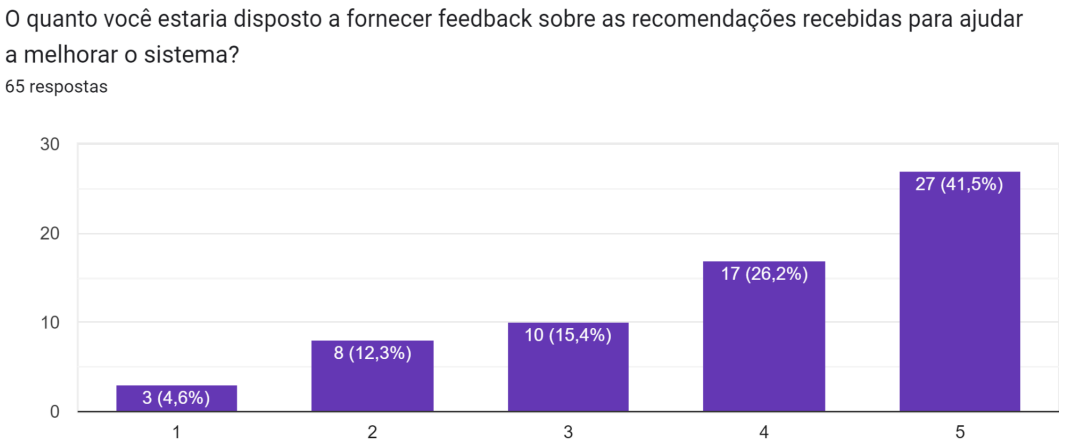
Figura 12 - Distribuição da opinião sobre relevância dos sistemas de recomendação.



Fonte: autoria própria

E, complementando a afirmação sobre o favorecimento dos leitores de mangá na utilização de sistemas de recomendação como auxílio na recomendação de obras personalizadas, a Figura 13 revela que um pouco mais de 67% dos voluntários na pesquisa estão dispostos a fornecer feedback para o sistema a fim de aprimorar o processo de recomendação para eles mesmos. Essa ação poderia ser aproveitada por meio de ferramentas implícitas para auxiliar no treinamento e na precisão do sistema de recomendação.

Figura 13 - Distribuição da disposição de feedback para aprimoramento de recomendação.



Fonte: autoria própria.

A partir da Figura 14, nota-se também que, com a amostra de leitores, foi compreendida uma alta taxa que afirma não utilizar sistemas de recomendação. Aqueles que afirmam utilizar, na verdade, descrevem sites que não empregam nenhuma solução complexa, mas apenas a base de avaliações e feedbacks dos próprios usuários, juntamente com ferramentas de pesquisa por filtros.

Figura 14 - Distribuição da utilização de sistemas de recomendação.

Fonte: autoria própria.

Com base nas informações reunidas durante o estudo preliminar, ficou evidente que os leitores buscam recomendações baseadas em suas preferências pessoais ou em tendências populares. Eles também desejam ter um sistema de recomendação que permita ajustar suas preferências, dar feedback sobre as recomendações recebidas, criar listas personalizadas de mangás e seguir ou interagir com outros leitores ou autores. Alguns leitores também demonstraram interesse em um sistema que utilize elementos de gamificação para tornar a experiência mais divertida e dinâmica.

4. Trabalhos Relacionados

Mesmo a área de recomendação de mangás sendo bastante específica, é possível encontrar alguns trabalhos as quais utilizam deste tipo de massa de dados para realização de estudos comparativos a fim de demonstrar a importância deste tipo de tecnologia para facilitar a busca por recomendações de qualidade.

Por exemplo, no trabalho de (Daniel Adrian, 2022), que desenvolveu um sistema de recomendação de mangás utilizando os algoritmos Naive Bayes e J48, que são justamente modelos de recomendação baseado em conteúdo. Com o objetivo de fornecer aos usuários recomendações personalizadas com base em suas preferências de gênero. No trabalho foi analisado os atributos dos mangás, como gênero, título, pontuação, popularidade, membros, favoritos e votos da pontuação e, com base nesses atributos, ele construiu modelos de recomendação utilizando os algoritmos Naive Bayes e J48. O Naive Bayes é um algoritmo de classificação baseado no Teorema de Bayes, ao qual utiliza de probabilidade condicional a fim de classificar os dados a partir de uma árvore de decisões com o objetivo de prever a classe de novos objetos baseados em seus atributos. Enquanto isso, o J48 é um algoritmo de árvore de decisão que busca criar um modelo para prever a variável alvo e realizar classificação de modo que consiga lidar com informações ausentes ou divergentes. Após a implementação dos modelos, Daniel Adrian realizou testes e avaliou o desempenho de cada algoritmo. Ele comparou a quantidade e a variedade de recomendações geradas por cada algoritmo, destacando o Naive Bayes, que produziu recomendações mais variadas, segundo ele, enquanto o J48 realizou recomendações mais precisas, só que em menor quantidade. No geral, o trabalho de Daniel Adrian concluiu a viabilidade de utilizar algoritmos de aprendizado de máquina para criar sistemas de recomendação de mangás, destacando a importância de considerar diferentes aspectos, como quantidade e variedade de recomendações ao avaliar o desempenho dos sistemas.

Uma outra abordagem realizada no mesmo tema é o trabalho de (Jill-Jenn Vie, 2017), que foi em busca de propor uma solução ao desafio do “início frio” em sistemas de recomendação de anime e mangás. A autora propôs o modelo BALSE (Blended Alternate Least Squares with Explanation), incorporando técnicas de aprendizado profundo. O BALSE é um sistema de recomendação híbrido que combina técnicas de filtragem colaborativa e baseada em conteúdo para melhorar a precisão das recomendações, especialmente em cenários de "início frio". Na explicação de Jill-Jenn Vie, o BALSE extrai informações de tags dos posters de mangás e animes usando uma rede neural convolucional, e em seguida, integra essas informações em um modelo de filtragem colaborativa, resultando em recomendações mais personalizadas e precisas. Essa abordagem híbrida permite ao BALSE superar as limitações dos métodos puramente baseados em filtragem colaborativa ou baseada em conteúdo, proporcionando recomendações mais diversificadas, melhorando bastante a qualidade das recomendações, em cenários de obras menos conhecidas.

Alguns outros trabalhos que, mesmo não sendo especificamente focados na recomendação de mangás, utilizam de técnicas e procedimentos semelhantes, aplicadas em diversas outras áreas, as quais se beneficiaram de sistemas que recomendam itens personalizados para cada usuário de maneira rápida e eficaz.

No trabalho de (Pedro Chamberlain Matos, 2021), o qual realizou um estudo de modelos de sistema de recomendações utilizando uma massa de dados de avaliações de filmes, mostra-se a importância dessas tecnologias de recomendação como solução para as grandes massas de informação existentes. Ele realiza uma visão detalhada sobre as diversas categorias de modelos de recomendação mais comuns. No trabalho, Pedro Chamberlain realiza um estudo comparativo de desempenho entre vários algoritmos, entre eles, alguns da categoria de filtragem colaborativa, sendo quatro variações do algoritmo de K-vizinhos mais próximos e o algoritmo SVD, e dois algoritmos de filtragem baseada em conteúdo, utilizando árvores de decisão e a técnica de Word2Vec. Nos processos realizados por Pedro Chamberlain, foi realizada uma análise exploratória da massa de dados obtida por ele de filmes, utilizando métricas para permitir a comparação entre os modelos, como o R-Score, soma das notas reais e erro de predição, além da sugestão de um modelo híbrido de recomendação utilizando SVD com refinamento por Word2Vec. Na conclusão, foi evidenciado como os modelos híbridos conseguem ter vantagem em relação aos algoritmos em sua forma básica, além de concluir que, para o cenário da massa de dados utilizada por ele, os algoritmos baseados em filtragem colaborativa tiveram um pouco mais de estabilidade e desempenho na hora da recomendação.

(Flávio de Holanda, 2017) também realiza um trabalho similar de desenvolvimento e comparação de métodos de recomendação. Em seu trabalho, é utilizado uma massa de dados já fornecida online pela GroupLens Research e, a partir disso, é realizada a implementação e validação de métodos de filtragem colaborativa SVD, item-item e user-user, baseada na métrica de desempenho RMSE. Em seu trabalho, é utilizada a biblioteca de Python chamada Suprise para instanciamento e treinamento dos modelos de filtragem colaborativa, sendo as validações feitas a partir da média de execução do RMSE para avaliar a taxa de aprendizado e o termo de regularização. Como resultado, foi observado que, entre as técnicas abordadas, o SVD teve um melhor desempenho. No entanto, foi discutido que uma possível solução para melhorar ainda mais o desempenho do modelo seria a utilização da variação SVD++, que é um modelo mais otimizado e comumente utilizado na Netflix Prize, uma competição na qual boa parte de seu trabalho foi baseado.

Já no trabalho de (Raimundo Nonato N., 2019), foi realizada a comparação entre medidas de similaridade no sistema de recomendação para a coleção ML100k de avaliações de filmes do GroupLens. As medidas utilizadas para realização de recomendação foram o Coeficiente de Correlação de Pearson e a Similaridade de Razão de Logaritmo da Verossimilhança. Durante seu trabalho, são abordadas as especificações de cada uma das medidas, assim como a utilização das métricas de desempenho MAE e RMSE. Durante o decorrer do trabalho, é explicada a importância da divisão da massa de teste e treinamento em 80/20, na qual 80% da massa de dados é para treinamento e 20% para testes e validação das predições. Por fim, o projeto mostra os resultados das recomendações realizadas por ambos os modelos e conclui que a medida LLR mostra maior acurácia em implementações de filtragem colaborativa baseada em usuário, enquanto o PCC fica bastante próximo em questão de desempenho quando utilizado em implementações de filtragem colaborativa baseada em item em cenário de maior número de avaliações.

5. Metodologia

Nesta seção, descreve-se a metodologia adotada no presente estudo com o objetivo de desenvolver um sistema de recomendação colaborativa de mangás, com base em dados coletados da internet, em uma comunidade de leitores, a fim de proporcionar recomendações personalizadas e relevantes para os usuários.

Todas as implementações de código realizadas neste trabalho serão feitas utilizando a linguagem de programação Python 3. Além da utilização de arquivos CSV (arquivos de valores separados por vírgula) para o armazenamento dos registros coletados com o intuito de serem utilizados nos modelos de filtragem colaborativa, também serão utilizadas bibliotecas Python de manipulação de objetos HTML, manipulação de massas de dados e funções para treinamento e geração das recomendações. Todas as bibliotecas Python serão introduzidas e será detalhado seu uso em seus tópicos específicos da etapa de desenvolvimento do projeto.

Como relatado no estudo sobre sistemas de recomendação de (F.O. Isinkaye. 2015), juntamente com os principais pontos descritos na Seção 2 deste trabalho, sobre o funcionamento e impacto que este tipo de tecnologia é capaz de proporcionar de forma benéfica à velocidade e qualidade de sugestões para os usuários de sistemas, foi elucidado que o processo para a implementação de um modelo de recomendação personalizada necessita respeitar certas etapas para reunir dados de qualidade que reflitam em resultados mais coerentes e precisos. Tal processo, como analisado na Seção 2, envolverá quatro etapas: a coleta inicial de dados, o tratamento dos dados brutos extraídos, a implementação dos modelos de recomendação e, por fim, a avaliação do sistema de recomendação com os resultados adquiridos, como ilustrado na Figura 15.

Figura 15 - Diagrama das quatro etapas do desenvolvimento de recomendações.

Fonte: autoria própria.

Como parte inicial do processo de desenvolvimento, é necessário reunir a massa de dados que será utilizada como base para que os algoritmos implementados sejam capazes de realizar a recomendação da maneira mais coerente e precisa possível. A qualidade e a precisão

de como essas informações refletem as entidades às quais estão contidas dentro do contexto de leitura de mangás, sendo elas, no geral, os leitores, as obras e as avaliações dos leitores diante das obras, caracterizam o relacionamento entre eles. Com este requisito em mente, no projeto será realizado o desenvolvimento de um rastreador web (webcrawler) que, a partir de uma fonte, sendo ela um site de comunidade de leitores de mangá que contém diversas informações relacionadas ao tema, coletará todos os dados que serão necessários para a implementação dos modelos de recomendação utilizando o processo descrito na Figura 16.

Figura 16 - Diagrama das etapas de processo dos rastreadores web.

Fonte: autoria própria.

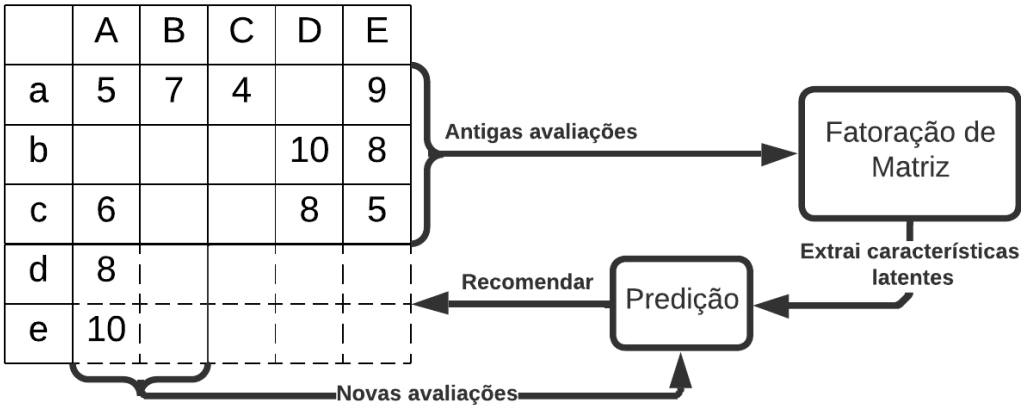
A fonte escolhida neste trabalho para a coleta dos dados é um fórum de comunidade em inglês, desenvolvido por um estadunidense e atualmente operado por uma empresa japonesa. Esse fórum conta com uma ampla gama de usuários diários, que comumente passam por mais de 10 mil acessos simultâneos e dispõe de um grande acervo de informações das obras de mangás, sendo atualizadas frequentemente com novos lançamentos a todo momento, e que não restringe as visualizações exigindo login. A partir da página principal de cada obra, a qual o fórum automaticamente realiza a identificação por um número inteiro incremental, iniciando do 1, é possível visualizar as informações retornadas pelo servidor do fórum dentro da sua estrutura HTML, como também fornece uma seção na qual é possível observar as avaliações que os usuários do fórum realizaram para a determinada obra, tanto o texto quanto a avaliação numérica de 0 a 10.

Portanto, devido às características dos dados que são fornecidos no fórum de comunidade dos leitores de mangá, na etapa de implementação do modelo de recomendação personalizada, como discutido anteriormente na Seção 2 de Referencial Teórico e embasado pelo trabalho descritivo de (F.O. Isinkaye. 2015), dos diversos tipos de abordagens para a implementação de modelos de sugestão, juntamente com a apresentação dos benefícios e desvantagens de cada um, serão implementados dois modelos de Filtragem Colaborativa. Os principais motivos da escolha são o benefício de se desempenhar bem neste tipo de cenário em que a esparsidade entre obras e usuários é bem grande, já que comumente os títulos mais populares possuem muitas avaliações, enquanto os menos populares são bem menos avaliados, desta forma os perfis de leitores poderão se complementar e aprimorar as sugestões de cada um. Além disso, também gerarão sugestões com tendências não apenas de manter sempre no preferencial do usuário, mas de permitir que outros gêneros e possíveis novos interesses surjam. E, além disso tudo, esta vertente de abordagem de modelo de recomendação é uma das mais utilizadas para este âmbito, segundo (F.O. Isinkaye. 2015), e também oferece uma implementação simples a partir do momento em que já possui a massa de dados tratada e pronta.

A partir desta premissa, foram selecionadas duas técnicas de recomendação por filtragem colaborativa para serem implementadas no projeto. O intuito foi selecionar uma alternativa de método de recomendação para as abordagens baseadas em modelo e em memória que possuem características que se beneficiam do cenário de massa de dados que está sendo coletado.

A primeira técnica implementada será a Singular Value Decomposition (SVD), uma abordagem que, com seu conceito matemática descrito na **Seção 2.2.3**, é amplamente reconhecida na filtragem colaborativa baseada em modelo por ser uma abordagem de álgebra linear e de fácil aplicação. Para a recomendação de mangás, seu processo é por meio da representação das relações e avaliações entre, usuários leitores e as obras. A partir desta estrutura estatística, o SVD aplica a decomposição de matriz com o intuito de preencher as relações inexistentes entre usuários e obras que ainda não foram avaliadas. Deste modo, o SVD é capaz de realizar a sugestão, predizendo as avaliações ainda não realizadas pelos usuários baseados nos padrões das relações, como mostrado na Figura 17.

Figura 17 - Diagrama da decomposição de matriz com SVD.



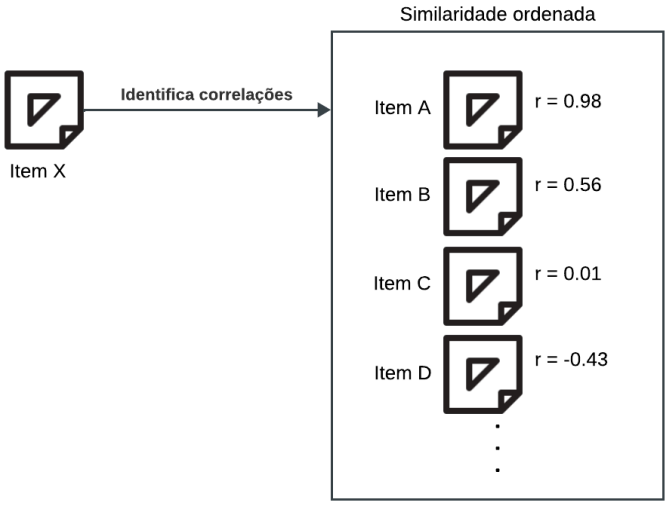
Fonte: autoria própria. (baseada em Ritik Gupta, 2023).

As vantagens, como defendido e analisado por (Flávio de Holanda, 2017), de aplicar o SVD vêm da sua eficácia em lidar com conjuntos de dados esparsos, onde as avaliações de usuários para itens são limitadas, o que é uma característica que pode auxiliar na recomendação de novos conteúdos que não possuem avaliações. Além disso, o SVD possui a capacidade de reduzir a dimensionalidade dos dados, ou seja, a técnica ao realizar a decomposição da matriz de relacionamentos, em seu processo, seleciona um subconjunto dos vetores singulares mais importantes para representar os dados, o que a torna eficiente computacionalmente, ocupa menos espaço, já que irá lidar com uma forma reduzida da massa de dados disponibilizada, como também é menos propensa a problemas de sobreajuste por não correr o risco de gerar ruídos nos dados na tentativa de reduzir em excesso, sempre capturando a maior parte da estrutura das informações. Em resumo, o SVD oferece uma maneira robusta de gerar recomendações personalizadas em cenários onde a esparsidade dos dados é um obstáculo e onde padrões precisam ser identificados para melhorar a precisão e coerência das recomendações.

Já a segunda implementação como alternativa no projeto é a Pearson's Correlation, uma abordagem de filtragem colaborativa baseada em memória que, como já visto na **Seção**

2.2.4, é uma medida estatística que funciona identificando a similaridade por correlação, neste cenário é entre os perfis de dois usuários, baseados em suas avaliações para as obras, considerando o quanto os dois são linearmente conectados, ou seja, a medida que as avaliações de um usuário aumentam em relação a uma obra, as avaliações do segundo usuário também aumentam de maneira proporcional, estabelecendo assim uma semelhança de perfis. Quanto maior a correlação de Pearson entre dois usuários, mais semelhantes são as suas preferências. Uma das características da medida de Pearson's Correlation, definida e estudada mais a fundo no trabalho de (Paranhos, R. et al, 2014), é que o seu valor pode variar também negativamente, o que significa que, para recomendações, a relação de similaridade entre dois usuários ou obras é dissimilar (especificamente em uma relação linear), ou seja, definida como uma correlação negativa ou zero, desta forma permite categorizar as sugestões resultantes mais profundamente, facilitando a interpretação de como as relações foram geradas pelo sistema, como visto na Figura 18.

Figura 18 - Demonstração da utilização do coeficiente de correlação de Pearson.



Fonte: autoria própria.

Diferentemente da utilização da medida de similaridade Cosseno, que utiliza o cálculo do cosseno do ângulo entre os vetores, a correlação de Pearson realiza uma média dos valores, ajustando-os, de maneira que não enfatize demais os valores não-zero, resultando na captura de relações onde as tendências de avaliação são diferentes. Tal característica funciona bem em cenários nos quais os dados não são tão densos e permite predizer possíveis novos interesses ao leitor. A utilização da medida de correlação de Pearson funciona também em cenários em que os usuários, pessoalmente, utilizam proporções numéricas de avaliações diferentes. A técnica realiza uma subtração da média de cada variável com o intuito de regular escalas de notas para que facilite a etapa de identificação de perfis semelhantes, já que realiza isso baseado em proporcionalidade.

Portanto, a medida de correlação de Pearson é uma alternativa para a implementação da recomendação de obras de mangá por ser uma técnica facilmente interpretável e de fácil aplicação, como também, devido a sua natureza matemática, é uma boa alternativa para

reduzir custo computacional, além de que é um modelo capaz de inferir as preferências do usuário com base em seu comportamento, no contexto deste trabalho seria com base no seu histórico de interações entre obras, identificando similaridades que não são tão óbvias, enriquecendo mais os resultados da sua predição.

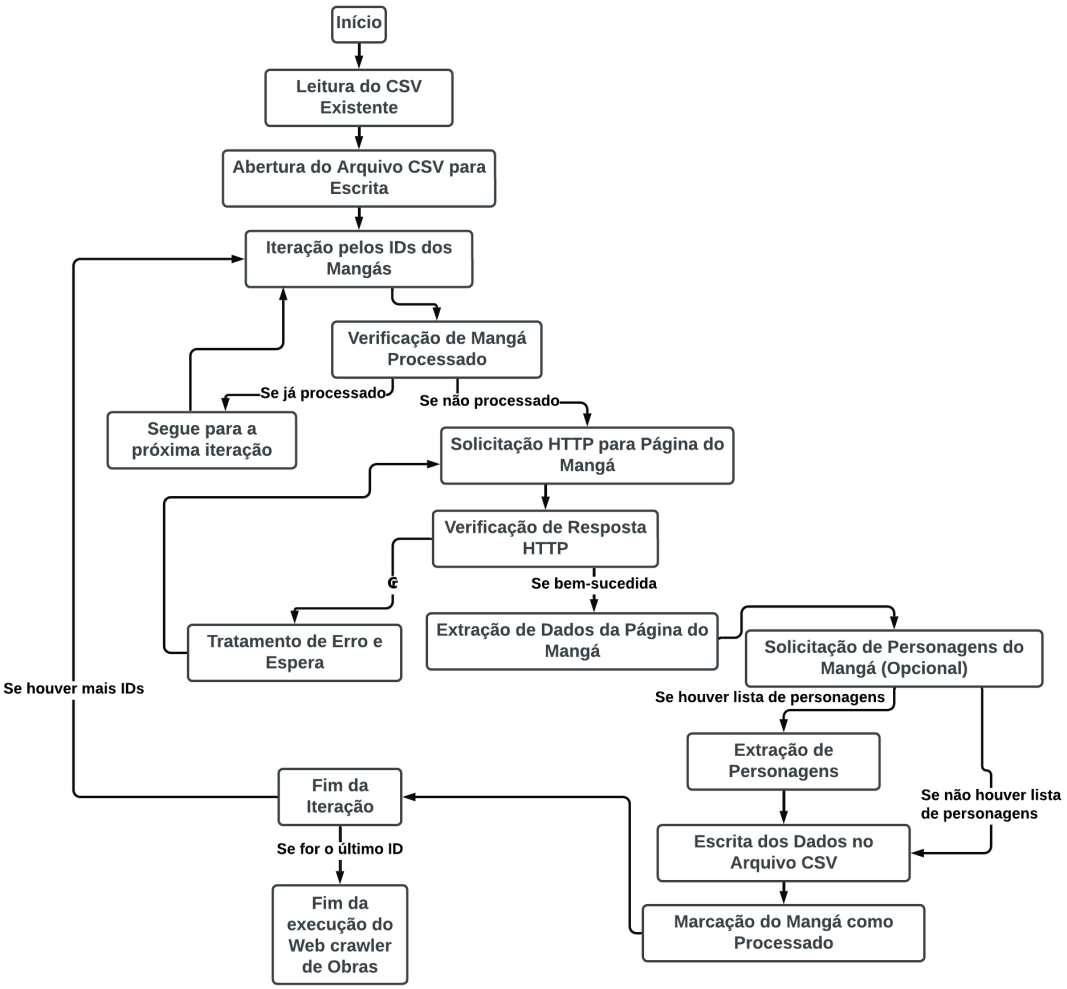
Seguindo estes pontos, os capítulos a seguir irão descrever o processo realizado para a obtenção da massa de dados a partir do rastreador web, a implementação e a execução da recomendação feitas com os algoritmos de recomendação com medida de correlação de Pearson e com a técnica SVD.

6. Extração dos dados

Como base para os modelos de recomendação, a criação do rastreador para a extração das informações essenciais foi realizada com dados disponibilizados no fórum de comunidade de mangás, a fim de reunir dados que possibilitem a realização de sugestões de novas obras para leitura a partir de um cenário real advindo da interação entre obras e usuários.

Para a realização da extração de maneira simples e que fosse possível interromper a execução do web crawler sem perder informações importantes e prejudicar os dados já coletados anteriormente, o código foi feito a partir da separação de responsabilidades, criando dois rastreadores diferentes. O primeiro, como ilustrado na Figura 19, é focado em extrair apenas as informações relacionadas às obras existentes, realiza uma iteração entre todas as obras registradas na fonte e gera um arquivo CSV. Ele insere os registros com base no número identificador incremental que o fórum disponibiliza e utiliza na navegação dos conteúdos.

Figura 19 - Fluxograma de execução das etapas do rastreador web de obras.

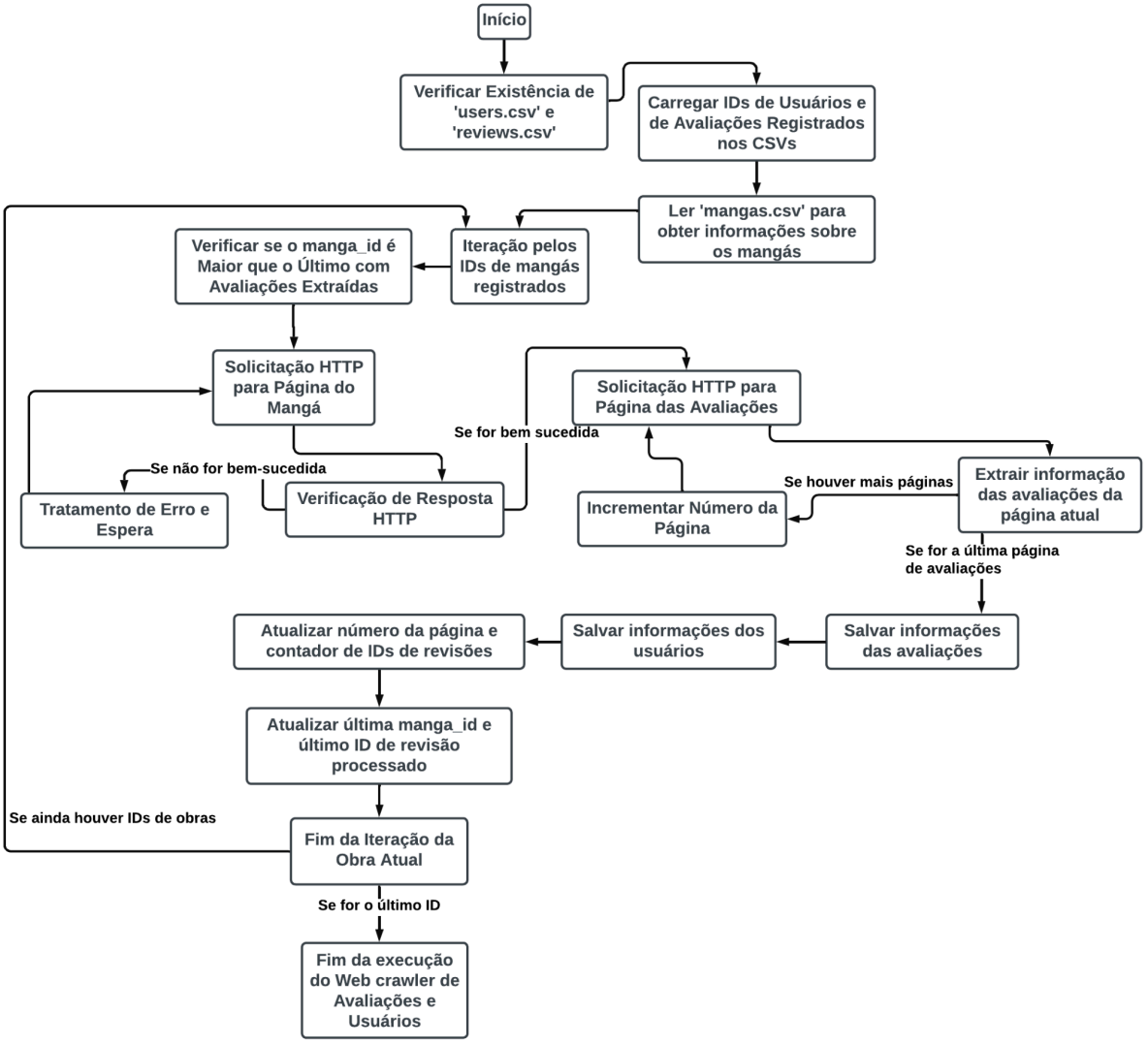


Fonte: autoria própria.

O segundo rastreador, que ficou exclusivamente responsável por coletar as informações das avaliações que as obras possuem, também é encarregado de coletar e registrar os usuários que realizaram tais avaliações. O código possui uma dependência

referente aos dados extraídos do primeiro rastreador, de obras, para garantir o relacionamento entre registros de ambas as tabelas. O rastreador coleta dados de críticas feitas pelos usuários. como também ficou responsável paralelamente por gerar um CSV básico de registro de todos os usuários, como observado na Figura 20, atribuindo a cada um um identificador incremental, como prevenção para evitar que um usuário acabe sendo registrado com identificadores diferentes dentro da massa de dados e facilitando caso surja a necessidade de cruzamento de informações entre avaliações, obras e usuários.

Figura 20 - Fluxograma de execução das etapas do rastreador web usuários e avaliações.



Fonte: autoria própria.

6.1 Extração das informações das obras

Antes de tomar como parte a implementação do rastreador web de obras, foi realizada uma análise e levantamento de quais atributos dos mangás seriam relevantes para extrair a partir do fórum de mangás. Na Tabela 1 é possível analisar o glossário destas informações,

como também o tipo de cada característica, para tomar como base na formatação e extração com o objetivo de facilitar a futura manipulação e geração das predições.

Tabela 1 - Glossário do mapeamento dos atributos relevantes dos mangás.

Atributo	Descrição	Tipo
id	Identificador único.	Inteiro
title	Título da obra.	String
type	Tipo da obra (por exemplo, Manga neste caso).	String
score_value	Pontuação média da obra.	Float
score_count	Número de avaliações que contribuíram para a pontuação média.	Inteiro
ranked	Posição no ranking.	String
popularity	Popularidade da obra.	String
volumes	Número total de volumes.	Inteiro
chapters	Número total de capítulos.	Inteiro
status	Status da obra(finished,publishing ou on hiatus)	String
genres	Gêneros aos quais a obra pertence.	Lista de Strings
themes	Temas abordados na obra.	Lista de Strings
demographic	Público-alvo da obra	String

serialization	Revista ou plataforma onde a obra foi originalmente publicada.	String
published	Período de publicação da obra.	String
authors	Autores da obra.	Lista de Strings
synopsis	Resumo ou sinopse da obra.	String
image_src	URL da imagem da obra.	URL (String)
background	Contexto ou informações adicionais sobre a obra.	String
alt_titles	Outros títulos pelos quais a obra é conhecida.	Lista de Strings
characters	Lista de personagens presentes na obra.	Lista de Strings

Fonte: autoria própria.

Já no início do desenvolvimento, como primeira preocupação, foi preciso definir como seria realizada a persistência dos dados extraídos pelo rastreador. De maneira simples, no início da execução do código, foi criada uma verificação para checar a existência do arquivo CSV responsável por armazenar as informações coletadas. Além disso, foi implementado um mecanismo para garantir que o rastreador iniciasse a busca a partir do último identificador único salvo, a fim de evitar múltiplos registros da mesma obra e impedir que o rastreador reiniciasse a extração do zero, como mostrado na Figura 21.

Após garantir a verificação da existência de obras já coletadas e salvas dentro do arquivo "mangas.csv", assim como na Figura 21, o código abre o arquivo em modo de leitura e escrita, criando-o caso ainda não exista no sistema, para iniciar o processo de extração. Em seguida, cria uma instância do objeto escritor e realiza uma verificação caso não haja nenhum registro, definindo assim as colunas que o arquivo CSV possui, identificando todos os atributos que podem ser atribuídos como identidade de cada obra.

Figura 21 - Pseudocódigo da inicialização do arquivo CSV de obras.

```
//Definindo variáveis
#url_base = "https://MY_FORUM"
#url_base_manga = #url_base + "/manga/"
#arquivo_manga = 'mangas.csv'
#ids_processados = inicializa como uma lista vazia

//Verificando se o arquivo CSV existe
se o arquivo #arquivo_manga existe:
    abrir #arquivo_manga para leitura:
        #leitor = criar leitor de CSV
        para cada linha de #leitor:
            adicionar o ID da linha a #ids_processados

abrir #arquivo_manga para adição:
    //Definindo os campos do cabeçalho
    #atributos_manga = ['id', 'title', 'type', 'score_value', 'score_count', 'ranked',
     'popularity', 'volumes', 'chapters', 'status',  'genres', 'themes', 'demographic',
      'serialization', 'published', 'authors', 'synopsis', 'image_src', 'background',
      'alt_titles','characters']

    #escritor_csv = criar escritor de CSV utilizando aspas duplas

    se o tamanho do arquivo #arquivo_manga for 0:
        #escritor_csv escrever cabeçalho #atributos_manga em #arquivo_manga
```

Fonte: autoria própria.

Após as etapas de verificação e preparação do arquivo de armazenamento, o rastreador entra na etapa de seleção e extração das obras no fórum de mangás. Ele itera dentro do intervalo de 1 a 165248, que representa o intervalo de identificadores do primeiro ao último mangá introduzido no fórum até o momento da execução do rastreador web (Figura 22). Durante essa iteração, o rastreador verifica se o identificador em questão já está presente na listagem de identificadores já processados. Se a obra já tiver sido processada, o iterador a ignora e passa para a próxima. Caso o identificador não exista dentro do arquivo, o rastreador monta a URL e realiza uma requisição GET, por meio do caminho /manga/{id}, para obter a página da obra no fórum.

Figura 22 - Pseudocódigo do rastreador de obras solicitando página web.

```
// Iterando sobre os IDs de mangá
para cada #id_manga no intervalo de 1 até 165247:
    se #id_manga está em #ids_processados:
        imprimir("mangá já existe na base.")
        pular para o próximo #manga_id

    // Solicitando manga para a fonte
    #url_manga = #url_base + #id_manga
    #pagina_manga = solicitar_pagina_web(#url_manga)
```

Fonte: autoria própria.

Um dos obstáculos encontrados durante o processo de extração das obras é que após uma certa quantidade ininterrupta de requisições para a fonte, a execução começa a ser detectada e bloqueada por cerca de 5 minutos. Para lidar com esse cenário, foi implementado um bloco de verificação relacionado ao status da requisição web, mostrado na Figura 23. Se o retorno da requisição estiver com código de status 429, que indica que o usuário realizou muitas requisições em um curto período de tempo, ou status 405, que indica que o usuário não está autorizado a utilizar o método de requisição, o processo do rastreador é pausado por 3 minutos. Durante esse tempo de espera, o rastreador tenta novamente realizar a requisição para a obra de mesmo identificador que não foi obtida da fonte. Esse aguardo é repetido a cada 3 minutos até que a requisição tenha um retorno com código de status 200, que indica que a requisição foi bem-sucedida. Após isso, o restante do fluxo da implementação é retomado.

Figura 23 - Pseudocódigo do rastreador de obras requisitando a fonte.

```
// Definindo a função para fazer solicitações ao website
função solicitar_pagina_web(#url):
    enquanto verdadeiro:
        #resposta_web = enviar solicitação GET para #url

        // Verificando o código de status da resposta
        se o código de status da resposta for 200:
            quebrar o loop
        senão, se o código de status da resposta for 429 ou 405:
            imprimir("A solicitação GET falhou com o código de status 429 ou 405")
            esperar 180 segundos
        senão:
            imprimir("A solicitação GET falhou com o código de status diferente")
            quebrar o loop
    retornar #resposta_web
```

Fonte: autoria própria.

Após a requisição bem-sucedida e com a página da obra em mãos, o rastreador utiliza a biblioteca BeautifulSoup em Python para facilitar a manipulação dos elementos HTML retornados. Com o BeautifulSoup, é possível pesquisar, selecionar e extrair elementos individuais da resposta da requisição.

Em relação aos atributos extraídos da obra, há um que não pode ser obtido apenas na requisição da página principal. A lista dos nomes dos personagens que fazem parte da história da obra possui uma página exclusiva. Para obtê-los, foi necessário implementar o método da Figura 24 que, a partir da chamada de método mostrado na Figura 23, executa a requisição para a página de personagens da obra específica, utilizando o BeautifulSoup para selecionar elementos HTML e retorná-los.

Figura 24 - Pseudocódigo do rastreador de obras que obtém a lista de personagens.

```
// Definindo a função para solicitar personagens de um mangá
função solicitar_personagens_manga(#verificar_lista_personagens, #id_manga, #pagina_manga):
    #pagina_personagens = inicializa vazio

    se #verificar_lista_personagens for verdadeiro:
        #caminho_personagens = obter caminho para personagens de #pagina_manga
        #pagina_personagens = solicitar_pagina_web(#url_base + #caminho_personagens)

    se #pagina_personagens não for vazio:
        #elementos_personagens = extrai lista de elementos de #pagina_personagens
        retornar #elementos_personagens

    retornar uma lista vazia
```

Fonte: autoria própria.

Nas próximas etapas do rastreador, é realizada a identificação e extração de todos os elementos que são caracterizados como atributos da obra de mangá solicitada. Para isso, foi implementado um código de extração de elementos HTML de forma genérica o suficiente para ser aplicado em todas as obras que precisam ser extraídas. Esse código trata as cadeias de caracteres para corrigir problemas decorrentes de erros de codificação ou problemas de formatação, como espaçamento, quebras de linha, novas linhas, parágrafos ou valores vazios.

Após a extração de todos os atributos da fonte, os dados são reunidos e um novo registro de mangá é escrito no arquivo "mangas.csv", finalizando a etapa após a adição do registro na lista de identificadores processados. Então, o processo itera para a próxima obra e o mesmo procedimento é repetido do início até que todos os identificadores necessários sejam requisitados e processados.

Devido à restrição da fonte de bloquear o acesso ao fórum por alguns minutos após uma certa quantidade de requisições em sequência e os processamentos necessários para extração e registro no arquivo CSV, foram necessárias de 6 a 7 semanas para obter a massa de dados completa de obras, consciente de que o rastreador foi executado em uma máquina pessoal durante o intervalo de horário das 09:00 AM às 10:00 PM.

Ao finalizar a extração, foi realizado um pequeno tratamento na massa de dados para verificar se houve algum problema relacionado à duplicidade de obras coletadas. Como os identificadores são os mesmos utilizados pela fonte para identificar unicamente todas as obras, a biblioteca pandas do Python foi importada para transformar o arquivo CSV "mangas.csv" em um objeto dataset e realizar a verificação. Após isso, foram executados métodos do pandas para excluir registros duplicados com base na coluna "id" e também foram removidos registros vazios que eventualmente poderiam ter sido inseridos na massa de dados. Como resultado, foi identificado que não houve nenhum registro salvo em “mangas.csv” com duplicatas, como também não foi encontrado nenhum registro vazio.

Como resultado, exemplificado na Figura 25, da extração e do procedimento de limpeza, a massa de dados bruta obtida registrou um total de **78.927** obras, com um arquivo contendo quase **47 MB** de conteúdo.

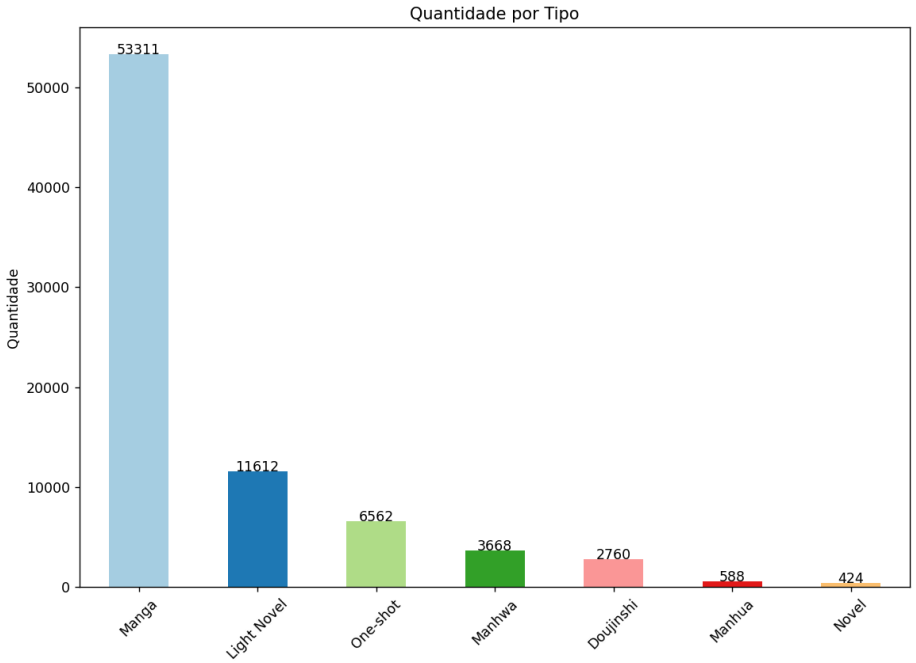
Figura 25 - Demonstrativo de parte dos dados coletados de mangás.

	A	B	C	D	E	F	G	H	I	J	
1	id	title	type	score_value	score_count	ranked	popularity	volumes	chapters	status	genres
2	1	Monster	Manga	9.15	92945	#5	#29	18	162	Finished	['Award Winning', 'Drama', 'Myster
3	2	Berserk	Manga	9.47	328161	#1	#1	Unknown	Unknown	Publishing	['Action', 'Adventure', 'Award Winni
4	3	20th Century Boys	Manga	8.95	84485	#13	#28	22	249	Finished	['Award Winning', 'Drama', 'Myster
5	4	Yokohama Kaidashi Kike	Manga	8.68	18012	#54	#205	14	142	Finished	['Award Winning', 'Drama', 'Sci-Fi',
6	7	Hajime no IppoHajime no	Manga	8.72	35049	#45	#169	Unknown	Unknown	Publishing	['Award Winning', 'Sports']
7	8	Full Moon wo Sagashite	Manga	8.04	18529	#667	#445	7	35	Finished	['Comedy', 'Drama', 'Fantasy', 'Ror
8	9	Tsubasa: RESERVoir Cl	Manga	8.3	36551	#291	#187	28	233	Finished	['Action', 'Adventure', 'Drama', 'Fan
9	10	xxxHOLiC	Manga	8.37	31813	#232	#174	19	213	Finished	['Comedy', 'Drama', 'Mystery', 'Sup
10	11	Naruto	Manga	8.07	267171	#612	#11	72	700	Finished	['Action', 'Adventure', 'Fantasy']
11	12	Bleach	Manga	7.85	229876	#1117	#13	74	705	Finished	['Action', 'Adventure', 'Award Winni
12	13	One Piece	Manga	9.22	363983	#4	#3	Unknown	Unknown	Publishing	['Action', 'Adventure', 'Fantasy']
13	14	RaveRave Master	Manga	7.85	23202	#1118	#313	35	298	Finished	['Adventure', 'Comedy', 'Fantasy']
14	15	Mahou Sensei Negima!N	Manga	7.93	35202	#895	#200	38	355	Finished	['Action', 'Adventure', 'Comedy', 'Fa
15	16	Love Hina	Manga	7.79	36204	#1311	#219	14	120	Finished	['Award Winning', 'Comedy', 'Roma
16	17	Kareshi Kanojo no Jijou	Manga	8.16	14018	#470	#435	21	108	Finished	['Comedy', 'Drama', 'Romance']
17	18	Kodomo no OmochaKod	Manga	8.3	10205	#292	#794	10	54	Finished	['Award Winning', 'Comedy', 'Dram
18	19	GetBackers	Manga	7.61	5497	#2165	#1283	39	344	Finished	['Action', 'Comedy', 'Drama', 'Myst
19	20	Hikaru no Go	Manga	8.11	22144	#551	#356	23	198	Finished	['Award Winning', 'Comedy', 'Dram
20	21	Death Note	Manga	8.7	221887	#49	#12	12	108	Finished	['Supernatural', 'Suspense']
21	22	Rurouni Kenshin: Meiji	Manga	8.56	46762	#103	#137	28	259	Finished	['Action', 'Drama']
22	23	Ranma ½Ranma 1/2	Manga	7.99	24407	#758	#312	38	407	Finished	['Action', 'Comedy', 'Romance', 'Ec
23	24	D.Gray-man	Manga	8.28	62328	#325	#63	Unknown	Unknown	Publishing	['Action', 'Adventure']
24	25	Fullmetal Alchemist	Manga	9.03	154540	#8	#20	27	116	Finished	['Action', 'Adventure', 'Award Winni
25	26	Hunter x Hunter	Manga	8.73	120755	#42	#24	Unknown	Unknown	Publishing	['Action', 'Adventure', 'Fantasy']
26	27	X	Manga	8.1	8264	#575	#701	18	158	On Hiatus	['Action', 'Drama', 'Fantasy', 'Super
27	28	Nana	Manga	8.79	41727	#36	#79	21	84	On Hiatus	['Award Winning', 'Drama', 'Roman
28	29	Paradise Kiss	Manga	8.29	20336	#311	#308	5	48	Finished	['Drama', 'Romance']
29	30	Ouran Koukou Host Clu	Manga	8.5	59018	#138	#91	18	87	Finished	['Comedy', 'Drama', 'Romance']
30	31	Lovely★ComplexLove★	Manga	8.33	27178	#268	#247	17	68	Finished	['Award Winning', 'Comedy', 'Dram
31	32	666 SatanO-Parts Hunte	Manga	7.44	16478	#3333	#478	19	78	Finished	['Action', 'Adventure', 'Comedy', 'D
32	33	Pita-Ten	Manga	7.38	4061	#3874	#2404	8	47	Finished	['Comedy', 'Drama', 'Fantasy', 'Ror

Fonte: autoria própria.

Na Figura 26 também é possível analisar a distribuição da quantidade por tipos de obras contidas na massa de dados do fórum, com a grande maioria pertencendo realmente à categoria de mangás. Porém, é possível ainda assim obras de Light Novel, que comumente são livros com somente texto que derivam posteriormente os mangás, além de obras de apenas uma publicação (One-shots), obras independentes (Doujinshi) e quadrinhos coreanos e chineses, manhwa e manhua respectivamente.

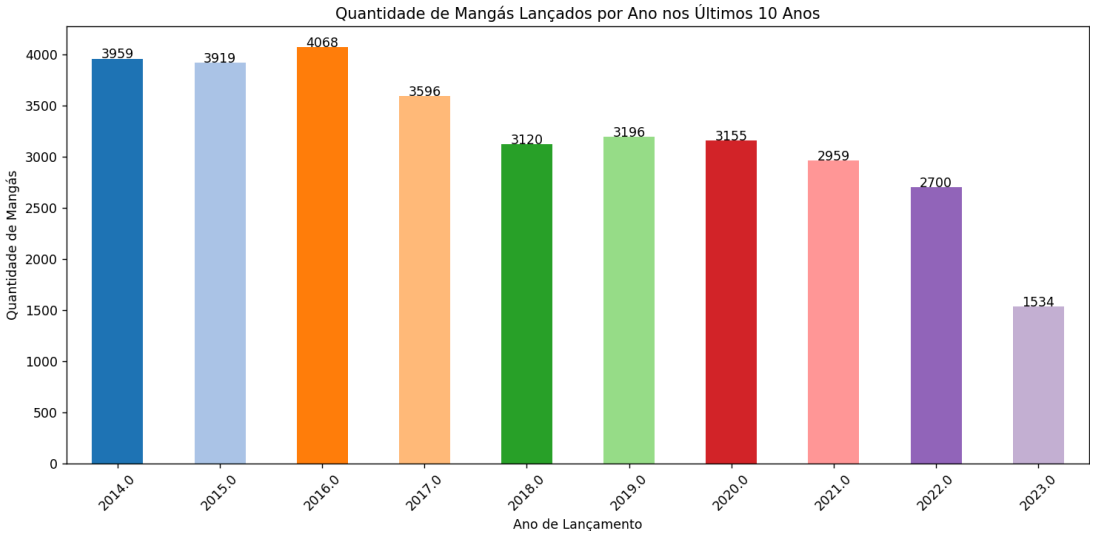
Figura 26 - Gráfico ilustrando a quantidade de registros de obra por tipo.



Fonte: autoria própria.

Outra análise interessante que é possível obter a partir do conteúdo extraído nas obras é a estatística da quantidade de mangás lançados por ano nos últimos 10 anos, ilustrado na Figura 27, a qual mostra uma média de 3000 títulos novos por ano.

Figura 27 - Gráfico ilustrando a quantidade de lançamentos por ano na massa de obras.



Fonte: autoria própria.

6.2 Extração das informações das avaliações

Como aplicado para as obras na Tabela 1, também foi realizado um mapeamento dos atributos relevantes referentes às avaliações dos usuários, tais características foram formatadas e utilizadas como base para a etapa de extração dos elementos na página de retorno HTML, na Tabela 2 é possível identificá-las, como também foi elencado suas descrições e tipos.

Tabela 2 - Glossário do mapeamento dos atributos relevantes das avaliações.

Atributo	Descrição	Tipo
review_id	Identificador único para cada revisão no conjunto de dados.	Inteiro
user_id	Identificador único do usuário que fez a revisão.	Inteiro
manga_id	Identificador único do manga que está sendo revisado.	Inteiro
rating	Classificação dada pelo usuário ao manga.	Inteiro
date	Data em que a revisão foi feita.	Data (Month Day, year)

review_text	Texto da revisão, onde o usuário descreve sua opinião sobre o manga.	String
total_reactions	Total de reações recebidas pela revisão.	Inteiro
nice	Quantidade de reações "Nice".	Inteiro
loveit	Quantidade de reações "Loveit".	Inteiro
funny	Quantidade de reações "Funny".	Inteiro
confusing	Quantidade de reações "Confusing".	Inteiro
informative	Quantidade de reações "Informative".	Inteiro
well-written	Quantidade de reações "Well-Written".	Inteiro
creative	Quantidade de reações "Creative".	Inteiro

Fonte: autoria própria.

Em relação à implementação, o rastreador web de avaliações é bem semelhante ao de obras, o seu intuito é, a partir do arquivo CSV “mangas.csv”, conseguir utilizar os mangás de identificadores já existentes na massa de dados para realizar uma requisição especificamente para a aba exclusiva de avaliações de usuários de cada obra. A peculiaridade desta extração é que cada mangá possui *n* números de avaliações que são divididas por *m* páginas, cada página mostra apenas um máximo de 20 avaliações por vez, ou seja, uma obra que possui 78 avaliações registradas para ela possui 4 páginas para poder extrair todas as avaliações. Então, foi preciso elaborar o rastreador de modo que checasse todas as páginas de avaliações de todas as obras.

O início da implementação do rastreador é bem similar ao do rastreador de obras, é realizado a verificação da existência do arquivo de armazenamento “reviews.csv” e verificado os registros que já estão salvos para evitar que o rastreador comece a coletar avaliações que já foram requisitadas como na Figura 21. Como experiência já do primeiro rastreador implementado, a etapa de pular para as avaliações já registradas foi feita de uma maneira um pouco diferente. Em vez de verificar se todos os registros até então estão realmente salvos, o código identifica e inicia a iteração a partir do último registro no CSV e obtém o identificador da avaliação e o da obra, já que as avaliações estão fortemente interligadas com as obras. Deste modo, a requisição poderá ser executada diretamente de onde foi pausado.

A partir disto, o rastreador realiza as etapas de iteração para cada identificador de obra no arquivo “mangas.csv” e monta a URL de requisição GET para a primeira página de avaliações e, de forma incremental, vai realizando requisições para a página n+1 até que a

fonte retorne uma mensagem de Não Encontrado, com código de status de requisição 404, como visto na Figura 20 do diagrama de fluxo do rastreador web.

Uma característica interessante deste rastreador de avaliações é que, mesmo sendo necessário muitas mais requisições para conseguir coletar todas as avaliações de cada obra, as operações HTTP entre páginas de avaliações não sofrem com problema de bloqueio por excesso de comunicação em pouco período de tempo. Assim, ambos os rastreadores de obras e avaliações possuem a tendência de ter uma complexidade de tempo de extração na mesma ordem de grandeza, já que as restrições de quantidade de requisições serão aplicadas apenas entre o fluxo de uma obra para a outra, independente se for para obter informações de mangás ou obter as avaliações do mangá.

Portanto, como as restrições de extração são similares, o tempo de execução do rastreador de avaliações acaba tendendo a ter a mesma complexidade de tempo para ser executado. Para formar a massa de dados de avaliações completa, teve um tempo de 6 a 7 semanas, considerando que o rastreador foi executado em uma máquina pessoal e que, no geral, no intervalo de horário das 09:00 AM às 10:00 PM.

Ao finalizar a extração, foi realizado um pequeno tratamento na massa de dados para verificar, como foi executado na massa de dados de mangás, se houve algum problema relacionado à duplicidade de avaliações coletadas. Como os identificadores utilizados, neste caso, foram gerados de maneira incremental para este projeto, em vez de fazer a verificação baseada na identificação numérica dada, foi feita a partir do conjunto de atributos **['manga_id', 'user_id', 'review_text']** a qual cada avaliação possui. Com isso, a partir da importação da biblioteca pandas, do Python, foi transformado o arquivo CSV “reviews.csv” em um objeto dataset para realizar a verificação. Como resultado, a massa de dados também não apresentou nenhum registro duplicado e nenhum registro vazio.

Com isto, a massa de dados bruta resultante possui **51.043 avaliações** de usuários, com um arquivo contendo quase **104MB** de conteúdo (Figura 28).

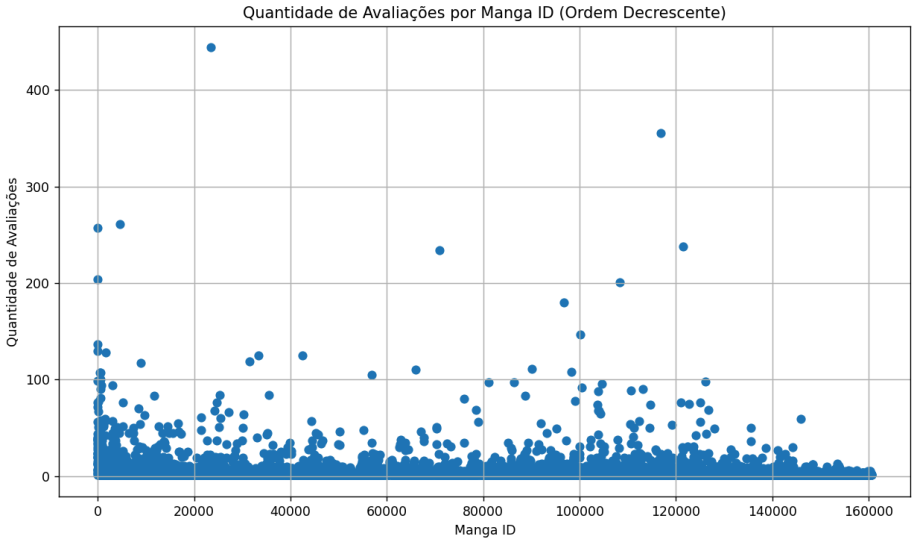
Figura 28 - Demonstrativo de parte dos dados coletados de avaliações.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	review_id	user_id	manga_id	rating	date	review_text	total_reactions	nice	loveit	funny	confusing	informative	well-written	creative
2	1	1	1	10	Jan 24, 2009	I rarely give t	482	462	5	3	4	0	8	0
3	2	2	1	9	Mar 22, 2012	"Look at me!	181	172	1	3	4	0	1	0
4	3	3	1	10	Jun 8, 2015	I loved this M	139	132	0	2	4	1	0	0
5	4	4	1	10	Mar 26, 2015	What makes	127	126	0	0	1	0	0	0
6	5	5	1	4	Dec 21, 2010	Not an overra	178	123	2	21	30	0	2	0
7	6	6	1	6	Sep 4, 2009	This manga i	48	39	1	6	0	0	2	0
8	7	7	1	9	Dec 25, 2007	The Viz Sign	55	55	0	0	0	0	0	0
9	8	8	1	2	Jan 18, 2020	Monster - the	39	20	5	11	3	0	0	0
10	9	9	1	10	Dec 14, 2010	Let me start	34	33	0	0	1	0	0	0
11	10	10	1	5	Aug 30, 2016	It kind of offe	26	22	2	1	1	0	0	0
12	11	11	1	4	Apr 15, 2021	Monster fails	18	10	3	4	1	0	0	0
13	12	12	1	2	May 22, 2023	I've read and	42	9	2	16	14	0	1	0
14	13	13	1	5	Apr 16, 2022	It is hard to r	8	1	1	1	1	4	0	0
15	14	14	1	4	Jul 30, 2022	I found Mons	11	5	3	3	0	0	0	0
16	15	15	1	9	Jan 9, 2013	Finally read M	16	14	0	0	1	0	1	0
17	16	16	1	9	Apr 26, 2012	REALLY fasc	19	18	0	0	1	0	0	0
18	17	17	1	9	Oct 22, 2008	Oh, MONSTE	17	16	0	0	1	0	0	0
19	18	18	1	10	Dec 28, 2013	Look at me, I	17	16	0	0	1	0	0	0
20	19	19	1	8	Jun 29, 2014	Although I ga	12	11	0	0	1	0	0	0
21	20	20	1	10	Apr 20, 2020	Do we all in I	11	10	0	0	1	0	0	0
22	21	21	1	10	May 24, 2016	There is only	10	9	0	0	1	0	0	0
23	22	22	1	8	Jan 30, 2011	Before Death	6	6	0	0	0	0	0	0
24	23	23	1	10	Oct 3, 2021	Despite most	6	6	0	0	0	0	0	0
25	24	24	1	7	Nov 22, 2015	I've heard a l	5	5	0	0	0	0	0	0
26	25	25	1	10	Dec 2, 2020	This is going	5	5	0	0	0	0	0	0
27	26	26	1	10	Aug 3, 2021	warning: spoi	5	5	0	0	0	0	0	0
28	27	27	1	10	May 22, 2020	This is a grea	5	5	0	0	0	0	0	0
29	28	28	1	10	Sep 18, 2020	Frighteningly	7	6	0	0	1	0	0	0
30	29	29	1	10	Mar 31, 2022	There's this r	1	0	0	0	0	0	1	0
31	30	30	1	9	Jun 29, 2016	It is importan	4	4	0	0	0	0	0	0

Fonte: autoria própria.

Ao analisar os dados extraídos de avaliações é possível compreender alguns pontos que tem relação direta no comportamento e precisão dos modelos de recomendação. Um destes pontos é a distribuição da quantidade de avaliações por obra registrada, como visto na Figura 29, a maior parte da massa de críticas disponível está abaixo das 30 avaliações por mangá, enquanto os títulos mais populares estão agrupados primariamente entre 50 e 150 avaliações. Também é possível observar que existem alguns outliers que se destacam, chegando a quantidades de 250 a 400 avaliações sozinhos.

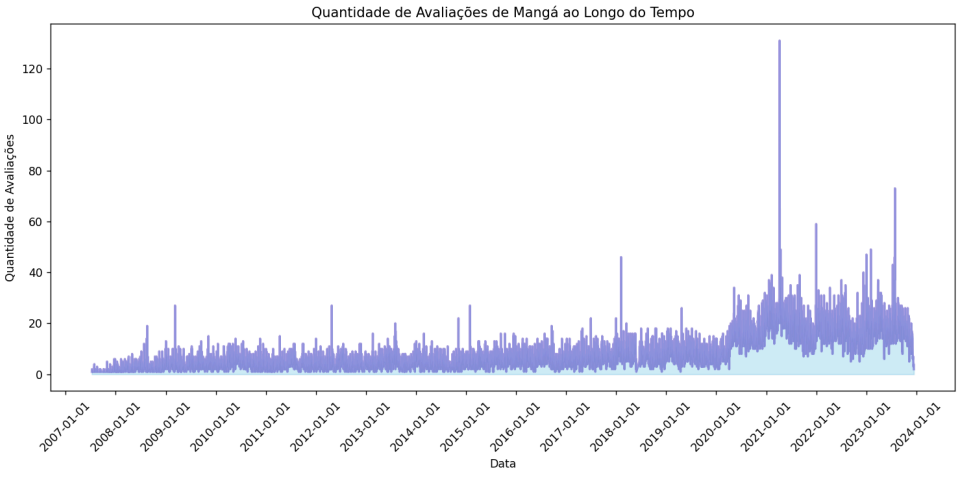
Figura 29 - Gráfico de dispersão de avaliações por ID de mangá.



Fonte: autoria própria.

Como também é possível entender o comportamento da quantidade de avaliações feitas durante os anos (Figura 30) e que mostra o crescimento recente, principalmente em específicos períodos as quais mostram picos enormes de avaliações criadas, o que evidencia o surgimento de obras que momentaneamente foram extremamente populares e que mobilizou boa parte da comunidade a interagir entre si.

Figura 30 - Gráfico de distribuição temporal de avaliações realizadas.



Fonte: autoria própria.

6.3 Extração das informações dos usuários

Por ser uma aplicação simples, de garantia à integridade entre relações realizadas pelas avaliações e obras, o glossário mapeado para a massa de dados de usuários se deu pela seguinte construção representada na Tabela 3.

Tabela 3 - Glossário do mapeamento dos atributos relevantes dos usuários.

Atributo	Descrição	Tipo
username	Nome de usuário que é utilizado como identidade única no fórum de mangás.	String
user_id	Identificador incremental único do usuário, utilizado para facilitar indexação.	Inteiro

Fonte: autoria própria.

Para a extração, também é realizada a verificação, no início da execução do código, de quais usuários já existem dentro do arquivo CSV "users.csv". Sempre que é identificado que um usuário ainda não foi adicionado, é realizada a obtenção do seu nome de usuário e, a partir de um mecanismo de identificador incremental, é realizada a sua inserção.

A extração dos usuários foi implementada de maneira bem simples e quem é responsável por isso é o rastreador de avaliações: enquanto está sendo realizada a obtenção das avaliações, o código também mantém salvo o usuário que realizou a avaliação (Figura 31), escrevendo-o no arquivo de armazenamento.

Figura 31 - Pseudocódigo do rastreador de avaliações, salvamento de usuário.

```
// Abrindo o arquivo CSV para adicionar novos usuários
#arquivo_usuarios = 'users.csv'
abrir #arquivo_usuarios para adição:
    #atributos_usuario = ["user_id", "username"]

    #escritor_csv = criar escritor de CSV

    se o tamanho do arquivo #arquivo_usuarios for 0:
        #escritor_csv escreve cabeçalho #atributos_usuario
    para cada #info_usuario em #usuarios_extraidos:
        #escritor_csv escreve linha com as informações do usuário
```

Fonte: autoria própria.

O fórum de mangás utiliza o próprio nome de usuário como identificador único para cada leitor, mas para evitar repetições muito grandes, foi optado por gerar estes identificadores numéricos. Como a complexidade de tempo de extração é compartilhada para avaliações e usuários, o tempo que durou em sua execução foi o mesmo, como discutido na Seção 5.2.

Para a etapa de tratamento da massa de dados resultante, como o identificador dos usuários na fonte é os próprios nomes de usuário, também foi realizada a varredura para

identificar registros duplicados e vazios dentro de “users.csv”. Porém, como resultado, também não foi identificada nenhuma correção necessária.

Como resultado, a massa de dados bruta resultante possui registrados **24.799 usuários**, ilustrado na Figura 32. Tal resultado será um dos atributos essenciais para a realização das recomendações a partir de filtragem colaborativa, justamente por ser o responsável por conectar o histórico de avaliações com as obras.

Figura 32 - Demonstração de parte dos dados coletados de usuários.

	A	B
1	user_id	username
2	1	BorisSoad
3	2	kurosaki_kabuto
4	3	crystaldawwn
5	4	Trojan_Invasion
6	5	Foolness
7	6	keasty
8	7	InformationGeek
9	8	ohooh
10	9	DynamicDylan
11	10	butekkusu
12	11	Totalphantasma
13	12	tragedyhero
14	13	Greaseboy
15	14	Kei_XII
16	15	Hikkykun
17	16	rokkugoh
18	17	Dong_Bong_Wong
19	18	N201_ASYLUM
20	19	Po_and_Dong
21	20	TheWildBeanz
22	21	humanofobia
23	22	keisha13ph
24	23	Fguyretftgu7
25	24	MangaGreat
26	25	Yarghov
27	26	WazEvergarden
28	27	Kelvin_0815
29	28	HotDadEren
30	29	Inspektical
31	30	syndel

Fonte: autoria própria.

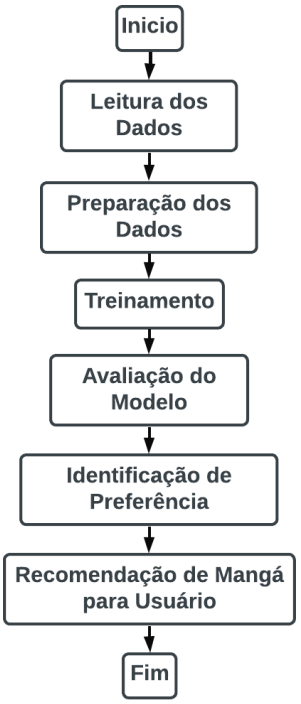
7. Recomendação de mangás

Com a massa de dados já coletada, torna-se possível utilizá-la para o treinamento e execução dos modelos escolhidos no trabalho. Os tópicos a seguir irão detalhar o processo de treinamento dos algoritmos de recomendação, assim como suas validações e resultados gerados no demonstrativo das predições.

7.1 Gerando recomendação com SVD

A primeira implementação foi em relação ao modelo SVD para geração de recomendações. Como ilustrado na Figura 33, o fluxo de aplicação se dá pelas principais etapas de manipulação, preparação e predição dos dados extraídos no Capítulo 7, Extração dos Dados.

Figura 33 - Fluxo de aplicação de recomendação utilizando SVD.



Fonte: autoria própria.

O processo de treinamento do modelo começa carregando os dados dos arquivos CSVs tratados, que contém informações sobre mangás, como identificadores únicos, títulos, e avaliações dos usuários sobre esses mangás.

Os dados são preparados para serem usados com a biblioteca Surprise. Esta biblioteca do Python é projetada justamente com ferramentas para construir, treinar e validar modelos de recomendação, que requerem um objeto Dataset contendo informações sobre usuários, itens e avaliações. O algoritmo SVD (Figura 34) é então treinado usando uma parte dos dados . Para o demonstrativo no trabalho, foi utilizado 20% da massa de dados extraída e testada para avaliar sua precisão na previsão de classificações de usuários para mangás.

Figura 34 - Pseudocódigo da montagem das predições SVD.

```
// Inicializando o algoritmo de filtragem colaborativa SVD
#svd = criar_SVD()

// Dividindo o conjunto de dados em 80% treinamento e 20% teste
#conjunto_treinamento, #conjunto_teste = dividir_conjunto_dados(#dados, tamanho_teste=0.2)
treinar #svd com #conjunto_treinamento
#previsoes = #svd realiza previsão de #conjunto_teste

// Coletando valores reais e valores previstos para métricas de desempenho
#valores_originais = listar #valor_original da predicao em #previsoes
#valores_previstos = listar #valor_previsto da predicao em #previsoes

// Calculando (MAE) e (RMSE)
#mae_svd = calcular_MAE(#valores_originais, #valores_previstos)
#rmse_svd = calcular_RMSE(#valores_originais, #valores_previstos)

// Criando um novo usuário para predição
#predicoes_usuario = aplica predição de #svd para determinado ID de usuário
#notas_estimadas = organiza #predicoes_usuario por ordem decrescente
imprimir(#notas_estimadas)
```

Fonte: autoria própria.

As previsões feitas pelo modelo são comparadas com as classificações reais usando as métricas (MAE) e (RMSE), como apresentado na Figura 34. Essas métricas fornecem medidas sobre o desempenho do modelo na previsão das classificações dos usuários realizando comparações entre um conjunto de dados original e um conjunto de dados previsto. Na aplicação deste trabalho, como está sendo utilizado notas de avaliações como parâmetro de recomendação do SVD, as métricas MAE e RMSE irão gerar valores referentes à distância de erro em que as previsões feitas estão em comparação aos dados originais utilizados como treinamento. Ou seja, uma métrica MAE ou RMSE de valor 5 significaria que as previsões em média estariam no valor de nota 5 de distância de erro.

Por fim, como representado na Figura 34, a última etapa de implementação é a realização da predição e geração das notas de estimativa para o usuário especificado, a partir desta previsão é possível realizar a organização das notas sugeridas para o usuário em relação às obras, de modo a ordená-las de maneira decrescente, evidenciando as de maior relevância como métrica positiva.

Após a implementação das etapas do SVD, foi realizada a execução do algoritmo com o intuito de gerar recomendações de leitura para o usuário especifico de **ID 1** e foram retornados os seguintes resultados em relação às métricas de desempenho:

Tabela 4 - Resultado das métricas MAE e RMSE no modelo SVD para usuário de ID 1.

MAE	RMSE
1.5427	1.9965

Fonte: autoria própria.

O valor da MAE, 1.5427, representa a média absoluta das diferenças entre as previsões do modelo e os valores reais, indicando o quão longe, em média, as previsões do modelo estão dos valores reais. Neste caso específico, uma média de 1.5427 significa que as previsões do modelo estão a uma distância média de nota de avaliação de 1.5427 dos valores reais.

Já para o valor de RMSE, 1.9965, representa a raiz quadrada da média dos quadrados das diferenças entre as previsões do modelo e os valores reais. Como o RMSE é uma métrica que penaliza mais fortemente erros maiores do que erros menores, ele é uma medida mais sensível a grandes desvios entre as previsões e os valores reais, o que resulta em um valor maior em comparação ao MAE. Isso significa que o resultado das predições de avaliações de usuários para obras estão variando com uma margem de 1.9965 de nota em comparação aos valores originais utilizados como massa de treinamento para o modelo.

Ou seja, os valores capturados de MAE e RMSE mostram que as predições de avaliações realizadas pelo modelo estão se distanciando dos valores originais da massa de dados por no máximo dois pontos de nota.

Agora, em relação à checagem do perfil do identificador 1, a quem está sendo realizada a recomendação de novas obras que se adequem ao seu perfil, pode-se observar que, as suas avaliações, com notas entre 7 e 10, como ilustrado na Figura 35.

Figura 35 - Resultado da filtragem das avaliações do usuário de ID 1.

title	rating	genres
Monster	10	['Award Winning', 'Drama', 'Mystery']
20th Century Boys	9	['Award Winning', 'Drama', 'Mystery', 'Sci-Fi']
Homunculus	9	['Drama', 'Horror', 'Mystery', 'Supernatural']
Goth	7	['Horror', 'Mystery']
Alive	6	['Horror', 'Supernatural']
Onanie Master Kurosawa	8	['Drama']

Fonte: autoria própria.

As obras as quais o leitor avaliou são caracterizadas pelos gêneros [Drama, Mystery, Sci-fi, Horror, Supernatural, Award Winning,]. Pode-se utilizar deste atributo para analisar as obras recomendadas e identificar as similaridades baseadas neste atributo. Porém, como já visto no entendimento do SVD, as sugestões vão muito além de apenas similaridade de gêneros. Como resultado da predição realizada pelo modelo, foi possível analisar que mais de 100 obras previstas pelo modelo mostraram um alto índice de nota de avaliação estimada. Na Figura 36 é possível observar que 32 obras possuem **nota estimada maior que ou igual a 9**, e não apenas isto, mas também é possível compreender a partir da análise que, entre as recomendações realizadas pelo modelo SVD, vários títulos possuem gêneros diversos que não existem dentro do conjunto de gêneros detectado na análise de perfil do usuário de ID 1, a partir das avaliações já realizadas por ele.

Figura 36 - Resultado da predição do SVD para usuário de ID 1.

title	type	genres	estimate_score
JoJo no Kim	Manga	['Action', 'Adventure', 'Mystery', 'Supernatur...	9.406589
Real	Manga	['Drama', 'Sports']	9.355916
Yagate Kimi	Manga	['Drama', 'Girls Love']	9.334997
Chikan Otok	Manga	['Comedy', 'Drama', 'Romance', 'Slice of Life']	9.314081
Berserk	Manga	['Action', 'Adventure', 'Award Winning', 'Dram...	9.287793
Pandora Hea	Manga	['Adventure', 'Fantasy', 'Mystery', 'Supernatu...	9.281173
Slam Dunk	Manga	['Award Winning', 'Sports']	9.247782
Skip Beat!	Manga	['Comedy', 'Drama', 'Romance']	9.215642
Neon Genesi	Manga	['Action', 'Drama', 'Mystery', 'Sci-Fi']	9.208576
Watashitach	Manga	['Drama', 'Romance']	9.168533
Watashitach	Manga	['Drama', 'Romance']	9.168533
Fruits Bask	Manga	['Award Winning', 'Drama', 'Romance', 'Superna...	9.152732
Haikyuu!!Ha	Manga	['Award Winning', 'Sports']	9.118736
The Boxer	Manhwa	['Drama', 'Sports']	9.118454
Yotsuba to!	Manga	['Award Winning', 'Comedy', 'Slice of Life']	9.114075
Dengeki Dai	Manga	['Comedy', 'Drama', 'Romance']	9.095128
Ashita no J	Manga	['Drama', 'Slice of Life', 'Sports']	9.091215
Kyou kara O	Manga	['Action', 'Comedy']	9.079268
Blue Period	Manga	['Award Winning', 'Drama']	9.078348
Saezuru Tor	Manga	['Boys Love', 'Drama', 'Erotica']	9.076634
Tsubasa: RE	Manga	['Action', 'Adventure', 'Drama', 'Fantasy']	9.073543
Monster	Manga	['Award Winning', 'Drama', 'Mystery']	9.065630
Death Note	Manga	['Supernatural', 'Suspense']	9.062276
Fullmetal A	Manga	['Action', 'Adventure', 'Award Winning', 'Dram...	9.050298
Koe no Kata	Manga	['Award Winning', 'Drama']	9.047680
Saikyou Den	Manga	['Action', 'Comedy', 'Drama', 'Suspense']	9.033183
ReLIFE	Manga	['Comedy', 'Drama', 'Romance', 'Slice of Life']	9.026478
One Piece	Manga	['Action', 'Adventure', 'Fantasy']	9.015038
Annarasuman	Manhwa	['Drama', 'Mystery', 'Romance']	9.013483
Omniscient	Novel	['Action', 'Adventure', 'Fantasy']	9.011817
Gakuen Alic	Manga	['Comedy', 'Drama']	9.004031
Vagabond	Manga	['Action', 'Adventure', 'Award Winning']	9.001426

Fonte: autoria própria.

Tendo em vista que as métricas MAE e RMSE demonstram um índice de ~1.5 a ~1.9 de variação de erro referente às notas de avaliações dos usuários. Tomando como partida estas 32 recomendações geradas com nota estimada de 9 ou mais, é possível compreender que elas estão destoando dos valores originais, podendo ser classificadas na verdade entre notas 7,5 e 7.

7.2 Gerando recomendação com Coeficiente de Correlação de Pearson

Na implementação da recomendação utilizando coeficiente de correlação de Pearson, foi realizada uma abordagem simples por falta de tempo para aprimoramento do trabalho, então a implementação foi aplicada somente na geração da medida de similaridade em sí, impossibilitando de aplicar métricas de desempenho por comparação de previsões como realizado anteriormente com MAE e RMSE. O conceito por trás é diferente, se comparada à aplicação feita com SVD no Tópico 7.1, da aplicação bruta do Coeficiente de Correlação de Pearson. A recomendação realizada leva em conta as avaliações para com as obras, porém a recomendação é baseada em semelhança entre obras. No início da implementação, foi gerado um DataFrame utilizando a biblioteca Pandas com a seguinte estrutura: as colunas da matriz são os identificadores das obras referentes ao arquivo “mangas.csv”, os índices das linhas são

os identificadores dos usuários, e os valores de coluna-linha são as notas dadas pelos usuários para as obras, construindo assim uma matriz de relacionamentos das avaliações, como ilustrado na Figura 37.

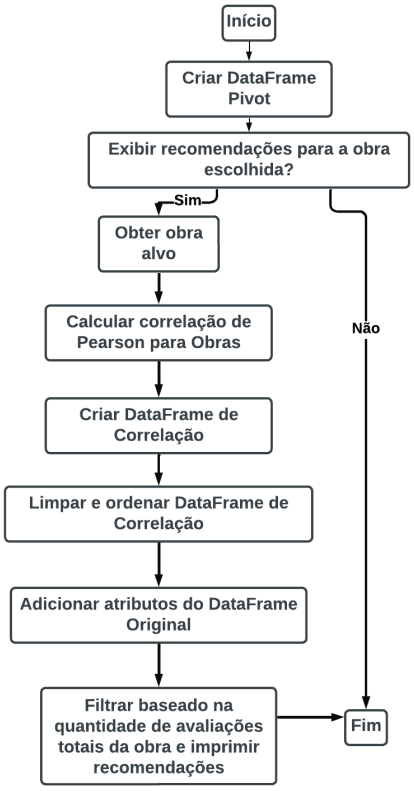
Figura 37 - Representação do dataset pivot para correlação de Pearson.

X	manga1	manga2	manga3	manga4
user1	rate(1,1)	rate(1,2)	rate(1,3)	rate(1,4)
user2	rate(2,1)	rate(2,2)	rate(2,3)	rate(2,4)
user3	rate(3,1)	rate(3,2)	rate(3,3)	rate(3,3)
user4	rate(4,1)	rate(4,2)	rate(4,3)	rate(4,4)

Fonte: autoria própria.

A partir disso, como representado na Figura 38, o código executa um método, cujo parâmetro é o identificador da obra que deseja buscar por recomendações similares, juntamente com um parâmetro que define a quantidade mínima de avaliações que uma obra deve ter para ser levada em consideração como resultado de sugestão da medida de correlação, já que quanto menor for a quantidade de avaliações que uma obra possui, menos precisa será a medida de similaridade quando correlacionada com outras obras.

Figura 38 - Fluxo de funcionamento da recomendação por correlação de Pearson.



Fonte: autoria própria.

O método de recomendação, assim como observado na Figura 38, realiza a seleção da parte da matriz que possui os valores referentes à obra que foi usada como alvo. Em seguida, aplica um método do Pandas chamado **Dataframe.corrwith()**, que calcula a correlação de pares entre linhas e colunas com o dataset da obra selecionada. A partir dessa correlação, é gerado um novo dataset com os valores da correlação aplicados em uma coluna para as obras e adicionando algumas informações extras para auxílio em análises, como a contagem de avaliações da obra, título da obra e a média de avaliação.

Como prática, foi executado o algoritmo utilizando como obra alvo de recomendação por similaridade o mangá de **ID 2, de título Berserk**.

Como resultado da sugestão realizada pela medida (Figura 39), além da própria obra com valor máximo, resultado esperado já que é a correlação mais próxima dele mesmo, houve resultados de outros valores com medida de valor máximo (1.0).

Como parte negativa da implementação, é dedutível que a quantidade de avaliações dentro de uma obra, para este tipo de método, faz bastante diferença, já que é necessária para gerar a medida de maneira precisa e fidedigna. Obras, por exemplo, que possuem poucas avaliações seriam evitadas utilizar como recomendação neste tipo de método por não garantir uma similaridade verídica ou imprecisa. No entanto, a utilização do Coeficiente de Correlação de Pearson garante uma lógica por trás da recomendação que pode servir como uma boa ferramenta de auxílio na escolha e recomendação de obras baseadas em item.

Figura 39 - Resultado da geração de correlação de Pearson por obra de ID 2.

title	genres	PearsonR	count	mean
Dorohedoro	['Action', 'Comedy', 'Fantasy', 'Horror']	1.000000	53	8.792453
Devilman	['Action', 'Adventure', 'Drama', 'Fantasy', 'Horror', 'Sci-Fi', 'Supernatural']	1.000000	26	7.076923
Hellsing	['Action', 'Horror', 'Supernatural']	1.000000	20	7.900000
Domestic na	['Drama', 'Romance']	1.000000	234	6.427350
Trigun Maxi	['Action', 'Adventure', 'Award Winning', 'Comedy', 'Drama', 'Sci-Fi']	1.000000	12	7.666667
Berserk	['Action', 'Adventure', 'Award Winning', 'Drama', 'Fantasy', 'Horror', 'Supernatural']	1.000000	257	9.163424
ES: Eternal	['Drama', 'Sci-Fi', 'Supernatural']	1.000000	13	8.538462
KiseijuuPar	['Action', 'Award Winning', 'Horror', 'Sci-Fi']	1.000000	27	8.666667
Psyren	['Action', 'Adventure', 'Romance', 'Sci-Fi', 'Supernatural']	1.000000	26	7.884615
Bakuman.Bak	['Comedy', 'Drama', 'Romance']	1.000000	63	8.428571
Eden no Ori	['Action', 'Adventure', 'Fantasy', 'Ecchi']	1.000000	52	6.942308
Eden no Ori	['Action', 'Adventure', 'Fantasy', 'Ecchi']	1.000000	52	6.942308
20th Centur	['Award Winning', 'Drama', 'Mystery', 'Sci-Fi']	1.000000	72	8.888889
Saint Seiya	['Action', 'Adventure']	1.000000	9	7.666667
Shijou Saik	['Action', 'Comedy', 'Drama', 'Ecchi']	1.000000	20	8.000000
Real	['Drama', 'Sports']	1.000000	24	9.666667
Ao no Exorc	['Action', 'Fantasy']	1.000000	20	8.400000
Haikyuu!!Ha	['Award Winning', 'Sports']	1.000000	45	9.044444
Ao no Exorc	['Action', 'Fantasy']	1.000000	20	8.400000
Dr. Stone	['Adventure', 'Award Winning', 'Sci-Fi']	1.000000	68	7.897059
RaveRave Ma	['Adventure', 'Comedy', 'Fantasy']	1.000000	35	7.971429
Lookism	['Action', 'Comedy', 'Drama', 'Supernatural']	1.000000	41	7.975610
Kingdom	['Action', 'Award Winning']	0.902829	46	9.108696
Kingdom	['Action', 'Award Winning']	0.902829	46	9.108696
Made in Aby	['Adventure', 'Drama', 'Fantasy', 'Sci-Fi']	0.866025	55	6.727273
Fullmetal A	['Action', 'Adventure', 'Award Winning', 'Drama', 'Fantasy']	0.848528	56	9.071429
Gantz	['Action', 'Drama', 'Horror', 'Sci-Fi', 'Supernatural']	0.836428	100	7.820000
Homunculus	['Drama', 'Horror', 'Mystery', 'Supernatural']	0.834966	52	8.250000
Claymore	['Action', 'Adventure', 'Fantasy', 'Horror']	0.790569	90	8.100000
UzumakiUzum	['Drama', 'Horror', 'Supernatural']	0.679366	107	7.700935
JoJo no Kim	['Action', 'Adventure', 'Award Winning', 'Mystery', 'Supernatural']	0.674200	60	8.583333
Vinland Sag	['Action', 'Adventure', 'Award Winning', 'Drama']	0.665016	81	8.296296

Fonte: autoria própria.

O algoritmo aplica a medida de similaridade para todas as obras do conjunto de dados existente porém, como observado nas primeiras ocorrências da medida, são poucas as recomendações com um índice de perfeita semelhança (PearsonR com valor 1) com o item alvo da recomendação, e após isto as seguintes recomendações em ordem decrescente vai diminuindo drasticamente, até que sobre apenas itens com medidas dissimilar. Em comparação com as recomendações realizadas pelo SVD, que retornou mais de 32 obras ainda com nota prevista acima de 9, e as posteriores ainda com valores interessantes de notas 8,5 e 8, as recomendações feitas pelo Coeficiente de Correlação de Pearson mostrou apenas 22 com medidas de PearsonR altas (1) e 12 registros entre 0.9 e 0.3, o restante foram todas dissimilares.

8. Conclusões e Trabalhos Futuros

Neste trabalho, foram elaborados os passos e mostrado um possível método de extração de informações para reunir massa de dados com o intuito de realizar recomendações já com valores iniciais. Foram exploradas duas abordagens distintas para a geração de recomendações de mangás: o modelo de Decomposição em Valores Singulares (SVD) e o Coeficiente de Correlação de Pearson. Ambas as metodologias apresentaram resultados coerentes, cada uma com suas próprias vantagens, desafios e formas de lidar com os dados para realização de sugestões.

O uso do SVD revelou-se eficaz na identificação de padrões latentes nos dados de avaliação dos usuários. Por meio da análise das previsões do modelo, foi possível observar sua capacidade de recomendar mangás com base nas preferências individuais dos usuários a partir da interação entre usuário e obra por meio do histórico. No entanto, as métricas de desempenho, como o Mean Absolute Error (MAE) e o Root Mean Squared Error (RMSE), mesmo mostrando um desvio relativamente baixo de no máximo 2 casas de distância das notas de avaliação originais, destacaram a necessidade contínua de refinamento do modelo para melhorar sua precisão, bem como a possibilidade de realizar testes com outros modelos para validar um possível ganho de desempenho em relação à massa de dados utilizada neste projeto.

Por outro lado, a abordagem baseada no Coeficiente de Correlação de Pearson demonstrou uma lógica direta ao recomendar mangás com base na similaridade das avaliações entre obras. Embora essa metodologia dependa significativamente do número de avaliações disponíveis para cada obra, ela oferece uma alternativa válida para a recomendação personalizada, especialmente em casos onde há uma grande quantidade de dados disponíveis e em cenários onde a esparsidade de avaliações é baixa em relação à quantidade de obras existentes.

Como avaliação dos resultados, foi possível identificar que a utilização do SVD consegue entregar muito mais recomendações, diferente das recomendações feitas pelo Coeficiente de Correlação de Pearson, que resulta em consideravelmente menos recomendações precisas de similaridade de obras, justamente por decair bastante ao decorrer da distribuição de semelhança. Por outro lado, o SVD também tem a capacidade de gerar recomendações mais diversificadas em termo de gênero de obra, oque é um benefício enorme quando se quer evitar cenários as quais o usuário fica estagnado em uma bolha e não é estimado a encontrar novos interesses em sua leitura.

Um dos problemas enfrentados foi relacionado à utilização do Coeficiente de Correlação de Pearson. Com o que foi estudado durante o trabalho, entende-se que seria uma melhor opção abordar esta medida utilizando algum algoritmo de recomendação como base, como por exemplo a utilização da medida de correlação de Pearson na implementação de um algoritmo KNN, para que seja capaz de aplicar métricas MAE e RMSE, com o objetivo comparativo para com o modelo de SVD. Tal abordagem pode ser objetificada como um futuro trabalho na área.

Este estudo contribui para o entendimento do campo de recomendação de mangás, fornecendo compreensão sobre diferentes técnicas e abordagens tanto para a obtenção de informações necessárias para a aplicabilidade dos modelos, como a compreensão dos cenários

onde determinados modelos conseguem se beneficiar mais. No futuro, recomenda-se a exploração de métodos híbridos que combinem o poder do SVD com a lógica baseada em similaridade do Coeficiente de Correlação de Pearson, com o objetivo de alcançar recomendações mais precisas para os usuários de plataformas de mangá.

Além disso, a investigação de outras técnicas de aprendizado de máquina e a coleta de dados adicionais podem enriquecer ainda mais a qualidade das recomendações oferecidas aos usuários, como a utilização de treinamentos por reforço baseados em feedbacks dos usuários do sistema, ou até mesmo a utilização de LLM's, como o ChatGPT, para o auxílio na estruturação e recomendação de novas obras baseadas em contexto e conteúdo.

9. Referências

1. STATISTA. Manga and anime market size worldwide 2017-2025. Disponível em: https://www.statista.com/statistics/1172810/manga-and-anime-market-size-worldwide/. Acesso em: 6 dez. 2021.
2. TECHNAVIO. Manga Market by Type and Geography - Forecast and Analysis 2020-2024. Disponível em: https://www.technavio.com/report/manga-market-industry-analysis. Acesso em: 6 dez. 2021.
3. BRUNIALTI, Lucas F.; PERES, Sarajane M.; FREIRE, Valdinei; LIMA, Clodoaldo A. M. Aprendizado de Máquina em Sistemas de Recomendação Baseados em Conteúdo Textual: Uma Revisão Sistemática. Goiânia: SBC, 2015. XI Brazilian Symposium on Information System, p. 203-212.
4. LUDERMIR, Teresa Bernarda. Inteligência artificial e aprendizado de máquina: estado atual e tendências. Estudos Avançados, v. 35, n. 101, p. 85-94, 2021. Acesso em: 16 fev. 2021.
5. ALVAREZ, E. B., SIRIANI, A. L. R., VIDOTTI, S. A. B. G., & CARVALHO, A. M. G. D. (2016). Os sistemas de recomendação, arquitetura da informação e a encontrabilidade da informação. *Transinformação*, *28*, 275-286.
6. COLMENERO-FERREIRA, Fernando; OLIVEIRA, Adicinéia Aparecida de. Os sistemas de recomendação na web como determinantes prescritivos na tomada de decisão. **JISTEM-Journal of Information Systems and Technology Management**, v. 9, p. 353-368, 2012.
7. ZHANG, Qian; LU, Jie; JIN, Yaochu. Artificial intelligence in recommender systems. **Complex & Intelligent Systems**, v. 7, p. 439-457, 2021.
8. BEREGOVSKAYA, Irina; KOROTEEV, Mikhail. Review of Clustering-Based Recommender Systems. **arXiv preprint arXiv:2109.12839**, 2021.
9. FELFERNIG, Alexander et al. An overview of recommender systems and machine learning in feature modeling and configuration. In: **Proceedings of the 15th International Working Conference on Variability Modelling of Software-Intensive Systems**. 2021. p. 1-8.
10. TARUS, John K.; NIU, Zhendong; YOUSIF, Abdallah. A hybrid knowledge-based recommender system for e-learning based on ontology and sequential pattern mining. **Future Generation Computer Systems**, v. 72, p. 37-48, 2017.
11. LUDERMIR, Teresa Bernarda. Inteligência Artificial e Aprendizado de Máquina: estado atual e tendências. **Estudos Avançados**, v. 35, p. 85-94, 2021.
12. BURCHFIEL, Anni.. What is NLP (Natural Language Processing) Tokenization?. TokenEx, 2022. Disponível em: https://www.tokenex.com/blog/ab-what-is-nlp-natural-language-processing-tokenization. Acesso em: 7 de setembro de 2023.
13. ROY, Kavika. Importance of Datasets in Machine Learning and AI Research. 2022. Disponível em: https://medium.com/datatobiz/importance-of-datasets-in-machine-learning-and-ai-research-ac3eb97ba875. Acesso em: 7 de setembro de 2023.
14. WOLNIAK, Radosław. The Design Thinking method and its stages. **Systemy Wspomagania w Inżynierii Produkcji**, v. 6, n. 6, p. 247-255, 2017.
15. ISINKAYE, Folasade Olubusola; FOLAJIMI, Yetunde O.; OJOKOH, Bolande Adefowoke. Recommendation systems: Principles, methods and evaluation. Egyptian informatics journal, v. 16, n. 3, p. 261-273, 2015.

16. KHANAL, Shristi Shakya et al. A systematic review: machine learning based recommendation systems for e-learning. Education and Information Technologies, v. 25, p. 2635-2664, 2020.
17. PINTO, Miguel Angelo Gaspar. Sistema Híbrido de Recomendação de Produtos com Uso de Filtros Colaborativos e Números Fuzzy. 2011. Tese de Doutorado. PUC-Rio.
18. SU, Xiaoyuan; KHOSHGOFTAAR, Taghi M. A survey of collaborative filtering techniques. Advances in artificial intelligence, v. 2009, 2009.
19. CAVALCANTI JÚNIOR, Flávio de Holanda. Avaliação de Técnicas de Filtragem Colaborativa para Sistemas de Recomendação. Trabalho de Conclusão de Curso, Universidade Federal de Pernambuco, Recife, 2017. Disponível em: https://repositorio.ufpe.br/bitstream/123456789/23425/1/TCC%20FL%C3%81VIO%20DE%20HOLANDA%20CAVALCANTI%20J%C3%9ANIOR.pdf.
20. PARANHOS, Ranulfo et al. Desvendando os mistérios do coeficiente de correlação de Pearson: o retorno. Leviathan (São Paulo), n. 8, p. 66-95, 2014.
21. CHAMBERLAIN, P. (2021). Estudo comparativo de algoritmos de sistemas de recomendação de filmes. Trabalho de Conclusão de Curso (Graduação em Ciência da Computação) – Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro. Disponível em: https://www.maxwell.vrac.puc-rio.br/57546/57546.PDF.
22. DO NASCIMENTO, Raimundo Nonato N.; DA MATA LIBÓRIO FILHO, João. Avaliação de Técnicas de Filtragem Colaborativa Aplicadas para Recomendação de Cursos em um Ambiente Virtual de Aprendizagem. Anais da Semana de Informática CESIT/UEA, v. 7, n. 1, p. 11-11, 2019.
23. MCKINNEY, Wes. Python para análise de dados: Tratamento de dados com Pandas, NumPy e IPython. Novatec Editora, 2018.
24. SALLAM, Rouhia M.; HUSSEIN, Mahmoud; MOUSA, Hamdy M. An enhanced collaborative filtering-based approach for recommender systems. Int. J. Comput. Appl, v. 176, n. 41, p. 9-15, 2020.
25. SURPRISE. (2024). Biblioteca Python para sistemas de recomendação. Disponível em: https://surpriselib.com/.
26. RAHARJO, Daniel Adrian. MANGA RECOMMENDATION USING NAIVE BAYES AND DECISION TREE ALGORITHM. 2022. Tese de Doutorado. Universitas Katholik Soegijapranata Semarang.
27. VIE, Jill-Jênn et al. Using posters to recommend anime and mangas in a cold-start scenario. In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). IEEE, 2017. p. 21-26.
28. WASNIK, A. (2020). Singular Value Decomposition (SVD) in Python. AskPython. Disponível em: https://www.askpython.com/python/examples/singular-value-decomposition. Acesso em: 07 de Março de 2024.
29. BROWNLEE, Jason (2019). How to Calculate the SVD from Scratch with Python. machinelearningmastery. Disponível em: https://machinelearningmastery.com/singular-value-decomposition-for-machine-learning/. Acesso em: 07 de Março de 2024.
30. STATISTCSHOWTO. Correlation Coefficient: Simple Definition, Formula, Easy Steps. Disponível em: https://www.statisticshowto.com/probability-and-statistics/correlation-coefficient-formula/. Acesso em: 08 de Março de 2024.
31. OLIVEIRA, Bruno (2019). Coeficientes de correlação. statplace. Disponível em: https://statplace.com.br/blog/coeficientes-de-correlacao/. Acesso em: 08 de Março de 2024.

32. GUPTA, Ritika (2023). How Singular Value Decomposition (SVD) is used in Recommendation Systems, “Clearly Explained”. Diposnível em: https://medium.com/@ritik_gupta/how-singular-value-decomposition-svd-is-used-in-recommendation-systems-clearly-explained-201b24e175db. Acesso em: 08 de Março de 2024.