UNIVERSIDADE FEDERAL DO MARANHÃO
CENTRO DE CIÊNCIAS EXATAS E TECNOLÓGICA
CURSO DE CIÊNCIA DA COMPUTAÇÃO

**ROSIENE BALTAZAR DE CARVALHO**

# ANÁLISE E RECOMENDAÇÃO DE LIVROS BASEADO EM GRUPOS DE USUÁRIOS

São Luís/MA
2021

**ROSIENE BALTAZAR DE CARVALHO**

# ANÁLISE E RECOMENDAÇÃO DE LIVROS BASEADO EM GRUPOS DE USUÁRIOS

Monografia apresentada ao curso de Ciência da Computação da Universidade Federal do Maranhão, como parte dos requisitos necessários para obtenção do grau de Bacharel em Ciência da Computação. Orientador: Prof. Dr. Mário Meireles Texeira

São Luís/MA

2021

Ficha gerada por meio do SIGAA/Biblioteca com dados fornecidos pelo(a) autor(a).
Diretoria Integrada de Bibliotecas/UFMA

CARVALHO, ROSIENE BALTAZAR DE.
ANÁLISE E RECOMENDAÇÃO DE LIVROS BASEADO EM GRUPOS DE USUÁRIOS / ROSIENE BALTAZAR DE CARVALHO. - 2021.
43 f.

Orientador(a): Mário Meireles Texeira.
Monografia (Graduação) - Curso de Ciência da Computação, Universidade Federal do Maranhão, São Luís - MA, 2021.

1. Agrupamento. 2. Avaliações de usuários. 3. Livros. 4. Recomendação. 5. Tags. I. Texeira, Mário Meireles. II. Título.

**ROSIENE BALTAZAR DE CARVALHO**

# ANÁLISE E RECOMENDAÇÃO DE LIVROS BASEADO EM GRUPOS DE USUÁRIOS

Monografia apresentada ao curso de Ciência da Computação da Universidade Federal do Maranhão, como parte dos requisitos necessários para obtenção do grau de Bacharel em Ciência da Computação. Orientador: Prof. Dr. Mário Meireles Texeira

Aprovada em ___/___/_____

---

**Prof. Dr. Mário Meireles Texeira**
(Orientador)
Universidade Federal do Maranhão

---

**Prof. Dr. Carlos De Salles Soares Neto**
Examinador 1

---

**Prof. Me. Carlos Eduardo Portela Serra de Castro**
Examinador 2

São Luís/MA

2021

# Agradecimentos

Vencer não é sinônimo de não fracassar, mas chegar ao fim e saber recomeçar, e nunca sozinhos. Por isso, quero externar minha gratidão a todos aqueles que estivem comigo nesta caminhada.

A Deus, que me sustentou em cada instante, que me guiou e me deu capacidade de chegar aqui. A glória somente a Ele.

Aos meus pais, que lutaram fortemente para que fosse possível o meu ingresso em uma universidade. E concluir, é minha forma de honrar-los.

A minha irmã, que esteve em cada momento e sempre me ensinou com seu exemplo que não podemos desistir, por me mostrar a importância da universidade.

A meu orientador, Mário Meireles, por sua paciência durante esse ano de produção da monografia, por não ter desistido de ser meu orientador, minha grande gratidão a você.

A minha querida amiga Mara Eduarda Alencar, que chegou no final dessa caminhada, mas foi minha grande auxiliadora na conclusão, que nos momentos de desânimo me ajudou a não parar, que no desenvolvimento deste trabalho me ajudou do início ao fim. É enorme minha gratidão a você. Obrigada por tanto.

Aos meus amigos que estiveram comigo na luta de todo dia no campus: Felipe Breno, obrigada por todas caronas que me possibilitaram chegar a ufma a tempo das aulas. Erickson Bruno, por sempre repetir "siow, vai dá"e meu amigo, deu, conseguimos. E por último, ao meu grande amigo Domingos Alves, muito do que aprendi em programação devo a você, sem dúvidas, você tornou essa caminhada mais leve, obrigada por tudo que você fez.

As crenteanes, que de tantas vezes em desanimo, foram os braços, palavras e orações que me ajudaram a continuar. A caminhada árdua da UFMA se tornou leve por ter vocês. Josane, Elen, Bruna, Priscila e Kelayne obrigada por tudo, levo todo aprendizado construído com vocês para minha vida.

Aos meu amigos que sempre me impulsionaram e motivaram a continuar: Mônica Danielle, sou grata por cada palavra que você externava, por sempre me lembrar que não poderia parar e que deveria formar. Ao Pr. Eric Karley, por me ensinar a glorificar a Deus nesta caminha de UFMA. À Edneia Reis, Carol Brito e Nayure Matias, vocês foram motivadores essenciais para que eu chegasse até aqui, de maneira indireta e direta, vocês facilitaram minha caminhada, minha gratidão a vocês.

Por último, quero agradecer a meu amigo Davi Nascimento, companheiro de

trabalho que em meio a tantas demandas, não deixou de me orientar neste trabalho, obrigada por cada vez que em tranquilidade me acalmou e disse que iria conseguir finalizar, não tenho palavras para agradecer o que você fez.

# Resumo

Recomendação de conteúdo surge da necessidade de se encontrar um item, seja uma notícia, livros ou filmes, em meio a bilhões de informações disponíveis na web que seja de interesse de um usuário. Seu principal alvo é proporcionar ao usuário facilidade de encontrar o que deseja de maneira mais rápida e eficaz. E desta feita, esse trabalho visa propor uma solução a este problema, através de uma recomendação de filmes baseada em conteúdo com a utilização de tags e de uma análise do comportamento dos grupos de usuários, visando o aperfeiçoamento na recomendação. Para este objetivo, utilizou-se o algoritmo de treinamento K-nn, que tem sido usado de maneira abrangente em recomendações, primeiro aplicou-se a relação de filmes e tags, que são as categorias dada aos filmes pelos usuários, dessa forma, gerando filmes candidatos. Logo após, foi aplicado a relação de usuários com filmes candidatos, que são as avaliações que cada usuário atribui ao conteúdo assistido, sendo assim, obtendo-se os filmes para recomendação.A abordagem ocorreu desta forma visando contornar o problema que surgiu com os sistemas de recomendações, que é a partida a frio, onde muitas vezes, o item escolhido pelo usuário não possui muitas avaliações. A metodologia foi aplicada à base de livros goodbook-10k, onde contem 10.000 livros com 980.000 avaliações de usuários. Desse modo, o objetivo do trabalho foi alcançado, que é efetuar recomendações de filmes aos usuários. Para analise do comportamento do usuários que são alvos da recomendação, utilizou-se um agrupamento utilizando o K-means para extrair as características dos grupos de usuários leitores, onde encontrou-se 10 grupos, mas todos expressaram um comportamento em destaque, todos são leitores predominantemente de livros ligados a ficção, crime e suspense. Portanto, contendo estas informações o sistema de recomendação pode ser treinado e cada vez mais aperfeiçoado de acordo com o comportamento dos usuários e suas leituras e dessa forma proporcionando uma experiência boa e direcionada ao conteúdo de interesse de maneira eficaz e rápida.

**Palavras-chaves:** Recomendação, livros, agrupamento, avaliações de usuários, tags.

# Abstract

Content recommendation arises from the need to find an item, be it news, books or films, amid billions of information available on the web that is of interest to a user. Its main target is to provide the user with the facility to find what they want more quickly and effectively. And this time, this work aims to propose a solution to this problem, through a recommendation of films based on content with the use of tags and an analysis of the behavior of the groups of users, aiming at the improvement in the recommendation. For this purpose, the K-nn training algorithm was used, which has been used comprehensively in recommendations. First, the list of films and tags was applied, which are the categories given to films by users, thus generating candidate films. Soon after, the relationship of users with candidate films was applied, which are the evaluations that each user attributes to the watched content, thus obtaining the films for recommendation. The approach occurred in this way aiming to circumvent the problem that arose with the systems of recommendations, which is the cold start, where often, the item chosen by the user does not have many evaluations. The methodology was applied to a goodbook-10k book base, which contains 10,000 books with 980,000 user reviews. Thus, the objective of the work was achieved, which is to make film recommendations to users. To analyze the behavior of the users who are targets of the recommendation, a grouping was used using K-means to extract the characteristics of the groups of reader users. 10 groups were found, but all expressed a prominent behavior, all of whom are predominantly readers of books related to fiction, crime and suspense. Therefore, containing this information, the recommendation system can be trained and increasingly refined according to the behavior of users and their readings and thus providing a good experience and directed to the content of interest in an effective and fast way.

**Keywords:** recommendation, books, grouping, user ratings, tags.

# Lista de ilustrações

# Lista de tabelas

# Sumário

# 1 INTRODUÇÃO

Os sistemas de recomendação (RSs) são ferramentas e técnicas que oferecem sugestões para itens com maior probabilidade de interesse para um usuário alvo. As sugestões dizem respeito a vários processos de tomada de decisão, como: quais itens comprar, quais músicas ouvir ou quais livros ler (BONIN, 2018). "Item"é o termo geral usados para denotar o que o sistema recomenda aos usuários.

Existem dois grandes desafios para uma recomendação de conteúdo. O primeiro é a escalabilidade do sistema, ou seja, o crescimento de forma a não impactar no seu desempenho, visto que com o excesso de informações nos dias atuais, torna-se necessário uma manipulação de milhares de dados em tempo real. A segunda é a de aperfeiçoar as recomendações geradas pelo sistema (MEDEIROS, 2013). Portanto, o desafio maior do sistema de recomendação é determinar uma classificação que represente bem o grau de interesse de um determinado usuário.

Muitos trabalhos foram feitos buscando vencer esses desafios da recomendação de conteúdo, como (MEDEIROS, 2013), que aplicou um experimento para averiguar a eficiência das recomendações, realizando uma comparação entre recomendação colaborativa baseada em conteúdo, colaborativa baseada em usuário e híbrida. Tendo como foco principal a recomendação colaborativa, ele utilizou os algoritmos de similaridade do cosseno, correlação e cosseno ajustado. O objetivo era obter o tipo de recomendação que produzisse uma recomendação escalável.

Por conseguinte, outros trabalhos que buscam propor melhorias nas recomendações de conteúdos, segundo (CAZELLA; NUNES; REATEGUI, 2010), propõe uma recomendação baseada em competências, que são qualificações necessárias de cada aluno, dessa forma, construindo perfis de usuários. (FERRO et al., 2010) desenvolveu um sistema para recomendar materiais didáticos aos usuários de acordo com seu perfil, para isso utilizou rankings nas filtragens colaborativas baseada em conteúdo e itens com maior evidência. (GARCIA; FROZZA, 2013) contruiu uma recomendação de conteúdo utilizando mineração de dados, clusterização e regras de associação. Outros que seguem a linha de recomendação de conteúdo baseado em itens, são (SAMPAIO, 2006); (LIMA, 2012); (COSTA; AGUIAR; MAGALHÃES, 2013).

Em vista disso, trabalhos com a abordagem de recomendação de conteúdo utilizando tags são mais raros, principalmente o que foi utilizado por (OLIVEIRA; COELLO, 2013), que faz uma abordagem de recomendação híbrida, utilizando etiquetagem social e filtragem colaborativa. O foco deste trabalho consiste em uma alternativa para buscar melhoria no desempenho do sistema de recomendação, que buscou reduzir os problemas de dados

esparsos como da partida a frio. A técnica para isto foi através da semelhança entre usuários e suas tags e depois ir aos livros para verificar se estes tem as tags candidatas e assim efetuar sua recomendação. (LIANG et al., 2009) utiliza a relação tridimensional entre usuário, item e tag para aperfeiçoar suas recomendações.

Em suma, para contornar o primeiro desafio na recomendação, se faz necessário uma análise dos usuários e dos perfis formados a partir das recomendações, isto é, um agrupamento dos usuários. Agrupamento é um método não supervisionado de aprendizado, que agrupa itens de forma que são mais similares quando comparado a outro grupo, segundo (RICCI F.; KANTOR, 2010). Dessa forma, alguns trabalhos buscaram solucionar essa questão. (SANTOS et al., 2017) desenvolve um sistema de recomendação baseada agrupamento, utilizando o método de agrupamento K-means. O (HEINZEN; MARTINS, 2018) que utiliza o método de agrupamento para analise de perfis de consumidores, com o objetivo de coletar e avaliar os itens consumidos por um usuário para então classificar o seus perfis de consumo.

Mediante o exposto, torna-se necessário a exploração dos métodos apresentados nos mais diversos cenários. Tendo em vista os problemas externados e soluções propostas por outros trabalhos, este trabalho irá por sua vez, propor uma recomendação de conteúdo baseada em itens utilizando tags, além disso, também uma análise apurada do comportamento desses usuários realizando seus agrupamentos e perfis, visando a análise do comportamento destes. Sendo assim, realizar o melhoramento frequente do sistema de recomendação.

## 1.1 Objetivos

O objetivo desde trabalho é propor uma recomendação colaborativa de livros e efetuar uma análise das características dos grupos de usuários em suas leituras.

Destaca-se como objetivos específicos desse trabalho:

- Desenvolver um sistema de recomendação de conteúdo baseado em itens
- Utilizar tags para aperfeiçoar a recomendação de conteúdo
- Realizar a relação de grupos e caracterizar sua relação, destacando seus principais comportamentos.

## 1.2 Estrutura do trabalho

Este trabalho está organizado em cinco capítulos, demonstração de modo claro os conteúdos conforme os parágrafos a seguir.

O Capítulo 1, expõe o cenário deste trabalho, descrevendo os trabalhos que são relacionados e suas características.

O Capítulo 2, é exposto a fundamentação teórica necessária para que se obtenha o entendimento necessário para este estudo.

O Capítulo 3, está descrita a metodologia abordada neste trabalho.

O Capítulo 4, neste capítulo estão apresentados e discutidos os Resultados obtidos neste trabalho.

O Capítulo 5, são discorridos a Conclusão e os trabalhos futuros.

# 2 Fundamentação teórica

Neste capítulo denota a fundamentação teórica disposta no corpo deste trabalho, essencial para a apresentação dos procedimentos utilizados para se obter os objetivos delineados pelo mesmo.

## 2.1 Recomendação de conteúdo

O sistema de recomendação é a solução para o problema de estimar a avaliação para uma determinada unidade que ainda não foi avaliada por um determinado usuário, ou seja, é um conjunto de algoritmos que utilizam técnicas de aprendizagem de máquina e recuperação da informação para gerar recomendações baseadas em algum tipo de filtragem (MEDEIROS, 2013). Esse conjunto de técnicas ajudam os usuários em suas tarefas de busca de informações, sugerindo itens (produtos, serviços ou informações) que melhor atendem às suas necessidades.

Um dos maiores desafios dessa área é a produção de recomendações de qualidade, manipulando, em algumas vezes, uma quantidade significativa de dados independente das condições adversas que estes dados se encontram. Segundo o autor supracitado, os sistemas de recomendação são frequentemente categorizados conforme a configuração de como eles obtêm essa estimativa, sendo elas: filtragem por conteúdo, filtragem colaborativa e híbrida.Vejamos a seguir:

1. Filtragem por conteúdo: se baseia na premissa de que os usuários gostariam de obter recomendações de itens semelhantes a itens preferidos do usuário no passado.

2. Filtragem colaborativa: fundamenta-se na técnica chamada filtragem social, segundo (SHARDANAND; MAES, 1995), onde as opiniões dos usuários da mesma corporação do usuário alvo da recomendação, são de fundamental importância no cálculo da recomendação a ser feita. Os usuários da mesma corporação são identificados por seu perfil histórico e similar ao usuário alvo.

3. Método Híbrido: surgiu para aumentar a eficiência dos sistemas de recomendação, técnica esta que utiliza a filtragem colaborativa e por conteúdo. Foram propostos vários métodos híbridos que combinam as duas abordagens, dentre elas algumas se destacam: produzir listas de recomendação de cada abordagem separada, após unir os resultados e produzir uma lista final; e utilizar pesos para os tipos de filtragem, por exemplo, valorizar itens que tem mais acessos.

Para tanto, neste trabalho utilizaremos a filtragem colaborativa baseado em itens.

## 2.2 Recomendação baseada em conteúdo

Nos sistemas de recomendação que se baseiam em filtragem colaborativa (FC), os usuários indicam através de suas avaliações o quanto gostaram de determinados itens. Através destas avaliações o sistema prevê qual será a nota que um usuário dará para um item ainda não avaliado.

A premissa FC é de que as melhores recomendações para um determinado usuário, também para um outro indivíduo, que possui preferência similares a ele. Portanto, a tarefa FC é identificar quais itens os vizinhos do alvo gostaram, e que ainda não foram consumidos por ele.

O processo da filtragem colaborativa pode ser generalizado em três passos segundo (QUEIROZ, 2003):

1. Captura dos dados de entrada: os usuários são observados por suas ações, que são a escolha de itens específicos e avaliações, sejam ela positivas ou negativas. Tudo isto é armazenado no seu perfil, informações estas que são coletadas tanto de forma explicita ou implícita.

2. composição de vizinhança: com os perfis dos usuários formados, é feito uma analise do usuário alvo com relação aos outros perfis e realizada a similaridade entre eles e dessa forma é formado aquilo que é denominado como vizinhança.

3. Geração da recomendação: Por ultimo, tendo os perfis e a vizinhança formada por eles, o sistema recomenda itens para o usuário alvo.

As técnicas de filtragem colaborativa, podem ser classificadas em duas classes. Primeiro, Baseado em modelo (model-based), e segundo, baseado em memória (memory-based) (JÚNIOR, 2017).

Filtragem baseado em modelo, são recomendações que utilizam as avaliações dos usuários e cria um modelo capaz de fazer recomendações. Essa abordagem utiliza-se de aprendizado de máquina, sendo capaz de representar características importantes dos usuários e dos itens.

Filtragem baseado em memória, são recomendações baseada em uma matriz, denominada como matriz usuário-item, com informações das avaliações, completa ou apenas amostral, cada item x usuário, representa a avaliação de um usuário a um determinado item. Esta matriz é mantida em memória. Nesta abordagem, cada usuário é parte de um grupo de usuários com interesses afins, ao identificar essa semelhança, é possível

recomendar novos itens. Essa abordagem é frequentemente usada em sistemas comerciais, por serem eficazes e de fácil implementação.

### 2.2.1 Filtragem Baseada em Usuários

O algoritmo baseado em usuários foi um dos primeiros algoritmos de filtragem colaborativa automático, isso é, que não requer intervenção humana para realizar as recomendações. Foi proposto em 1994, por Resnick et. al. como parte do sistema GroupLens. (ALEIXO et al., 2014)

Na filtragem colaborativa baseadas em usuários, a recomendação de um item para um usuário-alvo é determinada com base nas das avaliações de todos os outros usuários, que já tenham avaliado o item em questão. Os usuários que são levados em consideração, são aqueles que, no histórico, tem avaliações similares ao usuário-alvo, formando assim o que é denominado "vizinhança". Dizemos então, que para todo usuário-alvo encontrado na matriz usuário-item, são considerados vizinhos, onde as avaliações estão mais bem correlacionadas as anteriores feitas pelo usuário-alvo.

Filtros baseadas em usuários foram muito bem sucedidos no passado, porém, ao longo do tempo foram encontrados alguns desafios reais, como o problema da esparsidade (ALEIXO et al., 2014). O grande conjunto de itens para avaliar, principalmente em recomendadores comerciais, os usuários ativos compram menos de um por cento do total de itens, dessa forma, um sistema de recomendação baseado em vizinhança mais próxima, apresenta dificuldade em recomendar itens para um usuário em particular. Outra dificuldade encontrada é a escalabilidade (ALEIXO et al., 2014), pois algoritmos baseados em vizinhança aumentam com o numero de usuários e itens, com milhões de dados, um sistema normal de recomendação, enfrentaria sérios problemas de escala.

### 2.2.2 Filtragem Baseada em Itens

O primeiro algoritmo baseado em Itens foi proposto em 2001, por Linden et. (ALEIXO et al., 2014), onde a ideia utilizada é de que as avaliações aplicadas a itens similares ao item-alvo, que são efetuadas pelo próprio usuário, são mais consistentes para prever a avaliação deste item, do que as avaliações dos usuários similares ao item-alvo.

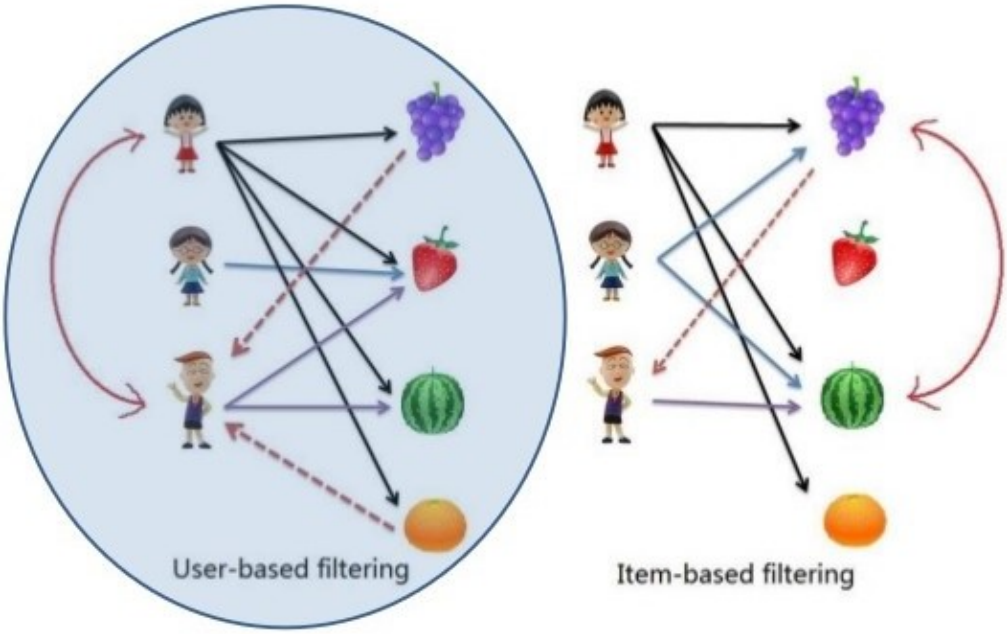
Dizemos então, se um grupo de usuários classificou dois itens da mesma forma, os dois itens, provavelmente são semelhantes. Portanto, se um usuário faz uma boa avaliação de um item, é provável que também irá fazer uma boa avaliação do outro item. Este é o princípio no qual a filtragem baseada em itens se fundamenta.

A Figura 1 demonstra a diferença entre filtragem baseada em usuários e baseada em itens. A esquerda destacada pelo circulo, está a filtragem baseada em usuário (User-based filtering), demonstrando que a moça de vermelho gosta de uva, maça, melancia e laranja.

O rapaz de azul, gosta de maça e melancia da mesma forma que a garota de azul. portanto, conclui-se que ele poderá gostar também de uva e laranja, pois há uma semelhança nas escolhas dos usuários.

A abordagem baseada em itens (Item-based filtering), a direita na figura 1 apresenta o comportamento segundo os itens. A moça de vermelho, novamente gosta de uva, maça, melancia e laranja, a moça de azul também gosta de melancia e uva. O rapaz de azul escolheu a melancia, logo é observado que os outros que escolheram melancia também escolheram uva, portanto, é provável que o garoto também irá gostar dessa fruta.

Figura 1 – Recomendação baseada em conteúdo




Fonte:(PIER, 2018)

Em ambas abordagens, é necessário a construção da vizinhança, que significa um grupo de usuários ou itens que são semelhantes. Para encontrar tal vizinhança, vários métodos de similaridade são utilizados, como a similaridade dos cossenos, que é amplamente utilizada, a correlação de Pearson, a amplificação de peso, distância euclidiana, dentre outros. Cada método tem como objetivo retornar os usuários ou itens semelhantes, formando assim a vizinhança. O tamanho desta vizinhança pode variar de acordo com a técnica, característica e domínio do aplicativo. Porém, o tamanho estabelecido tem impacto na qualidade da recomendação (KIM et al., 2010), se o tamanho da vizinhança for pequeno, é difícil capturar informações necessárias para ter um resultado mais real, enquanto que, se o tamanho for muito grande, enfrenta-se o problema de complexidade computacional,

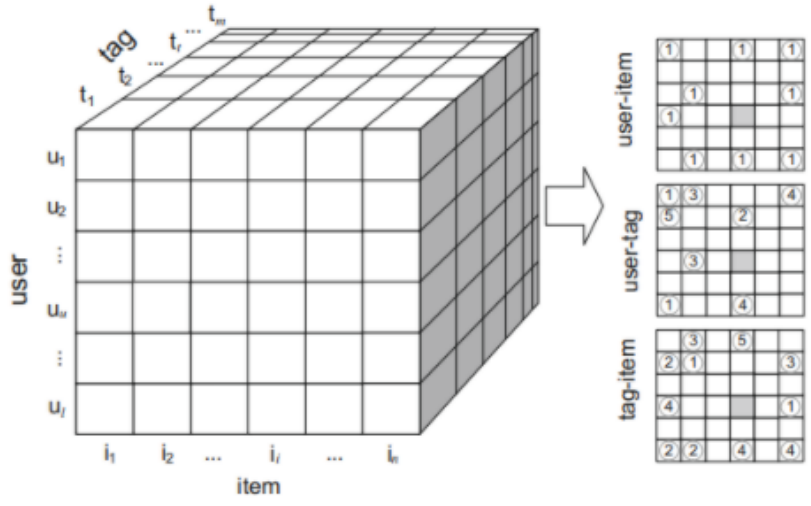
embora os resultados possam ser mais precisos (KIM et al., 2010).

### 2.2.3 Abordagem utilizando "tags"

A recomendação utilizando tags, surge em 2006, quando Hotho et al. Tentou explorar a estrutura das folksonomias para pesquisa e classificação (PENG et al., 2010). Seu método proposto foi o folkRank, uma adaptação do pageRank em sistema de recomendação social, e demonstrou ser capaz de superar várias linhas de base. (PENG et al., 2010) Filtragem colaborativa baseada em itens utilizando "tags", é a prática de permitir que os usuários associem "tags" (ou etiquetas) aos itens, com objetivo de caracterizá-los ou categorizá-los. Em algumas situações, como na Web, onde existem muitos itens para classificar, a marcação colaborativa é uma das maneiras mais úteis de categorizar conteúdo.

Uma das dificuldades enfrentadas pela filtragem colaborativa baseada em itens ou usuários, é a relação com usuários novos, que não avaliaram nenhum item, ou um número de itens insuficiente para criação de um perfil. Da mesma forma os itens novos, que foram avaliados por um número pequeno de usuários, dificultando assim a geração de uma recomendação de qualidade, este problema é conhecido como o problema da partida a frio (cold start problem) (KIM et al., 2010).

Figura 2 – demonstração da decomposição de uma matriz tridimensional para 3 matrizes bidimensional para representação dos usuários, tags e itens




Fonte: (PENG et al., 2010)

A utilização de "tags" é uma possibilidade para melhoramento do problema de dados esparsos, como da partida a frio, pois um sistema de recomendação que leva em consideração não somente os usuários e os itens, mas também as "tags", tem mais informações, e estas mais precisas, para fazer recomendação. A matriz usuário-item, agora

terá uma nova dimensão com as “tags”, tornando assim, uma matriz tridimensional. É possível ajustar essa matriz tridimensional em 3 matrizes bidimensionais, definida como: usuário-item, usuário-tag, tag-item. Como está demonstrado na figura 2, estes podem ser definidos da seguinte forma:

1. Usuário-item: Representa o mapeamento entre os pares de itens dos usuários e suas opiniões acerca deles. As linhas representam os usuários e as colunas os itens. Ou seja, uma posição na matriz indica se o usuário já acessou ou não um determinado item.

2. Usuário-tag: Representa o conjunto de tags e a quantidade de vezes que um determinado usuário fez uso de uma determinada tag.

3. Tag-item: Armazena a quantidade de vezes que um item foi rotulado por uma determinada tag.

## 2.3 K-Vizinhos mais Próximos - kNN

Para aplicação em sistemas reais a filtragem colaborativa necessita de algoritmos classificadores para realizar os cálculos de semelhança entre os itens ou usuários.

O K-vizinho mais próximo é um classificador muito utilizado para recomendação, devido sua fácil implementação e seu desempenho eficaz, contém uma simples interpretação nos resultados e é ideal para uso de banco de dados pequeno e médio.

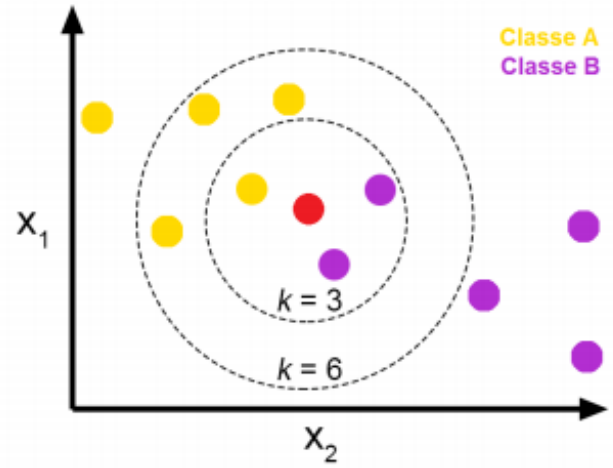
Ele realiza um conjunto de treinamento com vetores n-dimensionais e cada ponto no espaço n-dimensional é representado por cada elemento desse conjunto. Este classificador está entre os mais populares e simples algoritmos de aprendizado de máquina. Em resumo, para determinar a classe de um novo item no conjunto, o kNN busca k itens que estejam mais próximos deste novo elemento, ou seja, que tenha a menor distância.

Este algoritmo necessita de três parâmetros: O valor de K, ou seja, o número de vizinhos e a medida de similaridade, que será usada para calcular e encontras os vizinhos que estão mais proximos (DARMITON, 2020),e database para treinamento.

O primeiro paramênto, O valor K, deve ser diferente de impar e maior que 1, K se refere ao numero de distância para determinar o agrupamento, este é estabelecido de acordo com o proprio usuário. Para estabelecer o valor de K é necessário analisar que valores de K baixos, pode causar um processo de treinamento mais flexivel os quais terão resultados menos enviasado, mas com uma alta variância, porém, se atribuido um valor alto a K, é gerado vários vizinhos,que resultarão em uma fronteira de decisão mais complexa, levando a um resultado com baixa variância, porém um alto risco de enviesamento. O valor de K, deve ser analisado para cada base aplicada. (LUZ, 2019)

Na imagem 3 é demonstrado a classificação com o algorítmo k-NN (JOSé, 2018) onde se K é considerado 3, então o novo item pertence a classe B, pois os vizinhos mais próximos são em sua maioria desta classe. Porém, se considerar k=6, então o novo item pertence a classe A, pois o número de vizinhos aumentou, mudando o resultado, pois na nova vizinhança, contém mais vizinhos da classe A, logo o classificador irá classificar o novo elemento como desta classe A.

Figura 3 – Demonstração da classificação do k-NN




Fonte:(JOSé, 2018)

Para a escolha do valor K, existe várias maneiras de determina-lo de acordo com a tarefa empregada. Uma das alternativas é analisar o algoritmo em um conjunto de validação, adotando vários valores para K. Escolhendo aquele que alcançar melhor resultado nos testes.(DARMITON, 2020)

O segundo parâmetro, a medida de similaridade, muitas tem sido utilizada, as mais conhecidas são: distância euclidiana, similaridade dos cossenos e Correlação de Pearson. As distâncias euclidianas focam na magnitude e, no processo, mas não realizam uma boa medida do grau de similaridade ou dissimilaridade. Para isto, são mais utilizados a correlação de Pearson e a similaridade dos cossenos (BANIK, 2018). Tendo em vista essas descrições, neste trabalho iremos utilizar a similaridade do cosseno.

### 2.3.1 Similaridade do cosseno

A similaridade do cosseno realiza o cálculo do cosseno do ângulo entre dois vetores em um espaço n-dimensional. Se o resultado do cálculo da pontuação for 1, então pode-se concluir que os vetores são exatamente semelhantes, por outro lado, se o resultado for -1, então os vetores são completamente opostos e diferentes (BANIK, 2018). Matematicamente,

a similaridade do cosseno é definida da seguinte forma:

$$\cos(x, y) = \frac{x.y^T}{\|x\| \,.\, \|y\|} \tag{2.1}$$

A pontuação de similaridade do cosseno é mais apropriada para o cenário onde a similaridade é mais importante que a magnitude (BANIK, 2018). Sendo assim, este é mais apropriado na aplicação deste trabalho.

## 2.4 Algoritmos de agrupamentos

Assim como a recomendação de conteúdo se tornou um mecanismo de auxílio aos usuários, os direcionando a conteúdos que com grande probabilidade de ser de seu interesse, os algoritmos de agrupamentos também surgem para automatizar e organizar o grande número de informações disponíveis na rede. O processo de agrupamento é conhecido como clusterização, isto pode ser feito com objetos concretos e abstratos, organizando-os em classe similares.

Cluster é um conjunto de dados que são similares entre si, esta é uma tarefa denominada de agrupamento não-supervisionado, quando não há interferência externa no processo, ou seja, não há um usuário externo (humano), provendo informações para obter as classes. O agrupamento não supervisionado extrai características escondidas dos dados e desenvolve as hipóteses a respeito de sua natureza. (VALE, 2016) E de modo simplificado, agrupamento de dados são técnicas que visam fazer agrupamentos de dados segundo sua semelhança, e isto de forma automática.

### 2.4.1 K-means

K-means é uma heurística que visa minimizar a distância entre os itens de um conjunto de k centros. K se refere ao número de clusters que é definido pelo usuário, visando minimizar o momento de inércia de cada cluster, isto é, diminuir o grau de dificuldade de modificar o estado atual de movimento de um elemento. (HALLIDAY, 2012) O objetivo principal é identificar um agrupamento dos clusters de forma que, dentro de cada classe os dados sejam os mais próximos possíveis, ou seja, quanto mais próximos eles estão, mais parecidos são. A função para esta busca é dada por:

$$d\,(P, \chi) = \frac{1}{n} \sum_{i=1}^{n} d(p_i, \chi)^2$$

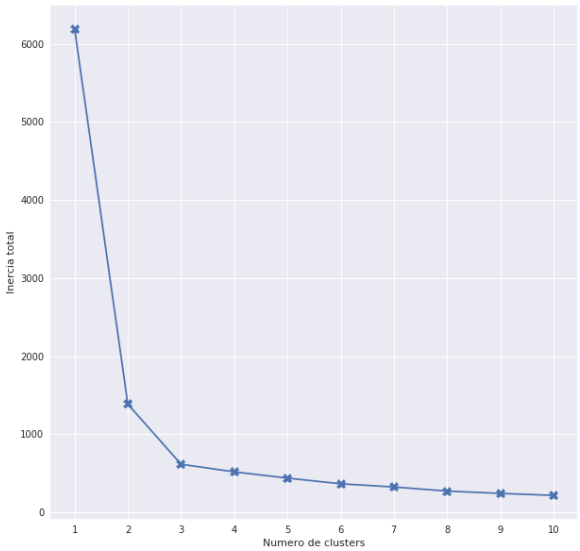
(2.2)

Este algoritmo tem a qualidade de ser rápido em sua execução, além de bem simples, o maior desafio deste método é definir o número k ideal, se colocado um número de clusters muito grande, isto pode fazer com que a diferença entre as classes sejam muito pequenas. Para solucionar este problema, é muito utilizado um método conhecido como método dos cotovelo (do inglês Elbow Method).

#### 2.4.1.1 O método do cotovelo

Este método funciona da seguinte forma: ele irá realizar diversas interações, iniciando com o valor de k=1, calcula a inércia total da melhor solução e então plota um ponto no gráfico onde se tem como coluna o número de clusters e linha de inércia total. Da mesma forma para k=2, calcula inércia e põe um ponto no gráfico, assim se repete até um ponto que em um determinado número k, ao adicionar mais um novo clusters não fará diferença e assim irá deformando um cotovelo no gráfico, daí o nome do método (SAMPAIO, 2018). Desta forma, é encontrado um k ótimo. A imagem 4 demonstra um exemplo da formação da linha do gráfico

Figura 4 – Diminuição da inércia com o crescimento do número de clusters




Fonte: (SAMPAIO, 2018)

Podemos observar no gráfico da imagem 4 que a inércia cai bastante com o um número de 2 e 3 clusters, porém quando se aumenta o número de clusters acima de 3, a inércia já não tem grande modificação. Diante disto, 3 clusters seria o ideal para o exemplo. É essa busca que o método do cutuvelo irá efetuar. Dessa forma, termos os dados necessários para enviar ao k-Means e assim agrupar nossos dados da melhor maneira.

# 3 Metodologia

Um dos maiores desafios da recomendação de conteúdo não é somente buscar itens semelhantes, mas com estes em mãos escolher os melhores para indicar ao usuário alvo o comportamento dos usuários em suas leituras, e as características de cada livro são pontos cruciais para refinar e obter um resultado mais proveitoso. Neste capítulo encontra-se a metodologia usada em todo esse trabalho, tendo em vista solucionar o problema da recomendação. Iniciando pelo pré-processamento da base de dados, a recomendação de conteúdo, agrupamento, extração de características dos grupos e classificação. Primeiramente, foi desenvolvida recomendação de conteúdo usando tags, em seguida, realizado o agrupamento e classificação dos dados.

## 3.1 Base de dados

O conjunto de dados goodbook-10k (ZAJAC, 2017), contém classificações de dez mil livros populares. A base contém 4 arquivos, o primeiro é o ratings onde contém classificações dos usuários para o livro que ele leu, sendo estas de um a cinco, onde todos os usuários fizeram ao menos duas avaliações. O segundo é o books, onde é fornecido as informações dos livros, estes extraídos de arquivos Goodreads, o arquivo contém dez mil itens. O terceiro é o bookTags, onde contém os gêneros atribuídos pelos usuários aos livros, além dos usuários avaliarem os livros que leram, estes também atribuíam novas tags aos livros, e por último as tags, onde contém ids de tags e os nomes destas. Um total de 980.000 avaliações de 53.424 usuários, para 10.000 livros. Na tabela 1 está; um resumo da base de dados.

Tabela 1 – Base de dados utilizada no trabalho

| | **Base de Dados** |
|---|---|
| **Books** | Arquivos que contém todos livros com suas características |
| **Ratings** | classificação dos usuários aos livros |
| **TagsBooks** | tags atribuídas pelos usuários aos livros |
| **Tags** | nome das tags |

As tags que são atribuídas pelos próprios usuários estão listadas na tabela 2 abaixo. Estas por sua vez, são aquelas que obtiveram mais de mil atribuições na base, e, portanto, são as principais.

Tabela 2 – Principais Tags da base

| **Principais tags da base** |
|---|
| 'to-read' 'favorites' 'owned' 'books-i-own' 'currently-reading' 'library' 'owned-books' 'fiction' 'to-buy' 'kindle' 'default' 'ebook' 'my-books' 'audiobook' 'ebooks' 'wish-list' 'my-library' 'audiobooks' 'i-own' 'adult' 'audio' 'favourites' 'novels' 'own-it' 'contemporary' 'read-in-2015' 'series' 'e-book' 'read-in-2016' 'read-in-2014' 'books' 'adult-fiction' 'e-books' 'read-in-2013' 'book-club' 'audible' 'fantasy' 'romance' 'audio-books' 'abandoned' 'novel' 're-read' 'have' 'audio-book' 'mystery' 'borrowed' 'adventure' 'read-in-2012' 'young-adult' 'english' 'did-not-finish' 'favorite' 'maybe' 'shelfari-favorites' 'drama' 'literature' 'general-fiction' 'read-2015' 'ya' 'all-time-favorites' 'classics' 'read-2016' 'read-2014' 'contemporary-fiction' 'favorite-books' 'dnf' 'read-in-2011' 'finished' 'read-in-2017' '5-stars' 'historical-fiction' 'paperback' 'historical' 'thriller' 'sci-fi-fantasy' 'american' 'suspense' 'reviewed' '4-stars' 'unfinished' 'read-2013' 'home-library' 'library-books' 'sci-fi' 'science-fiction' 'action' 'humor' 'family' 'history' 'non-fiction' 'calibre' 'crime' 'didn-t-finish' 'to-read-fiction' 'fantasy-sci-fi' 'nook' 'library-book' 'chick-lit' '20th-century' 'paranormal' 'school' 'classic' 'magic' 'mystery-thriller' 'teen' 'supernatural' 'recommended' 'nonfiction' 'favorite-authors' 'realistic-fiction' 'literary-fiction' 'bookclub' 'want-to-read' 'read-in-2010' 'tbr' 'unread' 'funny' 'bookshelf' 'scifi-fantasy' 'love' 'part-of-a-series' 'books-i-have' 'ya-fiction' 'mystery-suspense' 'kindle-books' 'own-to-read' 'read-2012' 'mysteries' 'must-read' 'need-to-buy' 'urban-fantasy' 'childhood' 'read-in-english' 'children' 'literary' 'horror' 'reread' 'childrens' 'young-adult-fiction' 'thrillers' 'read-2017' '2015-reads' 'on-my-shelf' 'british' '2016-reads' 'coming-of-age' 'children-s' 'kids' 'on-kindle' 'favorite-series' 'science-fiction-fantasy' 'science' 'friendship' 'on-hold' 'listened-to' 'children-s-books' '2014-reads' 'general' 'ya-books' 'my-favorites' 'scifi' 'action-adventure' 'humour' 'biography' 'speculative-fiction' 'contemporary-romance' 'high-school' 'to-read-non-fiction' 'women' 'philosophy' '3-stars' 'war' 'fantasy-scifi' 'mystery-crime' 'usa' 'shelfari-wishlist' 'comedy' 'juvenile' 'first-in-series' 'suspense-thriller' 'stand-alone' 'childrens-books' 'england' 'ya-fantasy' 'crime-mystery' 'kids-books' |

## 3.2 Pré-processamento dos dados

Na exploração dos dados goodbook-10k (ZAJAC, 2017) Foram observados dados duplicados e estes foram retirados. Além disso, os campos nulos e vazios foram substituídos pelo valor zero. O algoritmo K-nn depende de uma base de dados estruturada. Por isso foi necessário realizar modificações. Os dados precisam está no formato de uma matriz de ordem m x n.

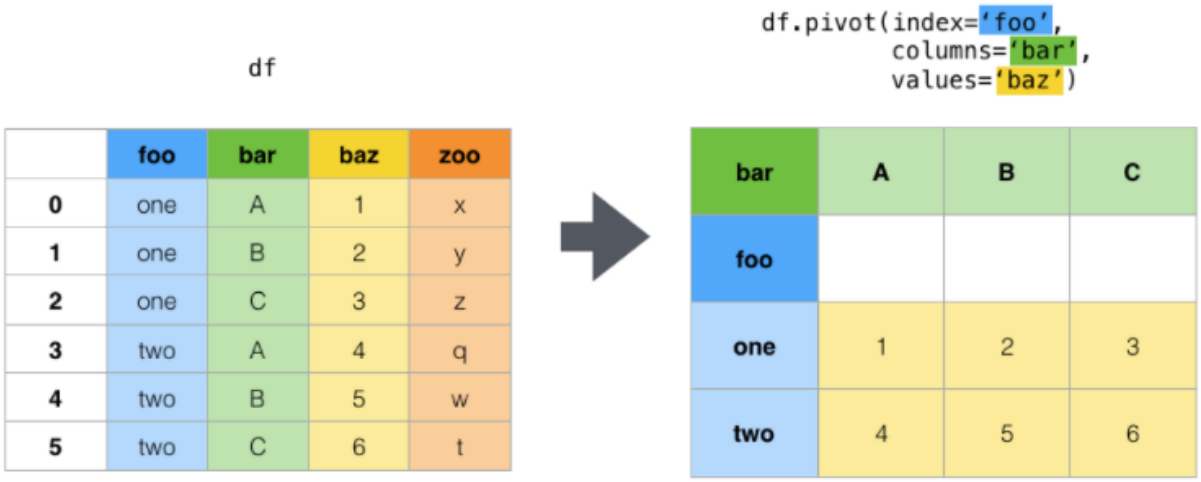
Para a remodelagem foi realizado o pivotamento, onde os índices da tabela são os usuários e os rótulos das colunas são os livros. A cada item da matriz resultante do pivotamento foram atribuídos valores referentes as classificações que cada usuário deu ao livro, ou seja, uma nota variando entre 0 e 4.

### 3.2.1 Pivot

Pivot é uma remodelagem nos valores dos índices e colunas, baseando-se nos valores escolhidos pelo usuário. O pivot recebe três parâmetros: o valor do índex, ou seja, as linhas da nova tabela. O valor para compor a coluna e por último, os valores da tabela. A figura 5 demonstra uma exemplificação de como ocorre esse processo. Para o pivot, foi escolhido os seguintes parâmetros da tabela df: Para índex o 'foo', para coluna o 'bar' e para valor o 'baz', ou seja, os dados da coluna 'baz' na tabela df, se tornará o valor do cruzamento dos

valores de 'foo' e 'bar'.

Figura 5 – Exemplo da execução de pivot sobre uma tabela




| | foo | bar | baz | zoo |
|---|---|---|---|---|
| 0 | one | A | 1 | x |
| 1 | one | B | 2 | y |
| 2 | one | C | 3 | z |
| 3 | two | A | 4 | q |
| 4 | two | B | 5 | w |
| 5 | two | C | 6 | t |

| bar | A | B | C |
|---|---|---|---|
| foo | | | |
| one | 1 | 2 | 3 |
| two | 4 | 5 | 6 |

Fonte: (PANDAS, 2020)

## 3.3 Arquitetura do sistema de recomendação

Para a recomendação de livros ser efetuada, foram necessárias cinco (5) etapas. A figura 6 demonstra os passos da arquitetura desenvolvida.

A arquitetura desenvolvida se divide em duas grandes partes. A primeira é a geração de livros candidatos utilizando as tags e a segunda é a recomendação final dos livros baseado nos conteúdos.

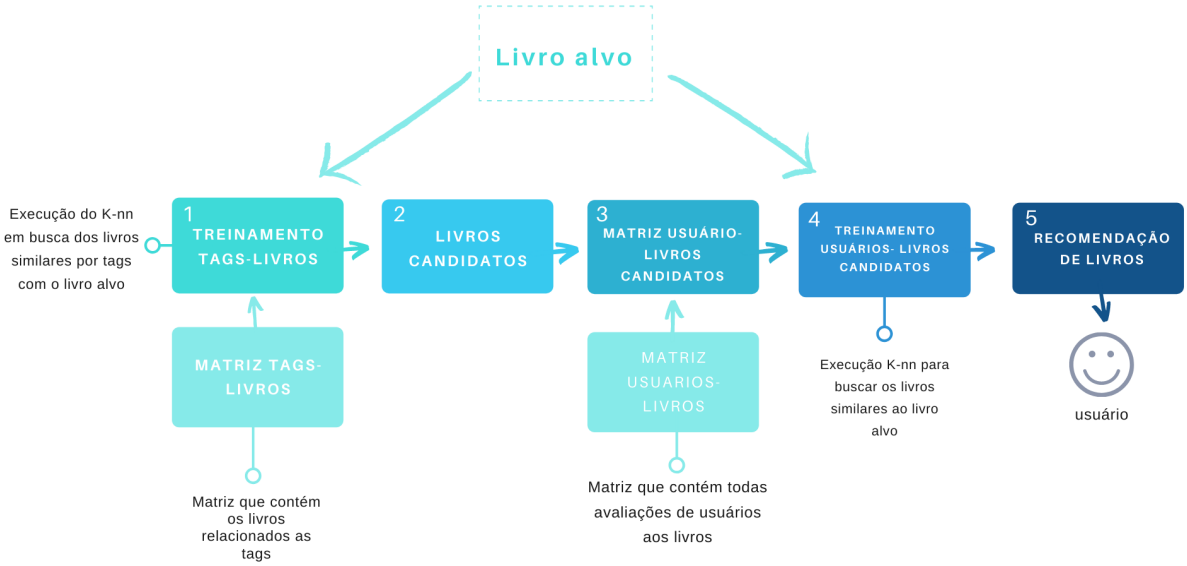
Na primeira parte ocorre a geração dos candidatos, que consiste nas etapas 1 e 2, demonstrado na figura 6: Geração da matriz tags-livros,

treinamento dos dados desta matriz e o resultado obtido, que são os livros candidatos. Na segunda parte, que consiste nos passos 3 ao 5, tem como objetivo efetuar a recomendação efetiva ao usuário em questão. Estes passos serão detalhados nas seções a seguir. O livro alvo inserido nas etapas 1 e 4, é o livro a qual o usuário está lendo no momento, é a partir deste que o sistema irá buscar outros similares.

## 3.4 Geração dos Livros candidatos

Livros candidatos, são aqueles que tem similaridade com o livro alvo, baseado em suas tags, estas por sua vez, atribuídas pelos usuários. É nesta etapa que as tags serão utilizadas, com o objetivo de melhorar o resultado da recomendação, visando contornar o problema da partida a frio, que ocorre quando um livro tem poucas avaliações.

Figura 6 – Arquitetura do sistema de recomendação proposto




Fonte: própria, 2020.

Para a geração destes candidatos, foram realizados três passos, demonstrado na figura 6: Geração da matriz tags-livros, treinamento dos dados desta matriz e o resultado obtido são os livros candidatos.

### 3.4.1 Matriz tags-livros

Para calcular os livros similares ao livro alvo, baseando-se em suas tags, é necessário uma matriz que tenha a relação de todos os livros com as tags atribuídas a eles. Para isto, foi utilizado o arquivo TagsBooks, exposto na seção 3.1

Na matriz de dados, na qual os índices correspondem as tags e as colunas aos livros, foram mapeados todos itens em que há uma relação tag e livro. Tal relação existe quando o respectivo item tem valor maior que 0. E a este item foi atribuído o valor de 1. Este valor é somado, sempre que encontrado uma coluna x e linha relacionada, isto é, quanto maior o número de linhas e colunas relacionadas, mais será contado e acrescentado o número 1. Isso é necessário devido o cálculo da similaridade, em que o item mais semelhante é aquele que possui mais relevância, normalmente com a melhor avaliação para este caso, onde mais vezes é atribuído uma determinada tag.

Abaixo um exemplo do processo. A tabela 3 demonstra alguns livros com os usuários, atribuindo tags aos livros, observe que alguns usuários atribuem as mesmas tags a um determinado livro, isto fará a relevância daquela tag para aquele livro.

Tabela 3 – Exemplo usuários atribuindo tags aos livros

| | **Harry Potter** | **The Closers** |
|---|---|---|
| **Usuário 1** | favorites,magic, fantasy | mystery, fiction, crime, crime-fiction |
| **Usuário 2** | fantasy, currently-reading | currently-reading, young-adult, fiction |
| **Usuário 3** | favorites,classics, fantasy, magic | crime, mystery, favorites |

A tabela 4 demonstra um exemplo de como ocorre o processo da relação e contagem de tags aos livros, levando em consideração os dados da tabela 3, utilizaremos apenas três tags para explicação.

Os três usuários atribuíram a tag 'fantasy' para o livro Harry Potter, portanto, essa tag tem muita relação com esse livro, caso o livro alvo possuir essa tag, esse livro terá grande probabilidade de ser candidato. Porém, observando o livro The Closers, a tag mais relacionada foi a 'mystery', portanto, o mesmo processo ocorreria caso o livro alvo tivesse esta tag.

Tabela 4 – Exemplo de relação quantidade de tags aos livros

| | **fantasy** | **favorites** | **mystery** |
|---|---|---|---|
| **Harry Potter** | 3 | 2 | 0 |
| **The Closers** | 0 | 1 | 2 |

### 3.4.2 Treinamento tags-livros

Nesta etapa ocorre a execução o K-nn, para que seja obtido os livros candidatos, como descrito na seção 2.3. O K-nn precisa de três parâmetros principais: a medida de similaridade, valor de K e a base de dados, esta última que foi demonstrada na seção anterior.

A medida de similaridade utilizada é a similaridade dos cossenos, como descrito na subseção 2.3.1, esta medida é mais apropriada para este trabalho, devido o foco ser a similaridade dos itens.

Para estabelecer o valor de K é necessário analisar que, valores de K baixos causam grande flexibilidade de dados e valores muito altos, causam risco de enviesamento. Por essa razão, o K utilizado nesse sistema é 4.000, um número intermediário, tendo em vista a base de 10.000 itens. O K-nn se comportou de forma adequada com este valor. De maneira mais simplificada, os parâmetros são:

- Valor de K: 4.000
- Medida de similaridade: cosseno
- Base de dados: matriz tags-livros

### 3.4.3 Livros candidatos

Após execução do treinamento foram obtidos os livros que são semelhantes pelo livro alvo. O livro alvo, está demonstrado na tabela 5. A primeira coluna da tabela contém o livro alvo e a segunda coluna lista as tags atribuídas a ele. As tags que estão em destaque, são as principais, isto é, aquelas que tem mais relevância para este livro.

Tabela 5 – Livro alvo do treinamento

| **Livro alvo** | **Tags livro** |
|---|---|
| The Closers | 'science-fiction', **'fiction'**, 'currently-reading' ,'space-opera', 'sci-fi-fantasy', 'aliens','to-read-fiction', 'did-not-finish' 'unread' 'classic-sci-fi',**'suspense'**, 'suspense-thriller','thriller', 'thriller-mystery', **'mystery'**, 'mysteries-thrillers', 'police' |

Para este livro alvo,sugerimos os principais candidatos do treinamento listados na tabela 6, em destaque estão algumas tags, para demonstrar a semelhança baseado em tags.

Tabela 6 – Livros candidatos

| **Livro** | **Tags** |
|---|---|
| Sundiver | **'fiction'**, 'scifi' , 'space-opera', 'favorites', 'aliens', **'mystery'**, 'science', 'space' 'fantasy','sf-fantasy', 'books-i-own','wish-list', 'sciencefiction', 'fantasy-scifi', 'speculative', 'fiction-to-read', 'gave-up-on', 'to-reread' 'future','genre-sci-fi','read-science-fiction |
| Eona: The Last Dragoneye | 'fantasy', 'young-adult', 'dragons','romance' ,'adventure', **'fiction'**, **'action'**, 'supernatural' 'paranormal','historical-fiction', 'maybe' 'sequels','mystery' |
| The Marriage of Opposites | 'currently-reading' 'historical-fiction' **'fiction'**,'romance','adult-fiction', 'france', 'literary-fiction', 'family-relationships', 'family', 'drama','friendship' 'general-fiction', 'fiction-historical', 'biographical-fiction', 'cultural' |

## 3.5 Recomendação de livros

Obtido os livros que são candidatos a recomendação, nessa etapa será efetuado o processo de recomendação baseado em itens. Para isto, serão realizados três passos, demonstrados na figura 6 : Geração da matriz usuários- livros candidatos, treinamento desta matriz e finalmente a recomendação dos livros ao usuário.

### 3.5.1 Matriz usuários-livros candidatos

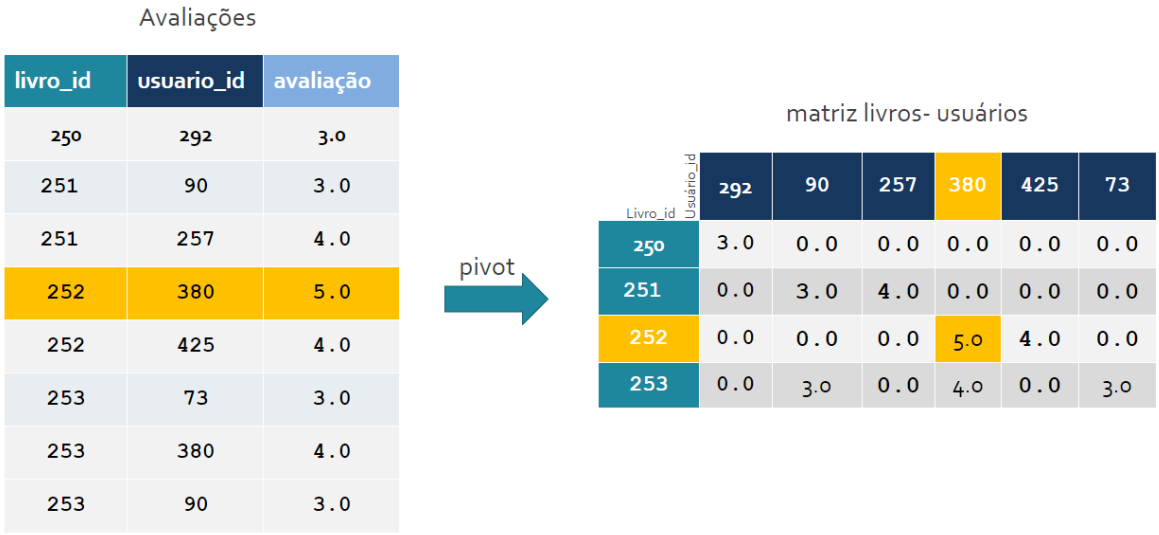
A matriz usuários-livros candidatos, será utilizada para o treinamento que o K-nn realizará para recomendação final do usuário. O arquivo da base utilizado é o 'ratings',

exposto na seção 3.1, nele estão contidas as avaliações dos usuários para os livros que eles leram.

Para realização do treinamento, serão levados em consideração somente as avaliações dos usuários realizadas para os livros candidatos, ou seja, um filtro é realizado na matriz book-usuario, onde obtém-se como resultado a matriz usuários-livros candidatos, que possui todas as avaliações de usuários atribuída aos livros candidatos.

Por fim, é necessário realizar o pivotamento nesse arquivo devido a necessidade do K-nn de receber uma matriz de dimensão M x n. O pivot foi construído da seguinte forma: Como índice é estabelecido o id do livro, para as colunas o id do usuário e para os valores é atribuído as avaliações. A figura 7 demonstra o resultado do pivot no arquivo em questão.

Figura 7 – Demonstração do pivot na tabela de usuário-item




Fonte: Própria, 2020

Observe a linha em destaque na tabela avaliações, em uma única linha se tem o usuário, livro e a avaliação, quando ocorrido o pivot, o livro passa a ser a linha em destaque e o usuário a coluna, que também está em destaque. A avaliação tornou-se o valor que nesse exemplo é uma avaliação com nota máxima 5.0.

### 3.5.2 Treinamento e recomendação dos livros

Nesta etapa, o K-nn é executado para o treinamento da base usuários-livros candidatos. Os parâmetros repassados ao K-nn são:

- O valor de K: 100 , com este valor o K-nn comportou-se de maneira adequada

- Medida de similaridade: Cosseno, tendo em vista o alvo do processo que é a similaridade.
- Base de dados: matriz usuários-livros candidatos

Tendo obtido os livros do treinamento, tem-se a lista em ordem decrescente de similaridade. Os 20 primeiros livros mais próximos/semelhantes, são recomendados ao usuário final.

## 3.6 Análise dos grupos de usuários

Após a recomendação, considerou-se importante analisar o comportamento dos usuários, normalmente estes têm comportamentos repetitivos, o que redunda em um padrão, e analisar esses comportamentos auxilia no aperfeiçoamento dos sistemas de recomendação. Tendo em vista essas questões, foi realizado uma análise do comportamento dos usuários, os organizando em grupos de leituras predominantes.

Para realizar o agrupamento dos usuários fez-se uso do algoritmo K-means. Para este propósito, foi necessário a utilização do método cotovelo, do inglês Elbow Method, tendo em vista a necessidade de conhecer o número de clusters, ou o número de grupos que a base de dados pode agrupar de forma a buscar o melhor resultado e isto de maneira antecipada, antes de executar o K-means. O método do cotovelo, exposto em 2.4.1.1, focaliza-se em encontrar o valor de K ideal para a execução do K-means.

O dataset enviado para o Elbow e o K-means foi o da figura 7. Após obter o número necessário de clusters, executou-se o algoritmo K-means, que é o algoritmo que agrupa de acordo com as características comuns de cada item.

### 3.6.1 Correlação de matrizes

Para obter o comportamento de leitura dos usuários buscando relação entre suas leituras, ou entre os leitores de determinados livros, foram necessários o uso da correlação de matrizes.

Matriz de correlação consiste em encontrar o coeficiente de correlação entre dois elementos. Cada célula da tabela representa um resultado de correlação entre os dois itens. A matriz obtida como resultado é uma matriz que contém um resumo dos dados, dessa forma, é possível encontrar padrões (BOCK, 2018). De modo informal, correlação é sinônimo de interdependência.

Coeficiente de correlação busca identificar o comportamento de um elemento, enquanto um outro está variando com o objetivo de detectar se existe alguma semelhança entre a variabilidade de ambos os elementos. O coeficiente quantifica a relação entre os

elementos. (OLIVEIRA, 2019). Existem diferentes formas de calcular o grau da correlação, mas o tipo mais utilizado é o coeficiente de correlação de Pearson, que se dá pela formula 3.1

$$\rho = \frac{cov(X,Y)}{\sqrt{var(X).var(Y))}} \tag{3.1}$$

onde $\rho$ assume valores entre -1 e 1, sendo 1 a correlação perfeita entre dois elementos, -1 correlação totalmente negativa e 0 significa que um elemento não depende do outro linearmente. (MUKAKA, 2012) Neste trabalho, será utilizado o coeficiente de correlação de Pearson.

### 3.6.2 Distribuições estatísticas

Para melhor visualização dos resultados obtidos pelo K-means e correlação de matrizes, fez-se necessário expor em gráficos estatísticos. Para isto, foi efetivada a distribuição estatística dos dados.

A distribuição estatística é responsável por definir uma curva no gráfico e toda a área que está sob esta curva determinará a probabilidade de ocorrer o evento relacionado a mesma, esta área sob a curva é sempre igual a 1. Existe diversos tipos de distribuição, seja elas continuas ou intercalares, existe uma chamada de distribuição de gauss que normalmente é a mais utilizada, que também é chamada de distribuição normal (JUNIOR, 2012).

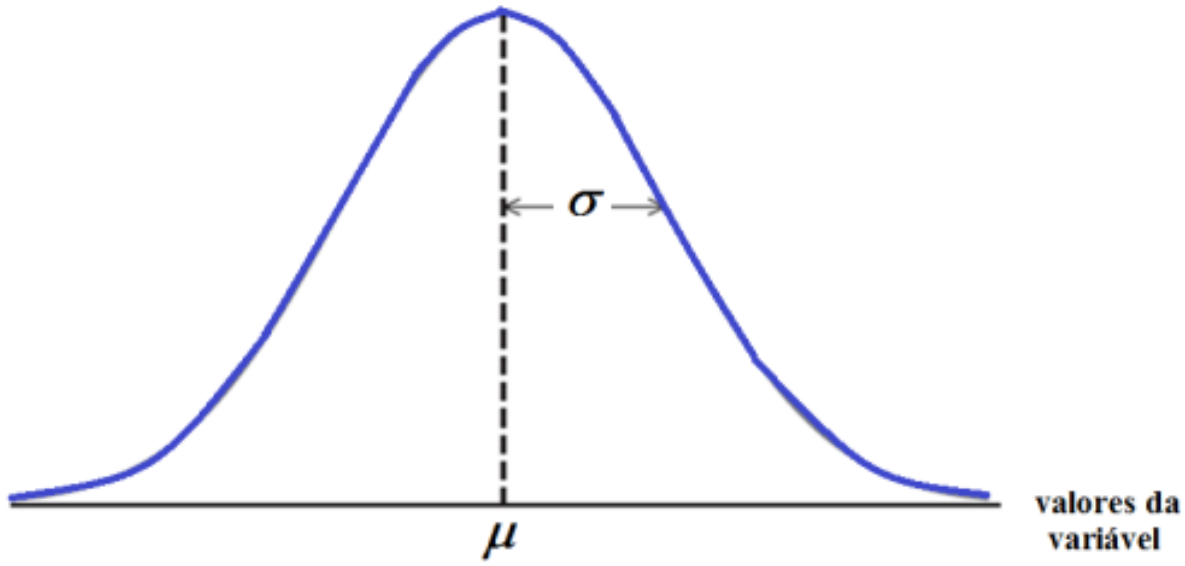
A curva de distribuição normal ou gausseana é normalmente representada por uma curva simétrica que fica em torno do seu ponto médio, formando assim um gráfico que lembra o formato de um sino (RODRIGUES, 2020). O que caracteriza esta distribuição é um acúmulo de valores ao redor de um valor central, simetria em torno deste e uma pequena quantidade de valores muito extremos. O gráfico da imagem 8 demonstra estes comportamentos.

A curva demonstrada na figura 3.2 é definida pela média $\mu$ e pelo desvio padrão $\sigma$, como demonstrado na equação 3.2. Para cada combinação de $\mu$ e $\sigma$ é formado uma curva diferente, a forma mais achatada ou mais alongada é determinado pelo desvio padrão $\sigma$, a média $\mu$ demonstra onde está centralizada a curva gausseana (UFMG, 2020).

$$f(x) = \frac{1}{\sqrt{2\sigma\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} \tag{3.2}$$

Nesse trabalho é utilizado a distribuição normal para encontrar os livros mais lidos.

Figura 8 – Demonstração do gráfico de distribuição normal




Fonte: (UFMG, 2020)

# 4 RESULTADOS E DISCUSSÃO

Neste capítulo são discutidos e expostos os resultados obtidos da metodologia apresentada no capítulo anterior.

## 4.1 Recomendação de livros

O livro utilizado como alvo, isto é, o livro que o usuário está lendo agora, é "The Closers"do autor Michael Connelly, como demonstrado na tabela 5 na seção de treinamento em 3.4.3. Os resultados da recomendação serão discutidos baseados neste livro alvo.

Na tabela 7 está a demonstração do resultado de uma recomendação obtida a partir do processo realizado com o livro alvo citado acima. A tabela expõe os 8 primeiros livros recomendados.

Tabela 7 – Recomendação colaborativa baseada em conteúdo utilizando tags

| **ordem** | **Livros candidatos** | **Livros recomendados** | **Similaridade** |
|---|---|---|---|
| 1 | Sundiver | White Fire | 1 |
| 2 | Eona: The Last Dragoneye | Takedown | 1 |
| 3 | Magyk | The Innocent Mage | 1 |
| 4 | The Song of Achilles | Dodger | 1 |
| 5 | Les Trois Mousquetaires | Bullet | 1 |
| 6 | We the Animals | Whispers Under Ground | 0,98175168 |
| 7 | Heretic | Divergent | 0,976897776 |
| 8 | The Fountains of Paradise | The Vile Village | 0,970559955 |
| 17 | **The Innocent Mage** | | |
| 980 | **Divergent** | | |
| 999 | **The Vile Village** | | |
| 1491 | **Dodger** | | |
| 1193 | **Bullet** | | |
| 2599 | **Takedown** | | |
| 3208 | **Whispers Under Ground** | | |
| 3764 | **White Fire** | | |

Na primeira coluna, está a ordem da recomendação, levando em consideração a similaridade. A segunda coluna da tabela 7 está a lista de livros que foram recomendados pelo treinamento. A terceira coluna exibe a recomendação dos livros efetuada ao usuário e na última coluna está a similaridade dos livros recomendados com o livro alvo, sendo o

máximo 1, que significa que os livros são exatamente iguais e -1 o mínimo, significando que os livros são totalmente diferentes.

O resultado demonstra que os livros recomendados, isto é, aqueles que tiveram boas avaliações em comum com o livro alvo, em sua maioria, não são os mesmo que tiveram tags semelhantes atribuídas nos livros candidatos. Mas a utilização das tags, refinou livros que são muito semelhantes ao livro alvo, auxiliando o treinamento da recomendação de livros a aperfeiçoar sua recomendação, sugerindo livros com número máximo de similaridade, como demonstrado na tabela 7.

Além do mais, se este livro tivesse somente duas avaliações, usando só a recomendação baseada em conteúdo, que leva em consideração somente as avaliações, ocorreria o problema da partida a frio, levando o sistema a recomendar livros que estão bem distante de ser parecido com o livro alvo, tornando a experiência do usuário comprometida. Com as tags, é encontrado outros livros que tem tags que são similares e dessa forma, é encontrado um grupo de livros que podem ser maior interesse do usuário.

## 4.2 Agrupamento de usuários

Diante dos resultados anteriores foi realizado um agrupamento para analisar as características e perfis dos leitores dos livros, utilizando o K-means, algoritmo de agrupamento que avalia e organiza em conjuntos os dados de acordo com suas características. Após executar o método dos cotovelos, que busca o número ideal de clusters para o K-means executar, teve-se como resultado 11 clusters, estes são os grupos nomeados de zero a dez nos gráficos adiante.

Também foi efetivado a execução da correlação gausseana da matriz de usuários x livros e obteve-se oito livros em destaque, estes são os mais lidos dentre todos os demais e são estes que distinguem os grupos de usuários, listados na tabela abaixo 8.

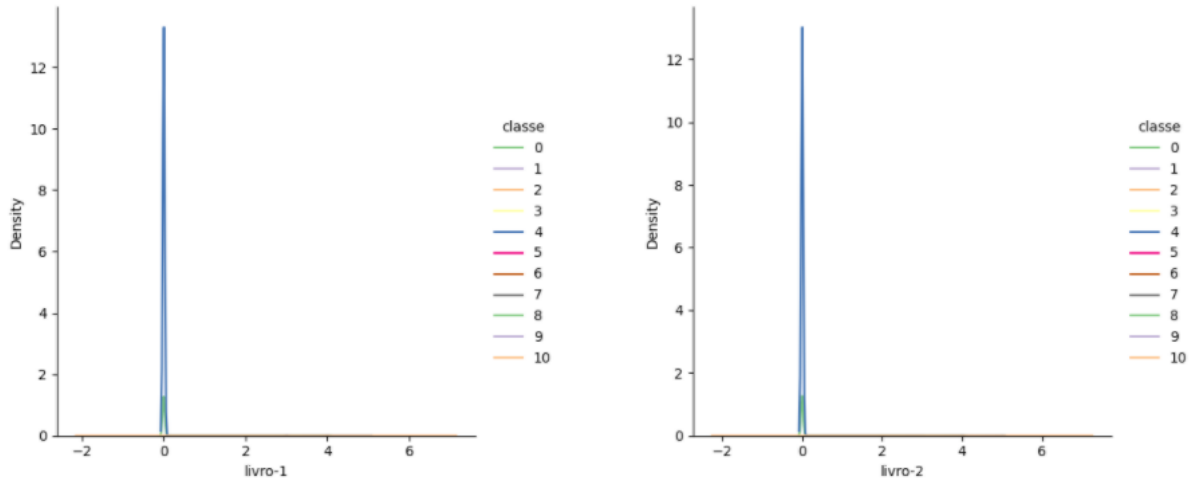
Por conseguinte, foi possível prescrutar que o livro que se destaca entre os leitores é o livro 8 - The Broken Kingdoms e o livro 6 - Intertwined, por suas densidades altas em dois grupos de usuários, estes por sua vez, os maiores entre os encontrados na base disponível. Portanto, podemos observar que os usuários são leitores principalmente de literaturas relacionadas a fiction (do inglês, ficção), mystery (do inglês, mistério) e crime.

Esses livros distinguem os grupos de usuários, após a análise através do K-means obteve-se dez grupos. Nos gráficos das imagens 9, 10, 11 e 12, são demonstrados os comportamentos destes grupos para os livros da tabela 8. Na coluna da tabela está demonstrado a densidade ou a frequência de leitura de cada grupo para o livro em questão. A classe se refere aos grupos. Esses gráficos são distribuições de dados que demonstram como os grupos são estatisticamente distintos.

Tabela 8 – Livros mais lidos

| **Descrição** | **Principais tags** |
|---|---|
| Livro 1 - The Closers | mystery, fiction, crime, crime-fiction, mystery-crime |
| Livro 2 - Such a Pretty Fat: One Narcissist's Quest To Discover | funny, biography, jen-lancaster, fiction, library, comedy |
| Livro 3 - Kill Me If You Can | mystery, fiction, thriller, patterson, suspense, default, crime-mystery, crime |
| Livro 4 - Drop Shot | mystery, currently-reading, harlan-coben, fiction, thriller, crime |
| Livro 5 - The City of Falling Angels | true-crime, book-club, default, owned-books, mystery , historical-fiction, crime, contemporary-fiction |
| Livro 6 - Intertwined | urban-fantasy, ghosts, vampire, owned, library, fiction |
| Livro 7 - The Lemonade War | reading,realistic-fiction, fiction, childrens |
| Livro 8 - The Broken Kingdoms | fantasy, currently-reading, fiction |

Figura 9 – Frequência de leitura do livro 1 e 2

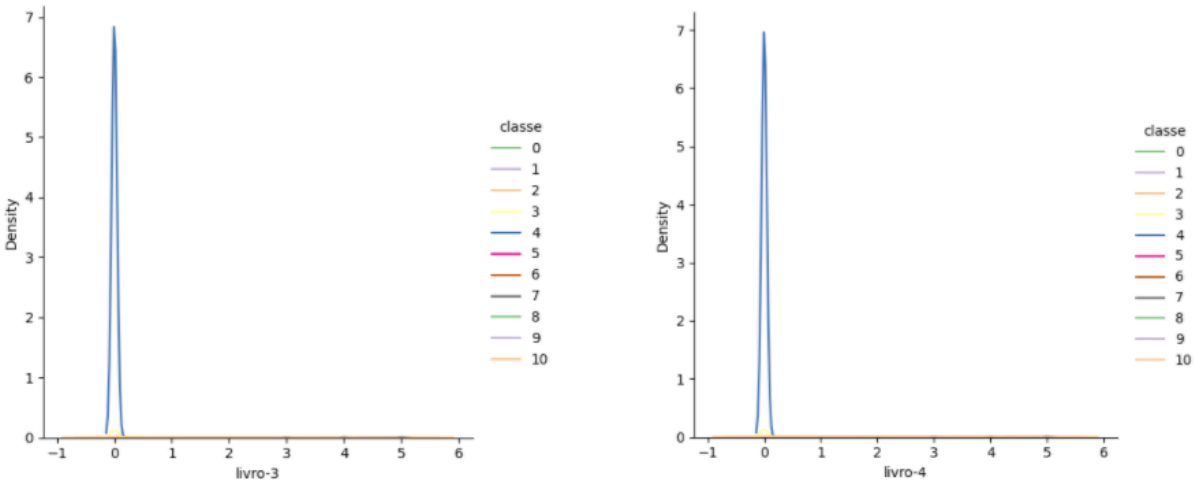



Fonte: Autor

Em vista disso, o grupo que mais se destaca em frequência de leitura é o grupo quatro (4), representado pelas linhas azuis no gráfico, em todos os livros sua densidade de leitura é alta. Seguido dele está o grupo oito (8), que no livro seis (6) e oito (8) se aproximam a densidade igual a dois (2). Estes grupos são principalmente leitores de livros de ficção.

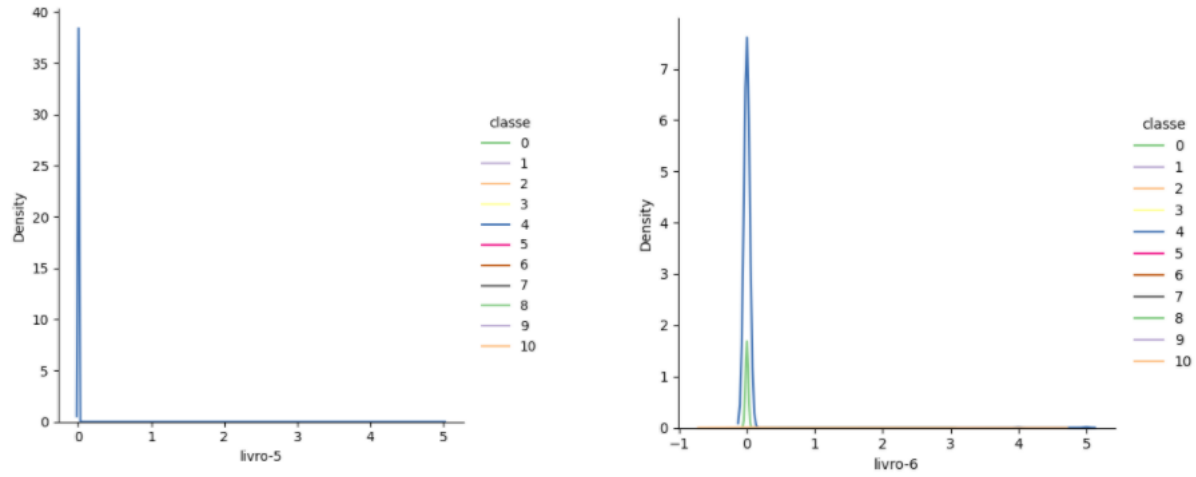
Portanto, mediante a analise realizada, foi possível concluir que nossos usuários tem prioridades por livros mais densos, descartando em sua grande maioria, livros de

Figura 10 – Frequência de leitura do livro 3 e 4




Fonte: Autor

Figura 11 – Frequência de leitura do livro 5 e 6



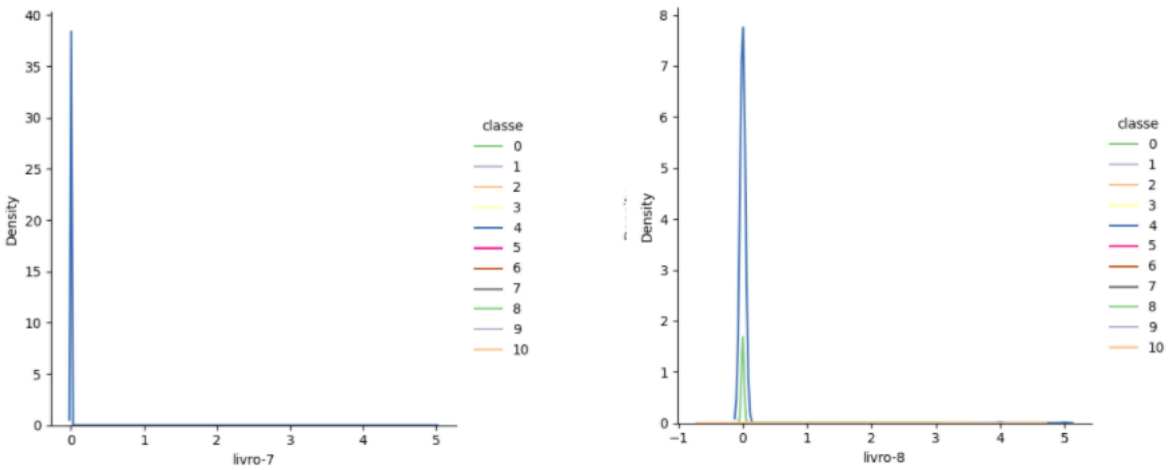

Fonte: Autor

comédias, animações ou romance.

## 4.3 Comparação com trabalhos relacionados

O sistema de recomendação construído nos trabalhos relacionados, (MEDEIROS, 2013), (CAZELLA; NUNES; REATEGUI, 2010), (FERRO et al., 2010) , normalmente se utilizam apenas da abordagem colaborativa, ou seja, segundo os usuários acessam alguns itens. Estes se tornam um grupo semelhantes, dessa forma, se outro usuário novo acessar um destes livros, logo os outros do grupo serão recomendados. Por outro lado,

Figura 12 – Frequência de leitura do livro 7 e 8




Fonte: Autor

algumas propostas anteriores buscam seu melhoramento utilizando tags (recomendação híbrida), como o (OLIVEIRA; COELLO, 2013), usando principalmente matrizes para associar os itens. Em vista disso, os parâmetros estão em uma tabela de difícil manutenção e escalonamento.

Em suma, este trabalho se diferencia por realizar treinamento visando as tags e após este resultado realizar o treinamento da base filtrada para realizar a recomendação dos livros. A utilização dessa estrutura, ajuda a contornar o problema da partida a frio e da matriz esparsa, dando mais assertividade a recomendação. Deste modo, a recomendação tende a aprender e modificar-se de acordo com o comportamento dos usuários, criando padrões e perfis de livros e usuários. Estes últimos por indicarem as tags (etiquetarem os livros) ajudam na eficiência da recomendação. Na tabela 9 abaixo, está a comparação de forma resumida, deste trabalho com os relacionados.

Tabela 9 – Comparação com os trabalhos relacionados

| **trabalhos relacionados** | **Trabalho atual** |
|---|---|
| Uso de classificado somente uma vez | Uso de classificador em todo treinamento da base |
| Uso de tags para comparação de matrizes | Utilização somente de tags para geração de itens candidatos |
| Somente agrupamento de usuários ou recomendação | Recomendação e analise de grupo de usuários |

Além disso, é acrescentado a análise dos grupos de usuários, tornando conhecido suas preferências, sendo assim, fornecendo base para mais um aperfeiçoado na recomendação. Isso é diferente dos demais trabalhos, pois, ou utilizam a recomendação, ou o agrupamento.

Em vista disso, este trabalho busca somar ambos para uma recomendação eficiente aos usuários.

# 5 CONCLUSÃO

A principal motivação para este estudo decorre da necessidade de que os usuários tenham recomendações que de fato sejam interessantes para eles e que irão ajudá-los no processo de decisão, tendo em vista a grande quantidade de conteúdo disponível na rede.

Para tal, no presente trabalho foi desenvolvido o método de recomendação de livros visando maior efetividade em sua sugestão. Utilizando-se do método baseado em conteúdo com a utilização de tags, método este que é pouco utilizado, mas bastante eficaz.

O K-nn foi utilizado para geração de livros candidatos, esta é a etapa que utiliza as tags onde foi gerado 4 mil livros candidatos. O k-nn também foi executado para a etapa da recomendação final, nesta etapa foi realizado o treinamento baseado nas avaliações dadas pelos usuários aos livros candidatos. Com a realização de dois treinamentos o sistema se torna dinâmico de acordo com a mudança de preferencia de usuário.

O sistema realiza o esperado, isto é, a recomendação de uma lista de livros mais semelhante ao livro que o usuário está lendo no momento. As tags auxiliaram de forma adequada na recomendação, ajudando ao usuário receber itens que são muitos semelhantes. A análise e caracterização dos grupos também foi efetuado e concluímos que entre os 11 grupos encontrados, estes são principalmente leitores de livros de ficção. Estes leitores formam grande parte da base utilizada.

Com isto, é possível em trabalhos futuros realizar novos treinamentos na recomendação, fazendo com que estes se modifiquem de acordo com o comportamento dos usuários, realizando eficácia e satisfação aos leitores. Portanto, essa pesquisa se deu com o objetivo de proporcionar uma boa experiência ao usuário, recomendando livros que de fato sejam de seu interesse.

# Referências

ALEIXO, E. L. et al. Item-based-adp: análise e melhoramento do algoritmo de filtragem colaborativa item-based. Universidade Federal de Goiás, 2014. Citado na página 16.

BANIK, R. *Hands-On Recommendation Systems with Python: Start building powerful and personalized, recommendation engines with Python*. [S.l.]: Packt Publishing Ltd, 2018. Citado 2 vezes nas páginas 20 e 21.

BOCK, T. *O que é uma matriz de correlação?* 2018. Disponível em: <https://www.displayr.com/what-is-a-correlation-matrix/#:~:text=A%20correlation%20matrix%20is%20a,a%20diagnostic%20for%20advanced%20analyses.> Acesso em: "01 dezemb. 2020". Citado na página 30.

BONIN, M. *O que são Sistemas de Recomendação? Veja exemplos*. 2018. Disponível em: <https://king.host/blog/2018/09/o-que-sao-sistemas-de-recomendacao/#:~:text=Sistemas%20de%20recomenda%C3%A7%C3%A3o%20s%C3%A3o%20t%C3%A9cnicas,escutar%20ou%20quais%20not%C3%ADcias%20ler.> Acesso em: 02 novemb. 2020. Citado na página 11.

CAZELLA, S. C.; NUNES, M. A. S.; REATEGUI, E. B. A ciencia da opinião: Estado da arte em sistemas de recomendação. capitulo 1. In: *XXX Congresso da Sociedade Brasileira de Computação*. [S.l.: s.n.], 2010. Citado 2 vezes nas páginas 11 e 36.

COSTA, E.; AGUIAR, J.; MAGALHÃES, J. Sistemas de recomendação de recursos educacionais: conceitos, técnicas e aplicações. *Jornada de Atualização em Informática na Educação*, v. 1, n. 1, 2013. Citado na página 11.

DARMITON. *k-vizinhos mais próximos: uma análise*. 2020. Disponível em: <https://maquinasqueaprendem.com/2020/06/22/k-vizinhos-mais-proximos-uma-analise/#:~:text=O%20algoritmo%20k%2Dvizinhos%20mais,frequ%C3%AAncia%20entre%20os%20k%20vizinhos.> Acesso em: 16 dezemb. 2020. Citado 2 vezes nas páginas 19 e 20.

FERRO, M. R. d. C. et al. Modelo de sistema de recomendação de materiais didáticos para ambientes virtuais de aprendizagem. Universidade Federal de Alagoas, 2010. Citado 2 vezes nas páginas 11 e 36.

GARCIA, C. A.; FROZZA, R. Sistema de recomendação de produtos utilizando mineração de dados. *Tecno-Lógica*, v. 17, n. 1, p. 78–90, 2013. Citado na página 11.

HALLIDAY, D. *Fundamentos de Física Volume 1 - Mecânica*. (9ª ed. Rio de Janeiro: LTC - Livros Técnicos e Científicos, 2012. volume 1. ISBN ISBN. Citado na página 21.

HEINZEN, R.; MARTINS, P. D. S. Sistema para identificação de perfil de consumidores utilizando análise de agrupamento (clusterização). *Ciência da Computação-Tubarão*, 2018. Citado na página 12.

JOSé, I. *KNN (K-Nearest Neighbors) 1*. 2018. Disponível em: <https://towardsdatascience.com/knn-k-nearest-neighbors-1-a4707b24bd1d>. Citado na página 20.

JÚNIOR, F. D. H. C. Avaliação de técnicas de filtragem colaborativa para sistemas de recomendação. 2017. Citado na página 15.

JUNIOR, P. D. G. de B. V. *Estatística: Tipos de Distribuição*. 2012. Disponível em: <http://www.cpaqv.org/estatistica/tipos_distribuicao.pdf>. Acesso em: 02 dezemb. 2020. Citado na página 31.

KIM, H.-N.; JI, A.-T.; HA, I.; JO, G.-S. Collaborative filtering based on collaborative tagging for enhancing the quality of recommendation. *Electronic Commerce Research and Applications*, Elsevier, v. 9, n. 1, p. 73–83, 2010. Citado 2 vezes nas páginas 17 e 18.

LIANG, H.; XU, Y.; LI, Y.; NAYAK, R. Tag based collaborative filtering for recommender systems. In: SPRINGER. *International Conference on Rough Sets and Knowledge Technology*. [S.l.], 2009. p. 666–673. Citado na página 12.

LIMA, I. R. R. Recomendação de artigos científicos utilizando filtragens colaborativa e híbrida. 2012. Citado na página 11.

LUZ, F. *ALGORITMO KNN PARA CLASSIFICAÇÃO*. 2019. Disponível em: <https://inferir.com.br/artigos/algoritimo-knn-para-classificacao/>. Citado na página 19.

MEDEIROS, I. *Estudo sobre sistemas de recomendação colaborativos*. [S.l.]: Recife, 2013. Citado 3 vezes nas páginas 11, 14 e 36.

MUKAKA, M. M. A guide to appropriate use of correlation coefficient in medical research. *Malawi medical journal*, v. 24, n. 3, p. 69–71, 2012. Citado na página 31.

OLIVEIRA, A. P. S. de; COELLO, J. M. A. *Desenvolvimento de Algoritmo Híbrido para Sistemas de Recomendação: Filtragem Colaborativa e Etiquetagem Social*. [S.l.]: Campinas, 2013. Citado 2 vezes nas páginas 11 e 37.

OLIVEIRA, B. *Coeficiêntes de correlação*. 2019. Disponível em: <https://operdata.com.br/blog/coeficientes-de-correlacao/>. Citado na página 31.

PANDAS. *Reshaping and pivot tables*. 2020. Disponível em: <https://pandas.pydata.org/pandas-docs/stable/user_guide/reshaping.html#reshaping-by-pivoting-dataframe-objects>. Citado na página 25.

PENG, J.; ZENG, D. D.; ZHAO, H.; WANG, F.-y. Collaborative filtering in social tagging systems based on joint item-tag recommendations. In: *Proceedings of the 19th ACM international conference on Information and knowledge management*. [S.l.: s.n.], 2010. p. 809–818. Citado na página 18.

PIER. *A GENTLE GUIDE TO RECOMMENDER SYSTEMS WITH SURPRISE*. 2018. Disponível em: <https://kerpanic.wordpress.com/2018/03/26/a-gentle-guide-to-recommender-systems-with-surprise/>. Acesso em: 16 dezemb. 2020. Citado na página 17.

QUEIROZ, S. Ricardo de M. *Group recommendation strategies based on collaborative filtering*. Dissertação (Mestrado) — Universidade Federal de Pernambuco, 2003. Citado na página 15.

RICCI F., R. L. S. B.; KANTOR, P. B. Recommender systems handbook. In: *Recommender systems handbook*. [S.l.]: Springer-Verlag New York, Inc., New York, NY, USA, 1st edition., 2010. Citado na página 12.

RODRIGUES, L. *Distribuição normal: o que é, para que serve e como calcular*. 2020. Disponível em: <https://www.voitto.com.br/blog/artigo/distribuicao-normal>. Acesso em: 02 dezemb. 2020. Citado na página 31.

SAMPAIO, I. A. *Aprendizagem ativa em sistemas de filtragem colaborativa*. Dissertação (Mestrado) — Universidade Federal de Pernambuco, 2006. Citado na página 11.

SAMPAIO, P. E. *Entendendo k-Means, agrupando dados e tirando camisas*. 2018. Disponível em: <https://medium.com/@paulo_sampaio/entendendo-k-means-agrupando-dados-e-tirando-camisas-e90ae3157c17>. Citado na página 22.

SANTOS, A. P. d. et al. Sistema de recomendação baseado em agrupamento usando propagação de afinidades. Universidade Federal do Amazonas, 2017. Citado na página 12.

SHARDANAND, U.; MAES, P. Social information filtering: algorithms for automating "word of mouth". In: *Proceedings of the SIGCHI conference on Human factors in computing systems*. [S.l.: s.n.], 1995. p. 210–217. Citado na página 14.

UFMG. *Introdução à Bioestatística*. Universidade Federal de Minas Gerais Instituto de Ciências Exatas Departamento de Estatística, 2020. Disponível em: <http://www.est.ufmg.br/~edna/bionutri/NUT-Aula07.pdf>. Acesso em: 02 dezemb. 2020. Citado 2 vezes nas páginas 31 e 32.

VALE. *Técnicas de agrupamento*. 2016. Disponível em: <https://www.maxwell.vrac.puc-rio.br/14382/14382_4.PDF>. Citado na página 21.

ZAJAC, Z. Goodbooks-10k: a new dataset for book recommendations. *FastML*, FastML, 2017. Citado 2 vezes nas páginas 23 e 24.