UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Explorando Formas de Calibração e Redução do Viés de Popularidade em Sistemas de Recomendação

Rodrigo Ferrari de Souza

Dissertação de Mestrado do Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional (PPG-CCMC)

ICMC USP SÃO CARLOS

Rodrigo Ferrari de Souza

Explorando Formas de Calibração e Redução do Viés de Popularidade em Sistemas de Recomendação

Dissertação apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP, como parte dos requisitos para obtenção do título de Mestre em Ciências – Ciências de Computação e Matemática Computacional. *VERSÃO REVISADA*

Área de Concentração: Ciências de Computação e Matemática Computacional

Orientador: Prof. Dr. Marcelo Garcia Manzato

USP – São Carlos
Junho de 2024

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

S719e

Souza, Rodrigo Ferrari de
Explorando Formas de Calibração e Redução do Viés de Popularidade em Sistemas de Recomendação / Rodrigo Ferrari de Souza; orientador Marcelo Garcia Manzato. -- São Carlos, 2024.
108 p.

Dissertação (Mestrado - Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2024.

1. Sistemas de Recomendação. 2. Calibração. 3. Justiça. 4. Vieses. I. Manzato, Marcelo Garcia, orient. II. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

Rodrigo Ferrari de Souza

Exploring Ways of Calibrating and Reducing Popularity Bias in Recommender Systems

Master dissertation submitted to the Institute of Mathematics and Computer Sciences – ICMC-USP, in partial fulfillment of the requirements for the degree of the Master Program in Computer Science and Computational Mathematics. *FINAL VERSION*

Concentration Area: Computer Science and Computational Mathematics

Advisor: Prof. Dr. Marcelo Garcia Manzato

USP – São Carlos
June 2024

RESUMO

SOUZA, R. F. DE. **Explorando Formas de Calibração e Redução do Viés de Popularidade em Sistemas de Recomendação**. 2024. 108 p. Dissertação (Mestrado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2024.

Em grande parte das aplicações de sistemas de recomendação é importante aumentar o engajamento dos usuários, de modo a apresentar novos conteúdos de seu interesse. Para isso, podem ser utilizados alguns algoritmos de recomendação, como os algoritmos de filtragem colaborativa, que promovem itens similares àqueles que os usuários se interessam, ajudando-os a descobrir novos tipos de conteúdo de que gostam. No entanto, trabalhos recentes mostraram que esse tipo de abordagem apresenta uma conexão entre injustiça, erro de calibração e viés de popularidade nos Sistemas de Recomendação. Ainda que o viés de popularidade promova o consumo de itens mais populares, esse fenômeno também afeta a calibração e justiça das recomendações, onde os gostos de certos usuários não são representados de maneira justa pelo sistema, enquanto outros usuários recebem recomendações consistentes com suas preferências. Nesse sentido, alguns dos trabalhos mais recentes em calibração focam apenas em fornecer recomendações mais justas, não considerando o viés de popularidade que pode amplificar o efeito de cauda longa. Embora outros trabalhos tentem reduzir o impacto do viés de popularidade, não levam em conta o nível de preferência dos usuários por essa característica. Para preencher essa lacuna de pesquisa, o nosso objetivo neste trabalho é estudar formas de calibrar o sistema para trazer recomendações coerentes com as preferências dos usuários e que reduzam o impacto do viés de popularidade. Assim, a proposta é a realização de um estudo sobre abordagens de calibração e de redução do viés de popularidade que tragam recomendações coerentes com os interesses dos usuários de acordo com diferentes níveis de popularidade, sem afetar consideravelmente o nível de satisfação dos usuários com o conteúdo recomendado. Esta pesquisa apresenta contribuições relacionadas à calibração, justiça, experiência do usuário e métricas de avaliação do sistema.

Palavras-chave: Sistemas de Recomendação, Vieses, Calibração, Justiça.

ABSTRACT

SOUZA, R. F. DE. **Exploring Ways of Calibrating and Reducing Popularity Bias in Recommender Systems**. 2024. 108 p. Dissertação (Mestrado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2024.

In most recommender systems applications, it is essential to increase user engagement to present new content of interest to them. For this, some recommendation algorithms can be used, such as collaborative filtering, which promotes items similar to those users are interested in, helping them discover new types of content they like. However, recent works have shown that this approach connects unfairness, calibration error, and popularity bias in Recommender Systems. While popularity bias promotes consumption of more popular items, this phenomenon also affects the calibration and fairness of recommendations, where the system does not fairly represent the interests of particular users. In contrast, other users receive recommendations consistent with their preferences. In this sense, some cutting-edge work in calibration only focuses on providing fairer recommendations to users, not considering the popularity bias that can amplify the long-tail effect. Furthermore, although other works try to reduce the impact of popularity bias, they do not consider users' preference level for this feature. To address this research gap, our objective in this study is to investigate methods of calibrating the system to provide recommendations that align with user preferences and mitigate the impact of popularity bias. The proposal involves conducting a study on calibration approaches and bias reduction strategies of popularity that yield recommendations consistent with users' interests across different levels of popularity, without significantly affecting users' satisfaction levels with the recommended content. This research yields insights and contributions pertaining to system calibration, fairness, user experience, and system evaluation metrics.

Keywords: Recommender Systems, Biases, Calibration, Fairness.

LISTA DE ILUSTRAÇÕES

Figura 1 – Estrutura de calibração proposta. A calibração por popularidade é aplicada de forma combinada a execução do BPR, resultando em uma lista calibrada de recomendações de acordo com as preferências do usuário sobre popularidade e gêneros. 45
Figura 2 – Curva representando a divisão dos itens em grupos de popularidade. 46
Figura 3 – O quadro à esquerda representa os dados observados. A abordagem cria uma relação par de itens específica para o usuário $i \succ_u j$ entre dois itens. No lado direito da tabela, o sinal de mais indica que o usuário u está mais interessado no item i do que no item j; o sinal de menos indica que o usuário prefere o item j ao i; o ponto de interrogação indica que não se pode inferir nenhuma conclusão entre os itens. 48
Figura 4 – Curva de Lorenz que indica a distribuição de popularidade nas duas versões do experimento. 57
Figura 5 – Fluxo da Abordagem de Calibração Personalizada 70
Figura 6 – Estrutura de calibração proposta. A saída da calibração de popularidade é a entrada para calibração de gênero, resultando em uma lista calibrada de recomendações de acordo com as preferências do usuário sobre popularidade e gêneros. 78
Figura 7 – Tela de seleção de gêneros preferidos do usuário. 101
Figura 8 – Tela de livros recomendados para o usuário com *nudge*. 102
Figura 9 – Tela de livros recomendados para o usuário sem *nudge*. 102
Figura 10 – Tela de detalhes do livro. 103
Figura 11 – Tela de seleção de filmes para avaliação e construção do perfil do usuário. . 108
Figura 12 – Tela de filmes recomendados para o usuário. 108

LISTA DE TABELAS

Tabela 1 – Comparação entre os trabalhos relacionados e a presente proposta de pesquisa 40
Tabela 1 – Comparação entre os trabalhos relacionados e a presente proposta de pesquisa 41
Tabela 1 – Comparação entre os trabalhos relacionados e a presente proposta de pesquisa 42
Tabela 2 – Estatísticas dos conjuntos de dados após realização do pré-processamento. . 50
Tabela 3 – Comparação da abordagem proposta com os outros trabalhos no conjunto de dados Yahoo Movies. O símbolo ▲ significa que a proposta teve um ganho significativo com relação aos outros trabalhos, com um p-*value* < 0.05 usando o *t-test* de Student; O símbolo • significa que não houve um ganho ou perda significativo; e o símbolo ▼ indica que o outro trabalho é estatisticamente melhor que a proposta. Cada par de símbolos se refere ao BPR e ao *PairWise*, respectivamente. 51
Tabela 4 – Comparação da abordagem proposta com os outros trabalhos no conjunto de dados Movie Lens 20M. O símbolo ▲ significa que a proposta teve um ganho significativo com relação aos outros trabalhos, com um p-*value* < 0.05 usando o *t-test* de Student; O símbolo • significa que não houve um ganho ou perda significativo; e o símbolo ▼ indica que o outro trabalho é estatisticamente melhor que a proposta. Cada par de símbolos se refere ao BPR e ao *PairWise*, respectivamente. 52
Tabela 5 – Número absoluto de livros favoritados pelos usuários nos grupos de controle e tratamento. 58
Tabela 6 – Número de usuários que colocaram ao menos um item de gênero diferente de suas preferências na lista de favoritos. 58
Tabela 7 – Porcentagem dos itens selecionados a partir da calibração pelos grupos de tratamento. 65
Tabela 8 – Estatísticas dos conjuntos de dados após realização do pré-processamento. . 72
Tabela 9 – Comparação das abordagens de calibração propostas com os outros trabalhos usando o conjunto de dados do Yahoo Movies. O símbolo ▲ significa uma melhoria estatisticamente significativa da abordagem proposta em comparação com os outros trabalhos, com um valor $p < 0,05$ usando o teste t de Student; o símbolo • não denota nenhum ganho ou perda estatisticamente significativo; e o símbolo ▼ indica que o trabalho da literatura é estatisticamente melhor que a proposta. Cada par de símbolos está relacionado aos trabalhos CP e Calibração por gêneros, respectivamente. 74

Tabela 10 – Comparação das abordagens de calibração propostas com os outros trabalhos usando o conjunto de dados do MovieLens 20M. O símbolo ▲ significa uma melhoria estatisticamente significativa da abordagem proposta em comparação com os outros trabalhos, com um valor $p < 0,05$ usando o teste t de Student; o símbolo • não denota nenhum ganho ou perda estatisticamente significativo; e o símbolo ▼ indica que o trabalho da literatura é estatisticamente melhor que a proposta. Cada par de símbolos está relacionado aos trabalhos CP e Calibração por gêneros, respectivamente. 75
Tabela 11 – Comparação das abordagens de calibração propostas com os outros trabalhos usando o conjunto de dados do Yahoo Songs. O símbolo ▲ significa uma melhoria estatisticamente significativa da abordagem proposta em comparação com os outros trabalhos, com um valor $p < 0,05$ usando o teste t de Student; o símbolo • não denota nenhum ganho ou perda estatisticamente significativo; e o símbolo ▼ indica que o trabalho da literatura é estatisticamente melhor que a proposta. Cada par de símbolos está relacionado aos trabalhos CP e Calibração por gêneros, respectivamente. 76
Tabela 12 – Estatísticas dos conjuntos de dados pré-processados. 80
Tabela 13 – Comparação dos algoritmos implementados e dos valores de *trade-off* definidos para cada um. 82
Tabela 14 – Comparação do método proposto com os métodos de outros trabalhos da literatura no conjunto de dados MovieLens 20M. Os melhores valores obtidos para os algoritmos de recomendação SVD++, NMF e VAE estão em negrito. Os resultados comparando a estrutura de calibração proposta com outros métodos são estatisticamente significativos. 83
Tabela 15 – Comparação da calibração proposta com os outros trabalhos da literatura no conjunto de dados do Yahoo Movies. Os melhores valores obtidos para os algoritmos de recomendação SVD++, NMF e VAE estão em negrito. Os resultados da comparação dos métodos propostos com outros métodos são estatisticamente significativos. 84
Tabela 16 – Comparação de todas as abordagens propostas neste trabalho no conjunto de dados do Yahoo Movies. Os melhores valores obtidos para cada métrica estão em negrito. Os resultados da comparação são estatisticamente significativos. 88
Tabela 17 – Comparação de todas as abordagens propostas neste trabalho no conjunto de dados do Movie Lens 20M. Os melhores valores obtidos para cada métrica estão em negrito. Os resultados da comparação são estatisticamente significativos. 88
Tabela 18 – Itens do Questionário Pós-Tarefa: Percepções de Qualidade e Justiça 105

SUMÁRIO

1 INTRODUÇÃO 17
1.1 Motivação 18
1.2 Objetivos e Questões de Pesquisa 20
1.3 Organização do Trabalho 20

2 CONCEITOS FUNDAMENTAIS 23
2.1 Sistemas de Recomendação 23
2.2 Abordagens Clássicas 24
2.3 Avaliação dos Sistemas de Recomendação 26
2.3.1 Formas para Avaliação dos Sistemas de Recomendação 26
2.3.2 Características dos Sistemas de Recomendação 26
2.3.3 Métricas para Avaliação dos Sistemas de Recomendação 27
2.4 Vieses 29
2.5 Justiça e Calibração 31
2.6 Considerações Finais 32

3 TRABALHOS RELACIONADOS 35
3.1 O Viés de Popularidade e Seu Impacto 35
3.2 Calibração de Sistemas de Recomendação 38
3.2.1 Calibração em Etapa de Pré-Processamento 38
3.2.2 Calibração em Etapa de Processamento 38
3.2.3 Calibração em Etapa de Pós-Processamento 39
3.3 Considerações Finais 40

4 UMA ABORDAGEM DE CALIBRAÇÃO NA FASE DE PROCESSAMENTO 43
4.1 Justificativa 43
4.2 Metodologia 44
4.2.1 Calibração por Popularidade 44
4.2.2 O Método BPR 47
4.2.3 BPR com Calibração por Popularidade 49
4.2.4 Configuração do Experimento 50
4.3 Resultados 51
4.3.1 Yahoo Movies 51

4.3.2 *Movie Lens 20M* *52*
4.4 Considerações Finais 53

5 UMA ABORDAGEM *NUDGE* 55
5.1 Justificativa 55
5.2 Metodologia 56
5.3 Resultados 57
5.4 Considerações Finais 58

6 UMA ABORDAGEM DE CALIBRAÇÃO EM PÓS-PROCESSAMENTO 61
6.1 Justificativa 61
6.2 Metodologia 62
6.3 Resultados 65
6.4 Considerações Finais 66

7 UMA ABORDAGEM DE CALIBRAÇÃO PERSONALIZADA 69
7.1 Justificativa 69
7.2 Metodologia 70
7.3 Resultados 74
7.4 Considerações Finais 76

8 UMA ABORDAGEM DE CALIBRAÇÃO DUPLA 77
8.1 Justificativa 77
8.2 Metodologia 78
8.3 Resultados 82
8.4 Considerações Finais 85

9 CONCLUSÃO 87
9.1 Contribuições 87
9.2 Comparação entre as Abordagens 87
9.3 Publicações 89
9.4 Limitações 92
9.5 Trabalhos Futuros 92

REFERÊNCIAS 93

APÊNDICE A IMAGENS DAS TELAS DO EXPERIMENTO COM USUÁRIOS DA ABORDAGEM *NUDGE* 101

APÊNDICE B QUESTÕES DE PESQUISA DO EXPERIMENTO DE CALIBRAÇÃO EM PÓS-PROCESSAMENTO 105

APÊNDICE C IMAGENS DAS TELAS DO EXPERIMENTO DE CALIBRAÇÃO EM PÓS-PROCESSAMENTO 107

CAPÍTULO

1

INTRODUÇÃO

Os sistemas de recomendação fazem parte da rotina das pessoas em todo mundo, influenciando as decisões que elas tomam ao acessarem serviços *online* com a apresentação de conteúdos específicos para cada usuário. Consequentemente, estes sistemas são cada vez mais predominantes nos diversos contextos atuais, incluindo comércio eletrônico, vídeos, música e muito mais. Como resultado da sua ampla adoção, os sistemas de recomendação tornaram-se um tema altamente relevante tanto na indústria como na academia (DELDJOO *et al.*, 2023). Dessa maneira, os sistemas são desenvolvidos com o objetivo de aumentar o engajamento dos usuários apresentando a eles conteúdos de seus interesses.

Para atingir esse objetivo, os sistemas constroem o perfil do usuário, que geralmente é composto por tópicos de seu interesse ou classificações previamente atribuídas aos itens consumidos. Nesse sentido, existem três abordagens utilizadas para obter essas informações do perfil do usuário: implícita, explícita ou híbrida (BALTRUNAS; AMATRIAIN, 2009). Na abordagem implícita, as interações do usuário com o sistema servem de informação sobre as suas preferências. No caso da abordagem explícita, o próprio usuário deve fornecer seus interesses, o que pode ser feito a partir de um formulário. Já na híbrida, ocorre uma mistura das duas abordagens anteriores.

Além dessa fase inicial de obtenção de dados, alguns algoritmos, como os de filtragem colaborativa, podem ser utilizados para promover itens que usuários com preferências parecidas ao do usuário alvo também gostaram. No entanto, os sistemas que usam esse tipo de algoritmo sofrem certas limitações como erros de calibração, imparcialidade e vieses, como o viés de popularidade. Esse tipo de viés acaba por promover itens mais populares em detrimento de itens menos populares, fato que torna o sistema injusto (KOWALD; SCHEDL; LEX, 2020).

A injustiça em recomendação é uma questão que vem sendo estudada em trabalhos recentes, sendo caracterizada quando um sistema favorece certos tipos de itens em detrimento de outros, seja por conta de aspectos como gênero, popularidade, idade, etc (DELDJOO *et al.*,

2023).

Há também outros vieses conhecidos que afetam as recomendações dos sistemas, como os vieses de posição, conformidade, exposição e confirmação (CHEN *et al.*, 2023a). Embora os sistemas de recomendação tenham avançado nos últimos anos, ainda precisam lidar com grandes desafios, como a injustiça e os vieses citados, e aspectos como bolhas de filtro, que é quando o usuário fica preso em recomendações somente de informações de seu interesse, o que impede de ver itens que discordem de suas visões.

Esses desafios estão diretamente ligados ao fato de que os sistemas de recomendação procuram aumentar a acurácia das recomendações de forma a aumentar a satisfação dos usuários recomendando itens de seu interesse para que utilizem o sistema por mais tempo. Assim, sistemas com uma acurácia alta tendem a prender o usuário em bolhas de filtro (KHENISSI; NASRAOUI, 2020), já que deixam de recomendar itens diferentes de suas preferências, o que diminui a diversidade do sistema. Ademais, aspectos como erros de calibração geram recomendações incoerentes com suas preferências.

Para lidar com essas limitações, os sistemas de recomendação podem ser calibrados de modo a entregar ao usuário proporções de itens de suas áreas de interesse coerentes com suas preferências (STECK, 2018). Em seu trabalho, Steck (STECK, 2018) apresenta uma abordagem que retorna itens com gêneros nas mesmas proporções das preferências dos usuários, ou seja, se o indivíduo tem uma preferência de proporção de 60% por filmes do gênero de ação e 40% por filmes de comédia, o sistema irá retornar recomendações de filmes que respeitem essa proporção.

Dessa forma, quando um sistema apresenta recomendações com itens com aspectos diferentes do interesse do usuário, é considerado um sistema mal calibrado, e quando os usuários lidam com diferentes níveis de calibração, é dito que o sistema é injusto a um grupo de usuários (ABDOLLAHPOURI *et al.*, 2020). Além disso, quando um sistema traz recomendações que atendam as preferências dos usuários é dito que é um sistema preciso, já que oferece satisfação ao usuário por recomendar itens que lhe agradam.

1.1 Motivação

Embora presentes nos mais diversos sistemas de mídia, os sistemas de recomendação ainda enfrentam desafios como vieses, injustiça e bolhas de filtro que reduzem a eficácia de suas recomendações. A injustiça presente nos sistemas de recomendação decorre de vários tipos de vieses que ocorrem naturalmente nos dados. Um desses vieses é o viés de popularidade, que tende a favorecer a recomendação de itens altamente populares em detrimento de itens menos conhecidos (CHEN *et al.*, 2023a).

Neste contexto, a noção de justiça dentro de um sistema de recomendação está ligada à sua capacidade de fornecer recomendações que se alinhem de forma consistente com as

preferências de todos os utilizadores. Portanto, é importante medir a capacidade do sistema de oferecer recomendações apropriadas a qualquer usuário, e uma abordagem promissora é empregar a calibração (STECK, 2018). Um tipo específico de calibração é quando se utiliza metadados de itens (por exemplo, gêneros) para fornecer um ranking de recomendação cuja distribuição de categorias esteja alinhada com a distribuição desses tópicos no perfil do usuário.

Em vista disso, o trabalho (SILVA; MANZATO; DURÃO, 2021) propõe um sistema que funciona em etapa de pós-processamento e gera recomendações conforme o nível de preferência dos usuários pelos gêneros dos itens, de forma a equilibrar a acurácia das recomendações. Da mesma forma, o trabalho de (STECK, 2018) também apresenta uma calibração das recomendações com o objetivo de retornar proporções de itens coerentes com as preferências dos usuários pelos gêneros dos itens.

O trabalho (GEYIK; AMBLER; KENTHAPADI, 2019), por sua vez, apresenta uma reclassificação das recomendações com o objetivo de reduzir os vieses do sistema e trazer recomendações mais justas no aspecto dos gêneros e da idade dos usuários candidatos a vagas de emprego. Apesar de apresentarem bons resultados e diminuírem a injustiça em suas recomendações, os trabalhos ainda não levam em conta o impacto do viés de popularidade em suas recomendações.

Já o trabalho (INGESSON, 2022) faz uma análise da percepção dos usuários sobre o aspecto de justiça em um sistema no domínio de música. Na proposta, são utilizados três algoritmos diferentes para retornar as recomendações, cada um deles com um nível de popularidade diferente, tendo como resultado o fato de que os usuários não notaram diferença na justiça entre os sistemas. Apesar do resultado desfavorecer o aspecto de popularidade na visão dos usuários, os sistemas avaliados não levam em conta o nível de interesse dos usuários por esse aspecto, podendo ser um caso em que não haviam usuários recrutados para o experimento que prefiram itens de nicho, sendo esse tipo de usuários os que costumam ser mais afetados pelo viés de popularidade.

Há ainda alguns trabalhos relacionados com a redução do impacto do viés de popularidade no sistema. O artigo (ABDOLLAHPOURI; BURKE; MOBASHER, 2019) propõe uma reclassificação dos itens utilizando o histórico de interação dos usuários com itens da cauda longa, dessa forma é possível medir o interesse dos usuários por itens de nicho. Apesar de utilizarem o aspecto de popularidade dos itens no processo de reclassificação, o sistema proposto não recomenda itens numa proporção consistente com as preferências dos usuários de forma calibrada. Outros trabalhos (YALCIN, 2021; LIN *et al.*, 2022; CHEN *et al.*, 2022; NAGHIAEI; RAHMANI; DEHGHAN, 2022; LESOTA *et al.*, 2021; YALCIN; BILGE, 2022) concentram-se apenas na redução do viés de popularidade, mas não abordam a justiça em relação à distribuição dos itens recomendados.

Sendo assim, há uma lacuna no estado da arte com relação a retornar recomendações que reduzam o impacto do viés de popularidade, tragam itens calibrados e que funcione de forma

agnóstica ao modelo de recomendação.

1.2 Objetivos e Questões de Pesquisa

Considerando o contexto apresentado, foram levantadas as seguintes questões de pesquisa:

- **RQ-1: Como as abordagens impactam a geração das recomendações?** A questão visa compreender como as abordagens afetam as recomendações geradas pelo sistema. Dada a relação entre justiça e relevância, as recomendações calibradas com justiça podem impactar negativamente a precisão do sistema.
- **RQ-2: De que forma as abordagens contribuem para a redução do viés de popularidade do sistema?** Enquanto a RQ-1 foca na geração das recomendações, essa questão de pesquisa foca exclusivamente no impacto que a abordagem possui diante do viés de popularidade. O objetivo é validar se a abordagem consegue contribuir para a redução do viés de popularidade.

Para validar as questões propostas, serão conduzidos experimentos para medir o efeito da calibração do sistema com base no aspecto de popularidade. Através desses experimentos, buscamos não apenas compreender como as abordagens afetam a geração e a justiça das recomendações (**RQ-1**), mas também avaliar se essas abordagens podem contribuir significativamente para a redução do viés de popularidade no sistema (**RQ-2**).

Assim, o objetivo do projeto é investigar como o sistema de recomendação pode ser calibrado de maneira a gerar recomendações que atendam aos diferentes níveis de preferências dos usuários. Para isso, pretende-se utilizar técnicas de calibração e a combinação delas com técnicas de redução do impacto do viés de popularidade, como a reclassificação das recomendações, a fim de gerar recomendações coerentes com as preferências dos usuários.

1.3 Organização do Trabalho

A estrutura desta monografia é a seguinte:

- No Capítulo 2, são discutidos os conceitos fundamentais sobre sistema de recomendação, vieses e calibração.
- No Capítulo 3, são apresentados os trabalhos relacionados.
- O Capítulo 4 detalha a primeira abordagem de calibração estudada neste trabalho. É uma abordagem que funciona em etapa de processamento e é uma modificação do algoritmo

BPR *(Bayesian Personalized Ranking from Implicit Feedback)* (RENDLE *et al.*, 2012). Essa abordagem funcionou como um estudo inicial e possibilitou o estudo de outras abordagens, já que trouxe conhecimento de abordagens em etapas de pós-processamento e também de um mecanismo de interface utilizado para guiar o comportamento dos usuários conhecido como *nudge*, que será detalhado no decorrer deste trabalho.

- O Capítulo 5 apresenta a proposta de redução do viés de popularidade por meio de *nudges* e os seus resultados. Esse estudo foi realizado em nível de interface por meio de um mecanismo que recomenda itens diferentes do perfil do usuário.

- O Capítulo 6 apresenta uma proposta de calibração em etapa de pós-processamento e os resultados após o experimento anterior, realizado com usuários. O intuito deste experimento foi verificar como o usuário percebe os itens calibrados. As conclusões do experimento permitiram validar que os usuários consideraram relevantes os itens retornados pela calibração e isso serviu de incentivo para estudar duas outras abordagens de calibração em etapa de pós-processamento.

- O Capítulo 7 detalha a proposta de calibração personalizada; essa técnica usa uma estratégia do tipo chaveamento, onde o usuário pode receber recomendações calibradas com base na popularidade ou no gênero dos itens.

- O Capítulo 8 apresenta a proposta de calibração dupla, que é uma evolução da calibração personalizada. Nessa abordagem, a estratégia utilizada é do tipo empilhamento pelo fato de empilhar em sequência primeiramente a calibração por gênero e depois a calibração dupla.

- Por fim, o Capítulo 9 apresenta as considerações finais, limitações, trabalhos publicados e ideias de trabalhos futuros.

CAPÍTULO

2

CONCEITOS FUNDAMENTAIS

O desenvolvimento de um sistema para gerar recomendações de itens aos usuários envolve conceitos relacionados a sistemas de recomendação, vieses, justiça e calibração. Este capítulo discute os conceitos fundamentais na literatura e visa ajudar a construção desta pesquisa de forma a avançar o estado da arte.

2.1 Sistemas de Recomendação

Com o avanço e popularização da internet em todo o mundo, os sistemas de recomendação foram sendo desenvolvidos e aprimorados de modo a sugerir novos itens para os usuários nos mais diferentes domínios, como em: sites de compras, filmes, notícias, buscas, dentre outros. Dessa maneira, os sistemas de recomendação podem ser definidos como softwares ou técnicas que tem como objetivo auxiliar os usuários na tomada de decisões (RICCI; ROKACH; SHAPIRA, 2015), apresentando a eles itens que sejam de seu interesse.

Esses sistemas são desenvolvidos com o objetivo de diminuir os esforços dos usuários para encontrar novos itens que atendam suas preferências ao mesmo tempo que tentam aumentar a satisfação do usuário durante a utilização do sistema. Para isso, o sistema gera recomendações de itens que podem ser baseadas no histórico de interações do usuário ou em interações de outros usuários com preferências semelhantes ao usuário alvo, podendo obter essas informações a partir de interações explícitas ou implícitas (RICCI; ROKACH; SHAPIRA, 2011).

A área a que esses sistemas pertencem é a de recomendação, que ganhou mais evidência após a realização de um evento em 2006 pela *Netflix*, chamado de *Netflix Prize*. O evento era uma competição que tinha como objetivo aprimorar a precisão dos sistemas de recomendação. Como resultado, novos algoritmos e metodologias expandiram a área de recomendação, além da melhoria dos algoritmos existentes (BENNETT; LANNING *et al.*, 2007).

A partir daí surgiram as principais abordagens dos algoritmos de recomendação, a

filtragem colaborativa e a filtragem baseada em conteúdo. A primeira consiste em gerar recomendações para o usuário alvo com base nas preferências de usuários que tenham gosto parecido ao dele. Já os algoritmos baseados em conteúdo utilizam informações dos dados, estruturados ou não estruturados, sendo esses dados informações dos itens como autores e gênero, por exemplo, juntamente com o histórico do usuário. Há também abordagens baseadas em conhecimento, em regras, sensíveis ao contexto, demográficas e híbridas. Na próxima seção serão discutidas algumas dessas abordagens.

2.2 Abordagens Clássicas

As abordagens empregadas em sistemas de recomendação são baseadas nos seguintes artefatos (RICCI; ROKACH; SHAPIRA, 2011):

- **Interações.** Os sistemas de recomendações continuamente armazenam as interações feitas pelos usuários com os itens do sistema, sejam elas explícitas ou não, com o objetivo de melhorar as recomendações de novos itens. Consistem em uma referência à escolha do item pelo usuário, relacionada ao contexto do domínio do sistema.

- **Itens.** O item é um objeto recomendado ao usuário atrelado ao domínio do sistema. Pode ou não ser associado a custos reais e costuma ter um valor positivo ou negativo. Caso seja do interesse do usuário, seu valor é positivo, caso contrário, seu valor é negativo.

- **Usuários.** Os usuários são vistos como únicos pelo sistema, tendo cada um deles um conjunto de informações específicas para terem recomendações coerentes com suas preferências. A forma que os itens serão recomendados para os usuários dependerá dos algoritmos e abordagens utilizadas pelo sistema.

Há várias abordagens para gerar recomendações e todas elas utilizam os artefatos destacados acima (BURKE, 2007). Essas abordagens geralmente são feitas utilizando as informações de usuários ou informações sobre os itens. Sendo assim, podemos dividir as abordagens existentes nas seguintes classes (AGGARWAL, 2016):

- **Filtragem Colaborativa.** É uma das abordagens mais comuns nos sistemas de recomendação e consiste na ideia que usuários com gostos similares no passado terão as mesmas preferências no futuro, sendo então um sistema baseado na interação de múltiplos usuários (AGGARWAL, 2016). É um sistema fundamentado nas avaliações de usuários ou itens com interações similares, tendo abordagens baseadas em memória e em modelos. As abordagens baseadas em modelo utilizam algoritmos e modelos matemáticos para prever as preferências dos usuários com base em dados de interações passadas. Exemplos incluem modelos de fatorização de matriz e redes neurais. As abordagens com base em memória

podem ser baseadas em usuários ou itens, sendo que a relacionada a usuários utiliza a avaliação de usuários com gosto similar ao do usuário alvo para predizer a avaliação sobre um item não visto. Já a que se baseia em itens analisa a avaliação de outros itens similares feitas pelo usuário alvo para predizer a nota do item não visto.

- **Filtragem Baseada em Conteúdo.** Esse tipo de abordagem avalia juntamente as interações anteriores dos usuários com os metadados dos itens (AGGARWAL, 2016), como gêneros e autores, por exemplo. Dessa forma, consegue resolver problemas como a partida fria.

- **Filtragem Baseada em Conhecimento.** A abordagem baseada em conhecimento procura gerar recomendações analisando o quão útil o item é para o usuário, avaliando a descrição do problema e a solução do problema para recomendar o melhor item que atenda o usuário (AGGARWAL, 2016). Não utiliza o histórico do usuário para gerar as recomendações e podem se basear em regras ou em caso, sendo que as baseadas em regras combinam as especificações do usuários com atributos dos itens respeitando as regras do domínio. Já aquelas baseadas em caso utilizam a similaridade entre as especificações dos usuários e as características dos itens.

- **Filtragem Híbrida.** A ideia da filtragem híbrida é combinar diferentes cenários e algoritmos, de modo a fortalecer a abordagem e as recomendações geradas (AGGARWAL, 2016). Essa técnica de combinar modelos para gerar um modelo mais robusto é conhecida como *ensemble*, e apesar de melhorar a efetividade das recomendações, esse tipo de abordagem pode se tornar muito complexa.

- **Demográfica.** É uma abordagem que utiliza informações demográficas do usuário para gerar novas recomendações, seja por meio da idade, gênero ou de sua localização (AGGARWAL, 2016). Como são informações sensíveis, esse tipo de abordagem pode gerar desconforto ao usuário do sistema.

- **Baseada em Comunidade.** Essa abordagem está relacionada à popularidade das redes sociais, dessa forma, o sistema combina informações do usuário e de seus amigos para gerar recomendações (AGGARWAL, 2016).

As abordagens apresentadas são formas de obter as informações dos usuários e personalizar o sistema de maneira a retornar recomendações específicas para cada usuário. Diante disso, é possível realizar avaliações dos sistemas de recomendação para validar as abordagens e comparar com outras abordagens existentes com o intuito de melhorar o sistema. A próxima seção discute formas existentes para avaliação e as características dos sistemas de recomendação.

2.3 Avaliação dos Sistemas de Recomendação

2.3.1 Formas para Avaliação dos Sistemas de Recomendação

Com o intuito de validar uma abordagem de recomendação e avaliar o sistema como um todo, costumam ser feitos experimentos das seguintes formas:

- ***Offline.*** É uma forma de avaliação realizada a partir de conjuntos de bases de dados de teste e é muito utilizada por conta de sua fácil reprodução e baixo custo (RICCI; ROKACH; SHAPIRA, 2015). Permite medir a capacidade de predição do sistema, mas não mede a reação dos usuários e aspectos como novidade e serendipidade (AGGARWAL, 2016).
- **Experimento com usuários.** Costuma ser conduzido por meio do recrutamento de um grupo de usuários que devem realizar algumas tarefas no sistema, tornando possível avaliar todas as interações realizadas por eles e o comportamento do sistema. É uma opção que corre o risco de haver resultados enviesados dependendo de como for realizada a seleção dos usuários (RICCI; ROKACH; SHAPIRA, 2015).
- ***Online.*** Realizado em sistemas já implementados ou comerciais, sendo menos suscetível ao viés gerado pelo processo de recrutamento dos usuários (AGGARWAL, 2016). Permite avaliar como o sistema influencia o comportamento os usuários.

2.3.2 Características dos Sistemas de Recomendação

Ao avaliar um sistema de recomendação, podem ser analisadas algumas características dos sistemas, como a acurácia e outras relacionadas à experiência do usuário. As principais características são destacadas a seguir:

- **Acurácia.** As métricas de acurácia são as mais utilizadas para avaliar sistemas de recomendação e estão relacionadas com a capacidade do sistema predizar avaliações e preferências dos usuários sobre itens (AGGARWAL, 2016).
- **Cobertura.** Mede a proporção de itens ou usuários que o sistema consegue recomendar (AGGARWAL, 2016).
- **Confiança.** Refere-se à confiança dos usuários no sistema, sendo possível ser medida em experimentos online e por meio de perguntas feitas ao usuário para validar se as recomendações recebidas por eles foram satisfatórias (AGGARWAL, 2016).
- **Diversidade.** É utilizada para medir a dissimilaridade entre as recomendações, sendo interessante quando o usuário prefere receber recomendações além de suas preferências (AGGARWAL, 2016). Quando há mais diversidade no sistema, há também maior novidade e surpresa, porém há uma redução na acurácia.

- **Escalabilidade.** Métrica para avaliar a quantidade de recomendações que um sistema pode fazer pelo tempo de resposta, além de estar relacionada a outras informações do sistema, como a memória disponível (AGGARWAL, 2016).

- **Novidade.** Essa característica é medida pelas recomendações que o usuário não conhecia (AGGARWAL, 2016).

- **Robustez.** É importante para a infraestrutura e escalabidade do sistema, medindo a estabilidade do sistema diante de muitas requisições e de falsas avaliações (AGGARWAL, 2016).

- **Surpresa.** Relacionada a recomendações bem sucedidas de itens inesperados para o usuário (AGGARWAL, 2016).

2.3.3 Métricas para Avaliação dos Sistemas de Recomendação

Para avaliar algumas das características de sistemas de recomendação apresentadas anteriormente, utilizaremos neste trabalho as seguintes métricas:

- ***Mean Average Precision* (MAP).** Métrica utilizada para medir a acurácia do sistema em todo o conjunto de usuários (PARRA; SAHEBI, 2013). Conforme a Equação 2.1, a métrica calcula a média do valor da precisão para todo conjunto de itens recomendados, sendo essa média representada por $AveP(n)$. Seus valores variam de 0 a 1, sendo que quanto maior, melhor.

$$MAP = \frac{1}{|N|}\sum_{n}^{N} AveP(n) \tag{2.1}$$

- ***Mean Reciprocal Rank* (MRR).** Mede a qualidade das listas de recomendação, avaliando o quão distante o primeiro item relevante está do início da lista (QIN, 2013). A Equação 2.2 representa essa medida, sendo que N é número de boas recomendações e $p(i)$ é a posição da recomendação na lista. Seus valores variam de 0 a 1, sendo que quanto maior, melhor.

$$MRR = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{p(i)} \tag{2.2}$$

- ***Mean Rank Miscalibration* (MRMC).** Métrica que computa o grau de calibração da lista conforme as preferências do usuário (SILVA; MANZATO; DURÃO, 2021). A Equação 2.3 obtém o valor de justiça, sendo ele normalizado a partir do pior caso de divergência da lista, dependendo da medida de divergência utilizada e representada por $F(p,q(\{\}))$. Na Equação 2.4 é calculada a soma das médias dos valores dos erros de calibração para cada

posição da lista. Por fim, a Equação 2.5 obtém o valor médio para todos os usuários. Seus valores variam de 0 a 1, sendo que quanto menor, melhor.

$$MC(p,q) = \frac{F(p,q)}{F(p,q(\{\}))} \tag{2.3}$$

$$RMC(u) = \frac{\sum_{j=1}^{N} MC(p,q(R^{*}@J))}{N} \tag{2.4}$$

$$MRMC(u) = \frac{\sum_{u \in U} RMC(u)}{|U|} \tag{2.5}$$

Nos experimentos envolvendo calibração detalhados nos Capítulos 7 e 8, a métrica foi utilizada para avaliar tanto a distribuição dos itens com base nos seus gêneros, bem como com base no aspecto de popularidade. Como o objetivo desta pesquisa é baseado em dois tipos de justiça (gênero e popularidade), propomos neste trabalho usar a média harmônica (ou pontuação F1) entre MRMC de gêneros e popularidade, onde os valores mais altos são melhores:

$$F1 = 2\frac{(1 - MRMC\ Genre) * (1 - MRMC\ Pop)}{(1 - MRMC\ Genre) + (1 - MRMC\ Pop)} \tag{2.6}$$

- ***Long-Tail Coverage*** **(LTC).** Mede a cobertura da cauda longa para todos os usuários, ou seja, verifica a quantidade de itens únicos que são expostos para todos os usuários (ABDOLLAHPOURI; BURKE; MOBASHER, 2018). Essa métrica varia no intervalo de 0 a 1, sendo que 0 significa que todos os itens recomendados são populares, enquanto que 1 significa que todos os itens recomendados são itens de nicho.

  A métrica é representada pela Equação 2.7, onde $\bigcup_{u \in Ut}$ é o conjunto de usuários da lista de recomendação, L_u é a lista de itens recomendados e Φ é o conjunto de itens pertencentes a cauda longa.

$$LTC = \frac{|(\bigcup_{u \in Ut} L_u) \bigcap \Phi|}{|\Phi|} \tag{2.7}$$

  A métrica foi aplicada nos experimentos de calibração detalhados nos Capítulos 7 e 8, onde a proposta do experimento teve seu desempenho comparado com outros algoritmos para verificar se a distribuição dos itens foi feita de forma mais equilibrada na cauda longa.

- ***Group Average Popularity*** **(Δ GAP).** Métrica que mede a variação da popularidade dos itens para os usuários, quanto maior esse número, mais o usuário prefere itens de nicho (ABDOLLAHPOURI *et al.*, 2019b). Essa métrica varia no intervalo $[-1, \infty]$, sendo que o valor negativo indica uma redução no viés de popularidade e o valor positivo indica um aumento do viés de popularidade no grupo de recomendações.

  Os níveis de popularidade de um sistema podem ser divididos em grupos, como em (ABDOLLAHPOURI *et al.*, 2019b), sendo um grupo com preferência por itens populares,

outro grupo com preferência por itens de nicho e um grupo com preferência pelos dois tipos de itens. Para medir a popularidade média dos itens avaliado pelos usuários para um grupo, pode ser utilizada a Equação 2.8, onde ϕ é a popularidade de um item e pu é a lista de itens do perfil do usuário u.

$$GAP(g) = \frac{\sum_{u \in g} \frac{\sum_{i \in pu} \phi(i)}{|pu|}}{|g|} \tag{2.8}$$

A partir disso, é possível representar o valor do GAP para o perfil dos usuários por meio de $GAP(g)_r$ e o GAP das recomendações por meio de $GAP(g)_p$. Assim, é possível calcular a variação de popularidade, como representada na Equação 2.9.

$$\Delta GAP(g) = \frac{GAP(g)_r - GAP(g)_p}{GAP(g)_p} \tag{2.9}$$

Por fim, como os valores ótimos de ΔGAP devem ser próximos de zero, propomos neste trabalho a utilização do Root Mean Squared Error (RMSE) entre os três grupos de usuários, onde os valores mais baixos são melhores:

$$RMSE = \frac{\sqrt{\Delta GAP_{BB}^2 + \Delta GAP_N^2 + \Delta GAP_D^2}}{3} \tag{2.10}$$

Nos experimentos envolvendo calibração, já realizados e detalhados nos Capítulos 7 e 8, foi utilizada a métrica Δ GAP para avaliar a variação da distribuição de popularidade dos itens em três grupos: mais populares (representado por ΔGAP_{BB}), itens de nicho (representado por ΔGAP_N) e o grupo que aceita tanto itens populares, quanto não populares (representado por ΔGAP_D).

Embora o objetivo principal seja atingir uma boa acurácia nas recomendações, as características apresentadas também devem ser levadas em conta ao se escolher uma abordagem para os sistemas de recomendação. Mesmo assim, os sistemas de recomendação ainda sofrem com certas limitações, como os vieses. A próxima seção irá discutir os principais vieses conhecidos e como eles impactam nas recomendações.

2.4 Vieses

Os sistemas de recomendação enfrentam um grande desafio ao apresentarem novas recomendações para os usuários. Trata-se do problema do viés, que ocorre quando o sistema apresenta aos usuários somente determinados itens e limitam suas interações, podendo deixar de oferecer outras opções de itens que sejam do gosto do usuário, além de poder sugerir recomendações diferentes de suas preferências. O viés é um grande desafio que pode deteriorar a eficácia

das aplicações e acarretar questões relevantes como injustiça e bolhas de filtro (ELSWEILER; TRATTNER; HARVEY, 2017).

Como é um problema relevante para a área, trabalhos recentes vêm estudando o assunto. Nesse sentido, foram definidos os seguintes vieses (CHEN *et al.*, 2023a):

- **Viés de Confirmação.** Nesse viés, os usuários tendem a consumir itens de acordo com suas crenças anteriores.
- **Viés de Conformidade.** Ocorre quando o usuário segue o comportamento de outros usuários, fazendo com que suas avaliações não representem aquilo que eles realmente acham. Ocorre, por exemplo, quando o sistema já apresenta a nota média de avaliação de um item e o usuário avalia o item seguindo essa nota.
- **Viés de Exposição.** Esse tipo de viés acontece quando o usuário vê somente parte do feedback de um item, não tendo todas as informações positivas e negativas sobre o mesmo.
- **Viés Indutivo.** Trata-se de quando o modelo de recomendação induz informações a partir dos dados de treinamento.
- **Viés de Posição.** Muito presente em listas de recomendações, esse tipo de viés acontece quando o usuário interage somente com os primeiros itens da lista, independente da qualidade dos outros itens.
- **Viés de Popularidade.** Esse viés faz com que os itens populares sejam recomendados mais frequentemente do que os itens não populares, diminuindo a serendipidade e personalização, além de afetar a experiência do usuário.
- **Viés de Seleção.** Ocorre quando o usuário é livre para selecionar os itens para avaliar, fazendo com que avalie somente itens que gosta, ou avalie somente os itens bons e ruins.

Os vieses são fruto do desequilíbrio de classes em aprendizado de máquina, que tem como resultado uma classificação injusta (ABDOLLAHPOURI *et al.*, 2020). A definição mais aceita sobre justiça é a que a considera como a capacidade dos sistemas de recomendação de fornecer desempenho consistente entre diferentes grupos de usuários (EKSTRAND *et al.*, 2018). Sendo injusto, o sistema está discriminando um determinado indivíduo ou grupo de indivíduos em detrimento de outros. Por exemplo, como resultado do viés de popularidade, ao recomendar somente itens de grande sucesso aos usuários, o sistema está sendo injusto com o grupo de usuários que tende a gostar de conteúdo de nicho.

Além da injustiça, os vieses podem ter como consequência a criação de bolhas de filtro, que consistem no isolamento de indivíduos de diversos pontos de vistas e conteúdos (NGUYEN *et al.*, 2014). As bolhas de filtro têm a capacidade de polarizar usuários e criar grupos

compartilhando a mesma visão, distanciando os usuários de informações que discordem de suas visões (GELFERT, 2018).

Dessa forma, um sistema de recomendação com bolhas de filtro inibe a criatividade e o aprendizado (PARISER, 2011). As bolhas de filtro são um problema bem relevante porque são invisíveis e as pessoas não notam que estão deixando de ver outros pontos de vista sobre a situação (LUNARDI *et al.*, 2020). Além disso, as pessoas tendem a consumir conteúdo com os quais concordam e evitar aqueles que discordam (BOZDAG *et al.*, 2014), bem como tendem a estabelecer conexões com aqueles que compartilham suas crenças e interesses (PASSE; DRAKE; MAYGER, 2018). Assim, os sistemas contribuem para aumentar a criação e efeito das bolhas de filtro.

Pelo fato de polarizar as pessoas, as bolhas de filtro representam um grande risco à democracia, já que as pessoas tendem a se posicionar seguindo os ideais dos conteúdos que visualizam nos sistemas de recomendação de suas aplicações, sem se dar conta da necessidade de entender outros pontos de vista sobre a situação para terem uma análise mais racional do que fanática. Por isso, é necessário que as pessoas tenham consciência das bolhas de filtro e se desafiem vendo outros pontos de vista (PARISER, 2011).

Na literatura, há varias abordagens diferentes para lidar com o problema das bolhas de filtro (LUNARDI *et al.*, 2020). Uma maneira de fazer isso é por meio da criação de algoritmos focados na diversidade (MUNSON; RESNICK, 2010), com o objetivo de apresentar ao usuário itens um pouco diferentes de suas preferências. Também é possível adicionar formas de visualização interativa no sistema, de forma a quebrar as bolhas de filtro (HELBERGER; KARPPINEN; D'ACUNTO, 2018), assim, os usuários serão expostos a itens diversos e terão mais entendimento sobre o mecanismo de filtragem e sobre as bolhas de filtro (NAGULENDRA; VASSILEVA, 2014).

Em razão de muitos pesquisadores acreditarem que os sistemas de recomendação devem ter uma grande acurácia (KONSTAN; RIEDL, 2012), as bolhas de filtro ainda persistem. Conforme (LUNARDI *et al.*, 2020), para eliminar esse problema, os sistemas devem focar além da acurácia, considerando então aspectos como: novidade, serendipidade, diversidade, dentre outros (RICCI; ROKACH; SHAPIRA, 2015).

Além do problema das bolhas de filtro, há também a questão da injustiça presente em sistemas de recomendação. Uma das estratégias utilizadas para lidar com isso é por meio da calibração dos sistemas. Na próxima seção será discutida essa abordagem de calibração.

2.5 Justiça e Calibração

Para combater a injustiça em sistemas de recomendação, existem algumas estratégias que podem ser usadas para o desenvolvimento de um sistema de recomendação mais justo, entre

elas, o mecanismo de *nudges* e a calibração das recomendações. O mecanismo de *nudge* é uma maneira de alterar a interface do sistema de modo a influenciar positivamente o comportamento do usuário no sistema e ajudá-lo a tomar decisões (JESSE; JANNACH, 2021). Assim, ele pode funcionar em um sistema de *e-commerce* de livros, por exemplo, por meio de um texto colorido indicando que o livro foi escrito por mulheres, com o intuito de promover esses livros e evitar o desfavorecimento deles em relação aos livros escritos por homens, que em um sistema injusto podem estar sendo favorecidos e recomendados mais vezes. Já a calibração visa aplicar alterações na geração de recomendações para melhorarem a eficácia dessas recomendações.

Em nossa pesquisa, realizamos um experimento com o mecanismo de *nudges*, como descrito no Capítulo 5, e que tinha o objetivo de verificar como os *nudges* auxiliam na redução do viés de popularidade de um sistema de recomendação. O experimento serviu para demonstrar como a interface pode ser utilizada para combater o viés de popularidade. Caso a opção seja alterar a geração das recomendações, a calibração é uma estratégia utilizada para reduzir a injustiça em sistemas de recomendação, pois tem a capacidade de fornecer recomendações aos usuários com proporções de itens em suas áreas de interesse que sejam consistentes com suas preferências (STECK, 2018).

Além disso, como verificado em (KAYA; BRIDGE, 2019), a calibração pode funcionar tanto em uma etapa de processamento quanto em etapa de pós-processamento, ou seja, logo após os dados do sistema já terem sido treinados com as preferências dos usuários e gerado recomendações que possam interessar a eles. Dessa forma, é possível aplicar a calibração de forma independente ao sistema que gerou as recomendações.

2.6 Considerações Finais

Os vieses encontrados nos sistemas de recomendações impactam na eficácia das recomendações e na utilização do sistema pelos usuários. Assim, identificar esses vieses e tentar amenizar o impacto deles nos sistemas tornaram-se os principais desafios para pesquisadores que visam melhorar os sistemas de recomendação.

Conforme verificado neste capítulo, muitos autores tendem a aumentar a acurácia do sistema de modo a aumentar a satisfação do usuário na aplicação, entretanto, isso acarreta problemas como as bolhas de filtro e injustiça. Esses problemas geram consequências na prática e até causam risco à democracia pelo fato de aumentar a polarização das pessoas, sendo então necessária uma conscientização a respeito desses problemas.

Nos trabalhos existentes, há algumas abordagens para reduzir o impacto desses vieses e da injustiça nos sistemas, como a calibração. Tendo isso em vista, a nossa proposta emprega a calibração para adequar a recomendação dos itens ao nível de preferência dos usuários, procurando dessa forma trazer recomendações coerentes com as preferências dos usuários e que reduzam o impacto do viés de popularidade no sistema. No próximo capítulo, serão discutidos

os trabalhos relacionados que abordam o viés de popularidade e que fornecem algum tipo de calibração, para diminuir o impacto desses problemas e melhorar os sistemas com recomendações de acordo com os interesses dos usuários.

CAPÍTULO

3

TRABALHOS RELACIONADOS

Neste capítulo, resumimos os trabalhos relacionados com base nos objetivos específicos que temos para este projeto e apontamos questões em aberto que pretendemos explorar mais neste estudo.

Considerando isso, a literatura revisada neste capítulo busca possibilitar o progresso do trabalho e conduzi-lo à inovação de ponta, destacando questões em aberto que precisam ser mais exploradas.

As buscas por trabalhos relacionados foram realizadas em um ambiente virtual de pesquisa, que organiza textos e metadados da literatura acadêmica. Para tanto, foram utilizadas palavras-chave como “Sistemas de Recomendação”, “Calibração”, “Viés de Popularidade”, “Viés e Injustiça” e outros sinônimos para busca dos artigos. Filtrando por períodos, foram listados os trabalhos relevantes mais recentes para este trabalho.

3.1 O Viés de Popularidade e Seu Impacto

O viés de popularidade é uma limitação bem conhecida na área de sistemas de recomendação e, por conta disso, alguns trabalhos recentes foram desenvolvidos com o objetivo de analisar o impacto desse viés em alguns sistemas existentes. Destacamos a seguir alguns desses trabalhos.

De início, o trabalho (DELDJOO *et al.*, 2023) deixa claro que os vieses existentes no sistema acabam refletindo nas recomendações como um todo. Assim, no trabalho (ABDOLLAHPOURI *et al.*, 2019a) os resultados obtidos sobre o viés de popularidade mostram que quanto mais um grupo de usuários é afetado por esse viés, maior será o erro de calibração do sistema no geral. Outra característica apresentada é em relação ao aspecto de popularidade, sendo que o grupo com menor interesse em itens populares acaba por ser o mais afetado pelo viés no processo de calibração, tendo menos itens apresentados consistentes com suas preferências.

Essa característica é corroborada também pelo artigo (KOWALD; SCHEDL; LEX, 2020), onde são feitos experimentos offline no domínio de música, os quais mostram que usuários que preferem itens de nicho recebem as piores recomendações e que os algoritmos de recomendação do estado da arte favorecem itens mais populares. O artigo (ABDOLLAHPOURI *et al.*, 2019b) também valida essa questão e ainda verifica que cada usuário possui um nível de interesse pelo aspecto de popularidade dos itens e que nem todos usuários são afetados da mesma forma pelo viés de popularidade.

Existe também uma outra questão muito importante verificada na literatura com relação a amplificação dos problemas em um sistema de recomendação enviesado. Nesse sentido, o artigo (MANSOURY *et al.*, 2020) mostra que sistemas que já possuem algum viés tendem a aumentar o impacto do viés existente de tal forma que a representação dos gostos dos usuários é alterada ao longo do tempo, reduzindo a qualidade das recomendações, havendo também uma homogeneização dos usuários. Ademais, por meio da realização de experimentos offline, é verificado que os usuários mais impactados pelos vieses pertencem ao grupo minoritário, isto é, são os usuários que preferem itens de nicho.

Por fim, o trabalho (WANG *et al.*, 2022) mostra uma outra visão sobre os vieses nos sistemas de recomendação, onde é feito um experimento com usuários utilizando dois sistemas de recomendação baseado em gêneros, um deles com calibração para redução de vieses e outro sem. Na pesquisa realizada, os usuários acabaram tendo uma preferência pelo sistema que apresentava os vieses, o que mostra ser necessário além de uma melhoria nos algoritmos existentes, também haver melhoria nos aspectos humanos e sociais.

Outro trabalho que analisa a percepção dos usuários com relação às recomendações do sistema é o (INGESSON, 2022), onde é feita uma análise da percepção dos usuários sobre o aspecto de justiça em um sistema de recomendação no domínio de música. Nesse trabalho, há três algoritmos com níveis de popularidade diferentes e o usuário recebe recomendações a partir de um deles e depois responde a um questionário de acordo com as recomendações recebidas. Como resultado, os usuários não perceberam diferença na justiça entre os algoritmos apresentados.

Embora seja um resultado ruim para a análise da popularidade na visão dos usuários, isso pode ter acontecido pelo fato do sistema proposto não levar em conta o nível de interesse dos usuários pelo aspecto de popularidade, bem como não terem sido recrutados usuários que se interessam por itens de nicho, que costumam ser os mais afetados pelo viés de popularidade. Além disso, o trabalho não apresentou métricas que validam que os algoritmos implementados na proposta realmente ajudam a melhorar a justiça no sistema, como aquelas apresentadas no Capítulo 2. Então pode ser o caso dos algoritmos não terem alterado a justiça das recomendações e por isso os usuários não sentiram diferença nesse aspecto.

Os artigos apresentados anteriormente são importantes para demonstrar como o viés de popularidade interfere na calibração e justiça de um sistema, seja favorecendo os itens

mais populares, seja diminuindo a calibração de um sistema, fazendo com que seja necessário desenvolver formas de reduzir o impacto desse viés e trazer sistemas mais justos. Sendo assim, diferente dos trabalhos que somente analisam esse viés, o presente trabalho visa propor um sistema que reduza o impacto desse viés e traga recomendações coerentes com as preferências dos usuários. Nesse sentido, há também pesquisas recentes que mostram formas de lidar com esse viés, reduzindo o seu impacto sobre o sistema, como os trabalhos destacados a seguir.

Para lidar com o viés de popularidade e reduzir seu impacto, o trabalho (GHARAHIGHEHI; VENS; PLIAKOS, 2021) implementa uma abordagem que faz uso de hipergrafos para reduzir o peso dos itens conforme sua popularidade nas iterações feitas pelo sistema, funcionando de forma dinâmica e por meio do aprendizado constante. Apesar de ter bons resultados no aumento da cobertura da cauda longa, a proposta gera uma redução na precisão das recomendações.

O artigo (ZHU *et al.*, 2021), por sua vez, mostra que algoritmos de recomendação baseados em fatoração de matrizes estão inerentemente ligados ao viés de popularidade e propõe também uma forma de redução de vieses com redução dos pesos dos itens na etapa de pós-processamento, além de uma diminuição na correlação da popularidade dos itens entre a popularidade dos itens e a predição na etapa de processamento. O sistema proposto também reduz o viés de popularidade, mas não retorna recomendações calibradas.

No trabalho (YALCIN; BILGE, 2021), a proposta para redução do viés é feita por um novo ranqueamento de recomendações a partir de duas estratégias: a primeira penaliza os itens mais populares durante o agrupamento dos itens, enquanto que a segunda aplica mais ênfase nas recomendações em grupo do que no aspecto de popularidade. Os resultados mostram que ambas as estratégias reduzem o viés, mantendo uma acurácia razoável nas recomendações.

A implementação de um novo ranqueamento em etapa de pós-processamento também é adotada em (ABDOLLAHPOURI; BURKE; MOBASHER, 2019), onde os itens da cauda longa são valorizados de forma a obter uma redução no viés com pouca perda na precisão. Essa abordagem de atribuição de pesos também é aplicada em (BORGES; STEFANIDIS, 2021), só que neste trabalho ela é feita a partir de redes neurais através de autoencoders variacionais, com a aplicação de pesos negativos nos itens mais populares.

Outro tipo de abordagem para redução do viés de popularidade é por meio do modelo de inferência causal, como proposto em (WANG *et al.*, 2021), onde é feito um ajuste nas recomendações de forma a melhorar as predições feitas pelo sistema, de modo que a partir de métricas de divergência, o histórico do usuário é analisado verificando se o usuário é mais suscetível a diversidade ou não, aplicando pesos diferentes nas recomendações em caso positivo. Como resultado, o sistema apresenta uma melhora na precisão e redução do viés.

Essa abordagem de inferência causal também é adotada em (ZHANG *et al.*, 2021), onde o viés de popularidade não é tratado como algo essencialmente ruim, fazendo uso da popularidade do item no treinamento do modelo para ajustar a pontuação das recomendações. O trabalho (WEI

et al., 2021) também implementa essa abordagem, removendo a popularidade do item na fase de treinamento para melhorar as predições.

Como verificado nos trabalhos apresentados, essas abordagens se preocupam somente com a redução do impacto do viés de popularidade no sistema de recomendação e não apresentam formas de calibrar os sistemas para retornar recomendações coerentes com as preferências dos usuários, além da grande maioria não se preocupar com o nível de interesse dos usuários pelo aspecto de popularidade. Sendo assim, a próxima seção mostra alguns estudos que visam retornar recomendações calibradas.

3.2 Calibração de Sistemas de Recomendação

A literatura recente explora três estratégias distintas para calibrar um sistema de recomendação para alinhá-lo com as preferências do usuário (PITOURA; STEFANIDIS; KOUTRIKA, 2022). Essas estratégias são elucidadas a seguir.

3.2.1 Calibração em Etapa de Pré-Processamento

Normalmente, esta etapa considera que as inconsistências de recomendação surgem dos dados usados para treinar o modelo de recomendação. Portanto, visa ajustar ou mesmo desconsiderar determinados aspectos dos dados de treinamento e envolve a aplicação de ajustes nos dados de treinamento para mitigar os vieses existentes (PITOURA; STEFANIDIS; KOUTRIKA, 2022).

O trabalho (AHANGER *et al.*, 2022) introduz duas abordagens para aliviar o viés de popularidade usando esta estratégia. Uma abordagem visa neutralizar o viés de popularidade nas recomendações, categorizando os itens em percentis de popularidade e avaliando a precisão dentro de cada percentil. Isto atenua eficazmente os vieses, especialmente nas distribuições de popularidade de cauda longa, equilibrando as recomendações entre itens populares e impopulares.

A outra abordagem envolve a criação de divisões de dados onde todos os itens têm um número igual de classificações de teste. Embora essas técnicas reduzam o viés de popularidade, elas diminuem simultaneamente a precisão do sistema. Sendo assim, optamos por empregar estratégias alternativas em nossas propostas.

3.2.2 Calibração em Etapa de Processamento

Esta etapa envolve a modificação ou introdução de novos algoritmos com o objetivo de reduzir vieses no modelo de treinamento (PITOURA; STEFANIDIS; KOUTRIKA, 2022). Esta estratégia envolve modificar ou implementar modelos e algoritmos para produzir resultados mais consistentes. A abordagem do BPR (*Bayesian Personalized Ranking from Implicit Feedback*) (RENDLE *et al.*, 2012) exemplifica esta estratégia ao atribuir pesos positivos e negativos aos itens

com base nas interações dos usuários. Utilizando a técnica do gradiente descendente estocástico, o sistema visa maximizar a qualidade dos itens recomendados.

A abordagem que funciona na fase de processamento também é adotada no estudo (BORATTO; FENU; MARRAS, 2021), onde é minimizada a correlação entre a relevância do item e a popularidade dele para promover os itens de forma mais igualitária na cauda longa. Como resultado, há também uma pequena perda na acurácia.

A literatura também explora vários métodos de calibração dentro desta estratégia. Por exemplo, (CHEN *et al.*, 2023b) implementa uma abordagem *pairwise* considerando a justiça entre grupos de itens, ajustando a lista de recomendações durante o treinamento. Por outro lado, no artigo (LIU *et al.*, 2022), o autor propõe uma técnica que utiliza redes neurais e grafos para reduzir discrepâncias na precisão das recomendações entre usuários semelhantes devido a informações inerentes aos seus perfis.

O trabalho (ZHU; WANG; CAVERLEE, 2020) demonstrou que a implementação do BPR aumenta a injustiça do sistema por favorecer alguns itens em detrimento de outros. O autor então propõe uma calibração para reduzir o impacto da injustiça. Já em (BEUTEL *et al.*, 2019), os autores propõem uma abordagem por meio da regularização pareada para melhorar a classificação dos itens durante a fase de treinamento a partir de métricas que avaliam a imparcialidade do sistema. Embora estas abordagens produzam resultados interessantes, tais estudos não abordam o viés de popularidade. Visando preencher esta lacuna, no Capítulo 4 será detalhada a proposta de modificação do BPR, de forma a calibrar o sistema para reduzir o viés de popularidade, sendo essa uma das propostas deste trabalho.

3.2.3 Calibração em Etapa de Pós-Processamento

Esta etapa altera a saída gerada pelo algoritmo de recomendação, permitindo-lhe abordar os vieses existentes na recomendação inicial. No entanto, esta técnica pode comprometer a precisão do sistema, modificando os resultados iniciais (PITOURA; STEFANIDIS; KOUTRIKA, 2022).

Com o objetivo de equilibrar um sistema calibrado com um que gera recomendações precisas, o trabalho (SILVA; MANZATO; DURÃO, 2021) propôs um sistema de calibração que funciona em uma etapa de pós-processamento, independente de qualquer algoritmo de recomendação e baseado em medidas de divergência. Dessa maneira, é apresentada uma abordagem com calibração baseada em gêneros e com pesos personalizados com o intuito de oferecer recomendações coerentes com as preferências dos usuários. Embora tenha resultados muito bons nas métricas de justiça, o sistema ainda sofre do viés de popularidade.

Seguindo a mesma ideia, Steck (STECK, 2018) apresenta uma técnica de calibração também baseada nos gêneros dos itens, que funciona em etapa de pós-processamento e visa atender a proporção de interesse dos usuários no aspecto de gênero dos itens. O trabalho

mostra o retorno de recomendações com proporções que respeitam as preferências dos usuários, mas também enfrenta o problema do viés de popularidade. O trabalho (GEYIK; AMBLER; KENTHAPADI, 2019), por sua vez, também apresenta um sistema com o objetivo de reduzir a injustiça baseado no gênero e idade dos usuários nas vagas de emprego oferecidas pelo sistema. Como resultado, a abordagem de reclassificação das recomendações obteve uma melhoria nas métricas de justiça.

Da mesma forma, alguns trabalhos de excelência do estado da arte também estudam a calibração das recomendações, de forma a reduzir a injustiça das recomendações e evitar o desfavorecimento de alguns itens. Sendo assim, o trabalho (KAYA; BRIDGE, 2019) segue a abordagem de (STECK, 2018) alterando a calibração por gêneros por uma baseada em subperfis de usuários e seus interesses, para retornar recomendações coerentes com as preferências dos usuários.

Há também estratégias para redução do impacto dos vieses que envolvem heurísticas (SEYMEN; ABDOLLAHPOURI; MALTHOUSE, 2021) e que conseguem reduzir o viés de popularidade e calibrar as recomendações de acordo com as preferências do usuário. Por fim, em (ABDOLLAHPOURI *et al.*, 2021), a proposta feita calibra o sistema dividindo os usuários em três níveis de preferência por popularidade e traz as recomendações conforme esse aspecto.

3.3 Considerações Finais

Conforme apresentado neste capítulo, existem diferentes abordagens para redução do viés de popularidade e para calibração dos sistemas de recomendação. A Tabela 1 destaca as principais características relacionadas com a proposta dos trabalhos apresentados.

As atuais abordagens para calibração das recomendações não aplicam a redução do viés de popularidade e não levam em conta o nível de preferência dos usuários por esse aspecto. Tendo isso em vista, a proposta do presente trabalho é estudar abordagens que reduzam o viés de popularidade e tragam recomendações calibradas conforme a preferência dos usuários. Para isso, pretende-se investigar alguns métodos existentes de calibração. Desse modo, os próximos capítulos detalham as diferentes abordagens desta pesquisa.

Tabela 1 – Comparação entre os trabalhos relacionados e a presente proposta de pesquisa

Trabalhos	**Agnóstico de Modelo de Recomendação**	**Reduz o Viés de Popularidade**	**Calibração das Recomendações**	**Calibração do Nível de Preferência por Popularidade**

Continua na próxima página

Tabela 1 – Comparação entre os trabalhos relacionados e a presente proposta de pesquisa

Trabalhos	**Agnóstico de Modelo de Recomendação**	**Reduz o Viés de Popularidade**	**Calibração das Recomendações**	**Calibração do Nível de Preferência por Popularidade**
(GHARAHIGHEHI; VENS; PLIAKOS, 2021)	✓	✓	✓	
(ZHU *et al.*, 2021)	✓	✓		
(BORATTO; FENU; MARRAS, 2021)		✓		
(YALCIN; BILGE, 2021)	✓	✓		
(ABDOLLAHPOURI; BURKE; MOBASHER, 2019)	✓	✓		
(BORGES; STEFANIDIS, 2021)	✓	✓		
(WANG *et al.*, 2021)	✓	✓		
(ZHANG *et al.*, 2021)	✓	✓		
(WEI *et al.*, 2021)	✓	✓		
(SILVA; MANZATO; DURÃO, 2021)	✓		✓	
(STECK, 2018)	✓		✓	
(GEYIK; AMBLER; KENTHAPADI, 2019)			✓	
(KAYA; BRIDGE, 2019)	✓		✓	
(ZHU; WANG; CAVERLEE, 2020)	✓	✓	✓	
(BEUTEL *et al.*, 2019)			✓	

Continua na próxima página

Tabela 1 – Comparação entre os trabalhos relacionados e a presente proposta de pesquisa

Trabalhos	**Agnóstico de Modelo de Recomendação**	**Reduz o Viés de Popularidade**	**Calibração das Recomendações**	**Calibração do Nível de Preferência por Popularidade**
(SEYMEN; ABDOLLAHPOURI; MALTHOUSE, 2021)	✓	✓	✓	
(ABDOLLAHPOURI *et al.*, 2021)	✓		✓	✓
BPR Modificado - Capítulo 4		✓	✓	✓
Abordagem com *Nudges* - Capítulo 5	✓	✓		
Calibração em Pós-Processamento - Capítulo 6	✓	✓	✓	✓
Calibração Personalizada - Capítulo7	✓	✓	✓	✓
Calibração Dupla Capítulo 8	✓	✓	✓	✓

Fonte: Elaborada pelo autor.

CAPÍTULO

4

UMA ABORDAGEM DE CALIBRAÇÃO NA FASE DE PROCESSAMENTO

Este capítulo descreve de forma detalhada a abordagem que utiliza a calibração em etapa de processamento e é organizado da seguinte forma: a Seção 4.1 apresenta a justificativa de realização dessa abordagem; a Seção 4.2 descreve a metodologia e o design do experimento; por fim, os resultados são apresentados e discutidos na Seção 4.3.

4.1 Justificativa

Para aumentar a justiça em Sistemas de Recomendação, a literatura mostra que é possível calibrar o sistema utilizando três tipos de estratégias: calibração em etapa de pré-processamento, calibração em etapa de processamento e, por fim, calibração em etapa de pós-processamento (PITOURA; STEFANIDIS; KOUTRIKA, 2022). Sabendo disso, a primeira estratégia estudada neste trabalho foi a estratégia de calibração em etapa de processamento.

Esse tipo de estratégia visa modificar algoritmos existentes ou introduzir novos algoritmos que resultem em classificações e recomendações justas, por exemplo, removendo preconceitos e discriminação durante o processo de treinamento do modelo. Normalmente, tais métodos visam aprender um modelo sem preconceitos, ao mesmo tempo que consideram a justiça durante o treinamento de um modelo, incorporando mudanças na função objetivo de um algoritmo por um termo de justiça ou impondo restrições de justiça (PITOURA; STEFANIDIS; KOUTRIKA, 2022).

A literatura apresenta diversos trabalhos relacionados com a estratégia em etapa de processamento. A abordagem BPR *(Bayesian Personalized Ranking for Implicit Feedback)* (RENDLE *et al.*, 2012) é uma técnica LTR (*Learning to Rank*) do tipo *pairwise* que procura posicionar itens relevantes no topo da lista de recomendação. Para isso, são feitas comparações entre pares de itens – um conhecido e outro desconhecido pelo usuário – de modo a maximizar

a diferença de suas respectivas representações. Apesar de não lidar com injustiça e vieses, o BPR possui uma flexibilidade em sua construção que permite utilização com outros modelos de recomendação, e também extensões para que outras condições (como vieses e injustiça) sejam impostas durante seu treinamento (BORATTO; FENU; MARRAS, 2021).

O trabalho (CHEN *et al.*, 2023b) implementa uma abordagem pareada considerando a justiça entre grupos de itens, ajustando a lista de recomendações durante o treinamento. Por outro lado, (LIU *et al.*, 2022) sugere a utilização de grafos e redes neurais para atribuir pesos aos itens recomendados, equilibrando-os de acordo com as preferências do usuário. Embora essas abordagens produzam resultados interessantes, ambos os estudos não abordam o viés de popularidade.

Outros trabalhos da literatura mostram abordagens em etapa de pós-processamento que calibram as recomendações de forma a melhorar a justiça em termos dos gêneros dos itens, como em (STECK, 2018) e (SILVA; MANZATO; DURÃO, 2021). Esse tipo de abordagem será estudada nos Capítulos 6, 7 e 8. Ademais, essa calibração poderia ser adaptada para levar em conta os aspectos de popularidade dos itens ao invés do gênero e possivelmente auxiliar a reduzir o viés de popularidade no sistema como um todo. Além disso, essa calibração poderia ser incorporada à etapa de processamento de forma a melhorar a eficiência do sistema nas recomendações.

Assim, a ideia desse estudo é combinar uma forma de calibração personalizada baseada na popularidade dos itens com o método BPR (RENDLE *et al.*, 2012) em etapa de processamento. O objetivo dessa combinação é trazer um sistema eficiente que traga recomendações coerentes com as preferências dos usuários, reduza o viés de popularidade e se aproveite do mecanismo de otimização de ranking das recomendações de acordo com a relevância dos itens. As próximas seções descrevem a implementação dessa combinação e os resultados dessa abordagem.

4.2 Metodologia

4.2.1 Calibração por Popularidade

Suponha que há um conjunto de itens $I = \{i_1, i_2, ..., i_{|I|}\}$, um conjunto de usuários $U = \{u_1, u_2, ..., u_{|U|}\}$ e um conjunto de itens candidatos para cada usuário $CI_u = \{i_1, i_2, ..., i_N\}$, onde N é o número de itens sugeridos pelo sistema de recomendação. Além disso, existem as informações dos usuários sobre as preferências de gênero e popularidade. A tarefa é explorar essas preferências para gerar uma lista de recomendações que aumente a justiça da popularidade.

Para tanto, propõe-se uma abordagem de calibração em etapa de processamento. Na prática, o método utiliza medidas de divergência na etapa de geração de recomendações para realizar uma calibração de acordo com diferentes níveis de popularidade de interesse do usuário. Como resultado, os usuários recebem uma lista de recomendações próxima ao seu perfil em

termos de popularidade. Essa calibração é incorporada ao BPR. A Figura 1 apresenta a estrutura de calibração de popularidade, cujos detalhes são descritos a seguir.

Figura 1 – Estrutura de calibração proposta. A calibração por popularidade é aplicada de forma combinada a execução do BPR, resultando em uma lista calibrada de recomendações de acordo com as preferências do usuário sobre popularidade e gêneros.

A calibração da lista de recomendações com base na popularidade dos itens consumidos pelo usuário no passado foi feita por meio de uma divisão de popularidade para agrupar os itens com base na quantidade de interações. A divisão de popularidade, introduzida em (ABDOLLAHPOURI *et al.*, 2021), é baseada no conceito de cauda longa dos sistemas de recomendação. A curva foi dividida em três partes. O ***Head*** **(H)**, com itens representando os 20% principais do total de interações passadas. A ***Tail*** **(T)** com itens que somam menos 20% das interações, e o grupo ***Mid*** **(M)**, que contém itens que não são nem ***Head*** **(H)** nem ***Tail*** **(T)**, como demonstra a Figura 2. Vale ressaltar que esta divisão por percentual foi escolhida com base no princípio de Pareto.

A calibração por popularidade foi uma adaptação da fórmula proposta por (STECK, 2018). Seu trabalho pressupõe que os itens podem ter mais de um gênero, o que não é válido no contexto de popularidade, onde um item possui apenas um nível de popularidade. Então, ao invés disso, foi calculada as somas dos pesos de cada tipo de popularidade sobre a soma de todos os pesos.

Assim, considerando t como a categoria de popularidade, $x(t|u)$ é definido como a distribuição alvo baseada na popularidade dos itens com os quais o usuário interagiu no passado.

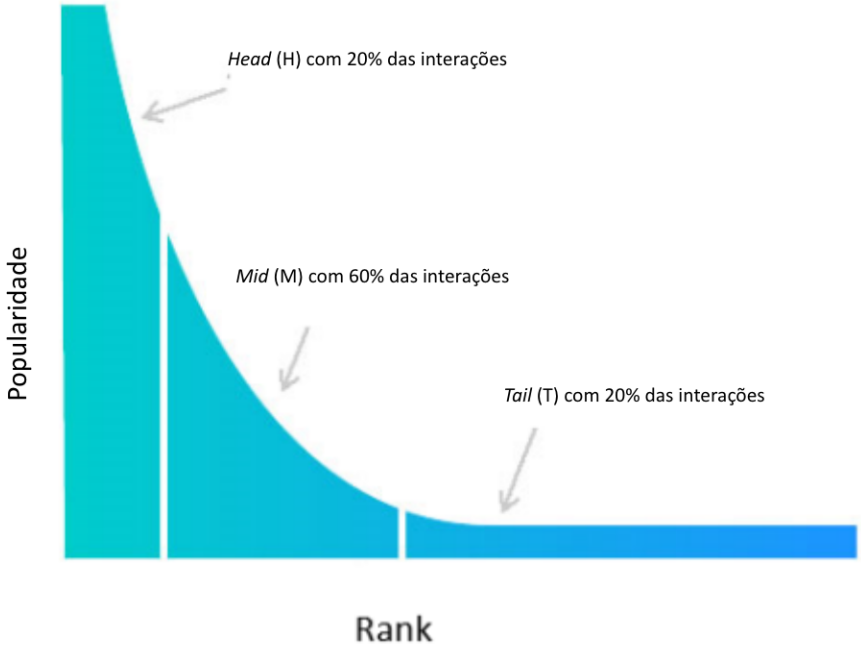


Figura 2 – Curva representando a divisão dos itens em grupos de popularidade.

Na Equação 4.1 os pesos r_{ui} são definidos como a classificação explícita ou implícita que o usuário u deu ao item i:

$$x(t|u) = \frac{\sum_{i \in I_u} r_{ui} \cdot x(t|i)}{\sum_{i \in I_u} r_{ui}} \tag{4.1}$$

onde I_u é o conjunto de itens interagidos pelo usuário u, e $x(t|i)$ é definido como 1 se o item i estiver na categoria de popularidade t. Então, para lidar com a distribuição de lista recomendada, a Equação 4.2 define $y(t|u)$ como:

$$y(t|u) = \frac{\sum_{i \in R_u^*} w_p(u,i) \cdot x(t|i)}{\sum_{i \in R_u^*} w_p(u,i)} \tag{4.2}$$

Neste caso, usamos os pesos $w_p(u,i)$ como a posição de classificação do item i na lista reordenada recomendada R_u^* para o usuário u.

Várias métricas avaliam a imparcialidade em sistemas de recomendação (VERMA; GAO; SHAH, 2020). Porém, nesse caso, utiliza-se a medida de divergência Kullback-Leibler pelas mesmas razões apontadas por (STECK, 2018) e exploradas por (SILVA; MANZATO; DURÃO, 2021). O Kullback-Leibler quantifica a desigualdade no intervalo $[0, \infty]$, onde 0 significa que ambas as distribuições são quase iguais e valores mais altos indicam injustiça.

Ademais, é adotada a regularização proposta por (STECK, 2018), que definiu $\alpha = 0.01$ como uma variável de regularização para evitar divisão por zero quando $y(t|u)$ vai para zero. Embora existam outras métricas de divergência, como Hellinger e Person Qui-Square, propostas

por (CHA, 2007) e explorado por (SILVA; MANZATO; DURÃO, 2021), foi utilizada apenas a Kullback-Leibler devido à sua simplicidade.

$$D_{KL}(x||y) = \sum_t x(t|u) \cdot log \frac{x(t|u)}{(1-\alpha) \cdot y(t|u) + \alpha \cdot x(t|u)} \tag{4.3}$$

A divergência de Kullback-Leibler é uma medida que quantifica a diferença entre duas distribuições de probabilidade, neste caso, entre a distribuição observada $x(t|u)$ e a distribuição de referência $y(t|u)$. No contexto da calibração por popularidade, $x(t|u)$ representa a distribuição empírica dos itens observados pelo usuário u, enquanto $y(t|u)$ representa uma distribuição de referência desejada, que é baseada na popularidade dos itens na base de dados.

4.2.2 O Método BPR

O BPR (RENDLE *et al.*, 2012) é uma abordagem eficaz para recomendação de itens em sistemas de recomendação baseados em feedback implícito. Ao modelar as preferências dos usuários por meio de características latentes e otimizar a função de perda, o modelo é capaz de aprender efetivamente as preferências dos usuários e gerar recomendações personalizadas, levando em consideração a ordem de preferência dos itens.

Mantendo a notação utilizada na seção anterior, as letras de indexação especial distinguem usuários e itens: um usuário é indicado como u e um item é referido como i, j; r_{ui} refere-se ao feedback explícito ou implícito de um usuário u para um item i. No primeiro caso, é um número inteiro fornecido pelo usuário indicando o quanto ele gostou do conteúdo; no segundo caso, é apenas um booleano mostrando se o usuário consumiu ou visitou o conteúdo ou não. A predição do sistema sobre a preferência do usuário u para o item i é representada por $\hat{r}_{ui}$, que é um valor de ponto flutuante estimado pelo algoritmo de recomendação. O conjunto de pares (u,i) para os quais r_{ui} é conhecido é representado por $K = \{(u,i)|r_{ui}\}$.

Em um modelo de fatorização tradicional, cada usuário u é associado a um vetor de fatores $p_u \in \mathbb{R}^f$ e cada item i com um vetor de fatores $q_i \in \mathbb{R}^f$. Uma regra de previsão seria:

$$\hat{r}_{ui} = p_u^T q_i \quad (1) \tag{4.4}$$

Conjuntos adicionais são $N(u)$, que indica o conjunto de itens para os quais o usuário u forneceu um feedback implícito, e $\overline{N}(u)$, que indica o conjunto de itens desconhecidos para o usuário u. Uma característica importante desse tipo de feedback é que apenas as observações positivas são conhecidas; os pares usuário-item não observados são interpretados como feedback negativo.

O trabalho (RENDLE *et al.*, 2012) discute um problema que surge quando um modelo de recomendação de itens é treinado apenas com esses dados positivos/negativos. Como as entradas observadas são positivas e as restantes são negativas, o modelo será ajustado para fornecer

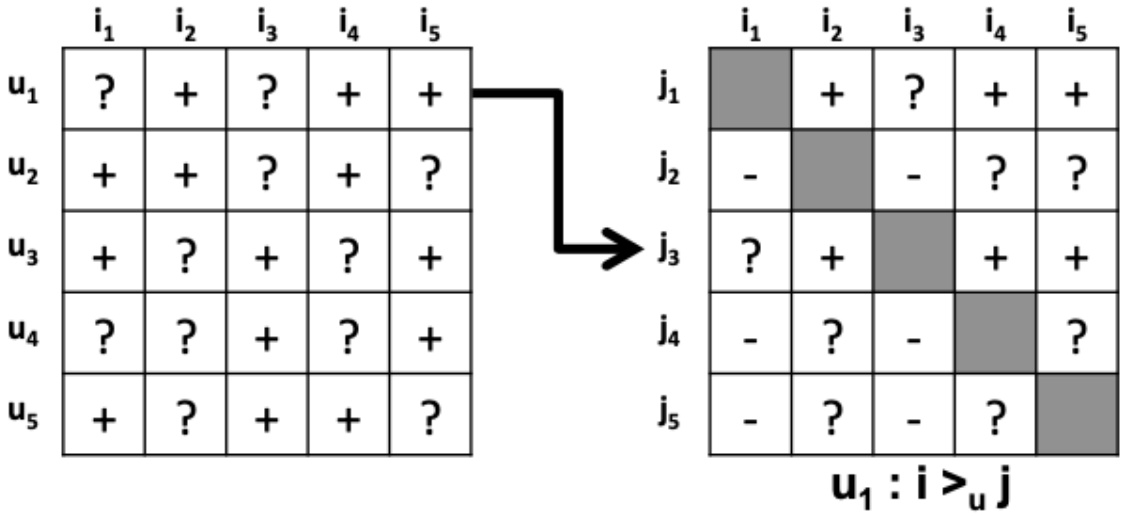


Figura 3 – O quadro à esquerda representa os dados observados. A abordagem cria uma relação par de itens específica para o usuário $i \succ_u j$ entre dois itens. No lado direito da tabela, o sinal de mais indica que o usuário u está mais interessado no item i do que no item j; o sinal de menos indica que o usuário prefere o item j ao i; o ponto de interrogação indica que não se pode inferir nenhuma conclusão entre os itens.

apenas pontuações positivas para os itens observados. Os elementos restantes, incluindo aqueles que podem ser de interesse para o usuário, serão classificados pelo modelo como pontuações negativas e a classificação não poderá ser otimizada, pois as previsões estarão em torno de zero.

Os autores propuseram um método genérico para aprender o comportamento do usuário para classificação personalizada (RENDLE *et al.*, 2012). Em vez de treinar o modelo usando apenas os pares usuário-item, eles também consideraram a ordem relativa entre um par de itens, de acordo com as preferências do usuário. Se um item i foi visualizado pelo usuário u e j não ($i \in N(u)$ e $j \in \overline{N}(u)$), então i é preferido a j. A Figura 3 mostra um exemplo do método. Quando i e j são desconhecidos para o usuário, ou equivalentemente, ambos são conhecidos, nenhuma conclusão sobre sua importância relativa para o usuário pode ser inferida.

Para estimar se um usuário prefere um item a outro, (RENDLE *et al.*, 2012) propuseram uma análise Bayesiana usando uma função de probabilidade $prob(i \succ_u j|u,\Theta)$ e a probabilidade anterior para o parâmetro do modelo $prob(\Theta)$. O critério final de otimização, BPR-Opt, é definido como:

$$BPR\text{-}Opt \sum_{(u,i,j)\in S_K} \ln \sigma(\hat{s}_{uij}) - \Lambda_\Theta \|\Theta\|^2$$

onde $\hat{s}_{uij}\hat{r}_{ui} - \hat{r}_{uj}$ e S_K é o conjunto de triplas (u,i,j) onde i está em $N(u)$ e j não está. O símbolo Θ representa os parâmetros do modelo, Λ_Θ é o conjunto de constantes de regularização, e σ é a função logística definida como $\sigma(x) = \frac{1}{1+e^{-x}}$.

Os autores também propuseram uma variação na técnica de descida de gradiente estocás-

tico, denominada LearnBPR, que amostra aleatoriamente de S_K para ajustar Θ. O Algoritmo 1 mostra uma visão geral do método de aprendizagem, onde α é a taxa de aprendizado.

Algoritmo 1 – Aprendizado via LearnBPR.

D_K Parâmetros ajustados Θ
Inicializar Θ com valores aleatórios

para cont = 1,...,#Iterações **faça** obtenha (u,i,j) a partir de S_K
$\hat{s}_{uij} \leftarrow \hat{r}_{ui} - \hat{r}_{uj}$
$\Theta \leftarrow \Theta + \alpha \left(\frac{e^{-\hat{s}_{uij}}}{1+e^{-\hat{s}_{uij}}} \cdot \frac{\partial}{\partial \Theta} \hat{s}_{uij} - \Lambda_\Theta \Theta \right)$

No presente estudo, definimos a abordagem BPR para considerar a regra de predição $\hat{r}_{ui}$ do modelo de fatorização simples definido na Equação 4.4. Portanto, aplicar a Equação 4.4 em $\hat{s}_{uij}$ resulta em $\Theta = \{p_u, q_i, q_j\}$, que devem ser aprendidos. Calculamos as derivadas parciais em relação a $\hat{s}_{uij}$:

$$\frac{\partial}{\partial \Theta} \hat{s}_{uij} = \begin{cases} q_i - q_j & \text{quando } \Theta = p_u \\ p_u & \text{quando } \Theta = q_i \\ -p_u & \text{quando } \Theta = q_j \\ 0 & \text{caso contrário} \end{cases}$$

Esses gradientes são então usados para atualizar os fatores de usuário e item em direção ao mínimo da função de perda, iterativamente, até que a convergência seja alcançada ou um número fixo de iterações seja concluído. Desse modo, o SGD permite ajustar os fatores de usuário e item de forma a maximizar a diferença entre as pontuações dos itens positivos e negativos, resultando em recomendações mais precisas e personalizadas.

4.2.3 BPR com Calibração por Popularidade

Com o objetivo de combinar um modelo em etapa de processamento com uma forma de calibração de popularidade para trazer recomendações que reduzam o viés de popularidade no sistema, foi pensado em alterar o algoritmo de aprendizado *LearnBPR* (Algoritmo 1). Assim, pode-se combinar o BPR com a calibração por popularidade, acrescentando no algoritmo a divergência de Kullback-Leibler implementada na calibração de popularidade.

Essa combinação pode possibilitar uma maior justiça nas recomendações em termos de popularidade, já que esse aspecto seria levado em conta na função de perda do BPR. A alteração é realizada na linha 5 do Algoritmo 1 somente quando $\Theta = p_u$:

$$\begin{aligned} p_u \leftarrow p_u + \alpha \Bigg(& \frac{e^{-\hat{s}_{uij}}}{1+e^{-\hat{s}_{uij}}} . (q_i - q_j) \\ & + \lambda \left(1 - \frac{D_{KL}(x \| y)}{D_{KLvoid}} \right) - \Lambda_{p_u} p_u \Bigg) \end{aligned} \tag{4.5}$$

onde λ é utilizado como coeficiente do impacto que a divergência terá no sistema, e D_{KLvoid} é definido como:

$$D_{KLvoid} = \sum_{t} x(t|u) \cdot log \frac{x(t|u)}{\alpha \cdot x(t|u)} \tag{4.6}$$

A razão para dividir a divergência $D_{KL}(x,y)$ por D_{KLvoid} é normalizar o valor da divergência, deixando o valor ajustado para uma escala específica. Essa normalização pode ser útil para realizar a calibração por popularidade entre diferentes usuários ou grupos, independentemente do número total de itens ou da escala de popularidade na base de dados, possibilitando a aplicação em diferentes contextos.

Ao considerar não apenas as preferências individuais dos usuários, mas também a popularidade relativa dos itens, a abordagem modificada pode levar a recomendações mais relevantes e personalizadas. Isso pode resultar em uma melhor experiência do usuário e maior satisfação com o sistema de recomendação.

4.2.4 Configuração do Experimento

A execução do experimento foi feita de forma *offline* utilizando dois conjuntos de dados do domínio de filmes. A Tabela 8 resume as informações dos conjuntos de dados utilizados.

- **Yahoo Movies**[1]: Este conjunto de dados é uma classificação de filmes do usuário, onde o usuário atribui classificações de um a cinco aos filmes que assistiu. Na etapa de pré-processamento, foram removidos apenas filmes sem gênero nos metadados. Em vez de binarizar a classificação como feito por (STECK, 2018), foi utilizado o *feedback* explícito como o peso $w_r(u,i)$ na Equação 4.1

- **MovieLens-20M**[2]: Neste conjunto de dados, semelhante a (STECK, 2018) e em contraste com o conjunto de dados do Yahoo Movies, foi feita a binarização das classificações retendo as interações onde a classificação era superior a 4. Além disso, devido a limitações de hardware, o tamanho do conjunto de dados foi reduzido, removendo filmes com menos de dez interações e usuários com menos de 180 filmes.

Tabela 2 – Estatísticas dos conjuntos de dados após realização do pré-processamento.

Conjunto de dados	# Usuários	# Avaliações	# Itens
Yahoo Movies	7,642	211,231	11,916
MovieLens 20M	12,603	3,984,599	10,417

[1] https://webscope.sandbox.yahoo.com/
[2] https://grouplens.org/datasets/movielens/20m/

O experimento foi executado três vezes em cada conjunto de dados para obter a média dos valores gerados pelas métricas e garantir a estabilidade dos resultados. Os conjuntos de dados de teste e treinamento foram escolhidos dividindo aleatoriamente o conjunto de dados em 70/30% de interações, seguindo respectivamente (ABDOLLAHPOURI *et al.*, 2021; SILVA; MANZATO; DURÃO, 2021). O desempenho da abordagem foi comparado com os seguintes trabalhos do estado da arte:

1. ***PairWise***: Proposto por (BORATTO; FENU; MARRAS, 2021), este método atua como uma etapa de processamento para redução de popularidade. Para o conjunto de dados do Yahoo Movies, foi aplicado $epoch = 100$, $batch = 1024$ e escolhido o melhor α variando no intervalo $[0,1]$. Para o conjunto de dados MovieLens, foi utilizado $batch = 2048$ e $epoch = 20$. A implementação seguiu aquela feita pelos autores[3].

2. **BPR**: Proposta em (RENDLE *et al.*, 2012), é um algoritmo de recomendação projetado para lidar com dados de feedback implícito, onde as interações entre usuários e itens são representadas como preferências binárias. Para os dois conjuntos de dados, foi aplicado o $batch$ = 1024.

4.3 Resultados

4.3.1 Yahoo Movies

Tabela 3 – Comparação da abordagem proposta com os outros trabalhos no conjunto de dados Yahoo Movies. O símbolo ▲ significa que a proposta teve um ganho significativo com relação aos outros trabalhos, com um $p\text{-}value < 0.05$ usando o *t-test* de Student; O símbolo • significa que não houve um ganho ou perda significativo; e o símbolo ▼ indica que o outro trabalho é estatisticamente melhor que a proposta. Cada par de símbolos se refere ao BPR e ao *PairWise*, respectivamente.

Algoritmo	LTC	MRMC Gêneros	MRMC Pop.	F1 Score	MRR	MAP	ΔGAP_{BB}	ΔGAP_{N}	ΔGAP_{D}	RMSE
BPR	0.409	0.629	0.687	0.340	0.002	0.001	-0.991	-0.881	-0.978	0.549
PairWise	0.140	0.696	0.661	0.321	0.012	0.038	-0.680	3.105	0.043	1.060
BPR Modificado	0.317 ▼▲	0.589	0.496	0.444 ▲▲	0.012 ▲•	0.004 ▲▼	-0.934	-0.142	-0.835	0.420 ▲▲

A Tabela 3 apresenta os resultados obtidos para o conjunto de dados Yahoo Movies. Analisando apenas a **precisão** dos modelos pela métrica MAP, notamos que a abordagem *PairWise* (BORATTO; FENU; MARRAS, 2021) atingiu o maior valor de MAP. No entanto, esta conquista significa que os itens não são muito diversos entre si, como mostram os seus resultados relativos a LTC, F1 e RMSE.

Em relação à **justiça dos gêneros** através do MRMC de gêneros, a Tabela 3 indica que a proposta de calibração combinada com o BPR produziu o melhor resultado, indicando que foi

[3] https://github.com/biasinrecsys/wsdm2021

Tabela 4 – Comparação da abordagem proposta com os outros trabalhos no conjunto de dados Movie Lens 20M. O símbolo ▲ significa que a proposta teve um ganho significativo com relação aos outros trabalhos, com um *p-value* < 0.05 usando o *t-test* de Student; O símbolo • significa que não houve um ganho ou perda significativo; e o símbolo ▼ indica que o outro trabalho é estatisticamente melhor que a proposta. Cada par de símbolos se refere ao BPR e ao *PairWise*, respectivamente.

Algoritmo	**LTC**	**MRMC Gêneros**	**MRMC Pop.**	**F1 Score**	**MRR**	**MAP**	ΔGAP_{BB}	ΔGAP_{N}	ΔGAP_{D}	**RMSE**
BPR	0.513	0.459	0.409	0.565	0.001	0.001	-0.912	-0.340	-0.790	0.419
PairWise	0.110	0.554	0.501	0.452	0.776	0.583	-0.997	-0.997	-0.996	0.575
BPR Modificado	0.464 ▼▲	0.453	0.330	0.596 ▲▲	0.002 ▲▼	0.001 •▼	-0.865	-0.060	-0.693	0.370 ▲▲

capaz de fornecer os mais equitativos itens em termos de gênero. O mesmo foi verificado em relação à **justiça de popularidade**, com a proposta tendo o melhor resultado do MRMC Pop.

Em termos de **cobertura de cauda longa**, a tabela indica que o modelo mais eficaz foi o BPR. O *PairWise* com pontuações mais altas no MAP obteve valores mais baixos para o LTC. Em relação à métrica **F1**, é possível observar que a proposta conseguiu alcançar o melhor resultado, indicando que a abordagem de calibração foi capaz de calibrar recomendações de acordo com gêneros e popularidade. Este aspecto é ainda validado ao analisar a métrica **RMSE**, onde a mesma abordagem obteve menor erro com a calibração, indicando que ela aborda os pontos de justiça mencionados e reduz o viés de popularidade do sistema.

Os resultados relatados na Tabela 3 mostram que a abordagem de calibração foi capaz de equilibrar recomendações de acordo com gêneros e popularidade, em oposição aos outros trabalhos, que são mais adequados para um único aspecto, como precisão, gêneros ou popularidade.

A Tabela 3 demonstra a importância de adotar métricas além da precisão na análise de algoritmos de recomendação. Reconhece-se a alta precisão do *PairWise*, conforme indicado pela métrica MAP. No entanto, os usuários que preferem itens de nicho, diversos e impopulares são afetados por recomendações injustas e tendenciosas produzidas por essas abordagens.

4.3.2 Movie Lens 20M

A Tabela 4 apresenta os resultados obtidos para o conjunto de dados do Movie Lens 20M. Analisando a **precisão**, assim como na base de dados anterior, o *PairWise* (BORATTO; FENU; MARRAS, 2021) superou as outras abordagens. No entanto, os resultados também indicam que estas abordagens devolvem recomendações injustas em termos de gênero e popularidade, e carecem de diversidade.

Em relação à **justiça dos gêneros** e à **justiça de popularidade**, a abordagem de calibração proposta obteve os melhores resultados, fato confirmado pela métrica F1 Score. Em relação à **cobertura de cauda longa**, o BPR obteve o melhor resultado entre todas as abordagens. Além disso, o *PairWise* (BORATTO; FENU; MARRAS, 2021) alcançou um valor baixo para essa métrica, apesar de ter uma alta precisão.

Com relação à **F1 Score**, pode-se observar que a proposta obteve os melhores valores, destacando seu alto desempenho em termos de justiça nos gêneros e popularidade. Ademais, a proposta também obteve o melhor resultado em **RMSE**, indicando que o sistema reduziu com sucesso o viés de popularidade para diferentes grupos de usuários.

A Tabela 4 relata resultados semelhantes aos do conjunto de dados Yahoo Movies, indicando que a proposta melhorou a justiça dos gêneros e a popularidade em ambos os conjuntos de dados. Embora a abordagem de calibração proposta não tenha alcançado alta precisão, obteve o menor erro de calibração de gênero e de popularidade, o que significa que o modelo fornece recomendações que respeitam o perfil do usuário tanto no gênero quanto no consumo de popularidade.

4.4 Considerações Finais

O objetivo do BPR é aprender representações latentes para usuários e itens que capturem suas preferências individuais. O procedimento de aprendizado do BPR envolve a otimização de uma função de perda que visa maximizar a ordenação correta dos pares de itens positivos e negativos para cada usuário. Isso é feito por meio de gradiente descendente estocástico, onde os gradientes da função de perda são calculados para atualizar os vetores latentes dos usuários e dos itens.

A modificação proposta, que combina o BPR com a calibração por popularidade, visa melhorar a justiça nas recomendações, considerando não apenas as preferências individuais dos usuários, mas também a popularidade relativa dos itens. Isso é alcançado incorporando a divergência de Kullback-Leibler na função de perda do BPR, levando a recomendações mais relevantes e personalizadas. Os experimentos realizados em dois conjuntos de dados diferentes mostram que a abordagem modificada obtém resultados comparáveis ou melhores em relação aos métodos do estado da arte, tanto em métricas de classificação quanto em métricas de popularidade e justiça.

No entanto, é importante ressaltar que a abordagem proposta ainda pode ser aprimorada em vários aspectos. Por exemplo, a escolha dos parâmetros do modelo, como o tamanho do lote e o número de épocas, pode afetar significativamente o desempenho do sistema. Além disso, a implementação de técnicas adicionais de regularização ou otimização pode ajudar a evitar o *overfitting* e melhorar a convergência do modelo. Futuras pesquisas podem explorar essas direções para desenvolver ainda mais a abordagem proposta e melhorar sua eficácia em uma variedade de cenários de recomendação.

Com relação à questão de pesquisa **RQ-1**, os resultados apresentados mostram que a precisão do sistema se manteve praticamente igual ou ligeiramente superior ao BPR original em ambas as bases. Com relação a cobertura de cauda longa nas duas bases, a proposta teve resultados ligeiramente inferiores aos do BPR original, porém foram superiores ao *PairWise*.

Já no que se refere à questão de pesquisa **RQ-2**, o experimento mostrou resultados promissores na redução do viés de popularidade do sistema. Nas duas bases, foi demonstrado que as métricas de justiça em termos de gênero e popularidade foram melhores na abordagem proposta. Além disso, a abordagem também foi superior na métrica RMSE, indicando que reduziu com sucesso o efeito do viés para diferentes grupos de usuários.

CAPÍTULO

5

UMA ABORDAGEM *NUDGE*

O capítulo anterior apresentou uma abordagem que tinha o objetivo de verificar o impacto da calibração em etapa de processamento. Nesse sentido, o presente capítulo visa analisar uma abordagem em nível de interface que, por meio do mecanismo de *nudges*, tenta incentivar os usuários a interagirem com itens diferentes de seus perfis, de modo a promover a diversidade.

Assim, este capítulo descreve de forma detalhada a abordagem que utiliza *nudges* para redução do viés de popularidade e é organizado da seguinte forma: a Seção 5.1 lista a justificativa de realização dessa abordagem; a Seção 5.2 descreve a metodologia e o design do experimento; por fim, os resultados são apresentados e discutidos na Seção 5.3.

5.1 Justificativa

Os Sistemas de Recomendação vão muito além de somente recomendar itens para os usuários, eles podem ser incrementados de modo persuasivo para que o usuário fique mais tempo interagindo com o sistema e, assim, traga mais retornos para seus clientes. Dessa forma, os sistemas podem ser ser planejados com o objetivo de influenciar o comportamento do usuário (YOO; GRETZEL; ZANKER, 2012). Nos sistemas digitais, isso pode ser feito por meio de interfaces chamadas *nudges* (WEINMANN; SCHNEIDER; BROCKE, 2016).

Nudge é um conceito introduzido por (LEONARD, 2008) e que envolve guiar sutilmente o usuário para um comportamento específico, sem que isso limite suas opções. Ele pode ser implementado de diferentes maneiras em sistemas tecnológicos, seja sugerindo itens similares, indicando itens que usuários com o mesmo perfil também gostaram, ou até mesmo com indicativos de que aquele item é o último restante (CARABAN *et al.*, 2019).

Sabendo disso, os *nudges* podem ser utilizados também para auxiliar o usuário a interagir com itens diferentes do seu perfil, possibilitando a descoberta de itens interessantes além de suas preferências. Nesse sentido, foi planejado um experimento para medir o impacto dos *nudges* para

auxiliar o usuário a interagir com itens diversos e como eles podem ajudar a reduzir o viés de popularidade de um sistema. Este experimento foi publicado na revista *UMUAI - User Modeling and User-Adapted Interaction 2022* (ALVES *et al.*, 2023).

5.2 Metodologia

Para desenvolvimento do experimento, um estudo empírico foi realizado com usuários para medir o impacto que os *nudges* têm para influenciar usuários a interagirem mais com conteúdo diverso. Foi proposto um aplicativo desenvolvido para Android, utilizando o domínio de livros, de forma a recomendar alguns livros para os usuários, tendo duas versões do mesmo aplicativo para análise, uma versão de controle e outra versão de tratamento. Em ambas as versões eram apresentadas aos usuários uma lista de livros de acordo com os gêneros preferidos que eles escolhiam na tela inicial do sistema.

A versão de controle não tinha *nudges*, mas a versão de tratamento apresentava *nudges* nas recomendações para indicar os itens diversos. Na versão de tratamento, foram utilizados quatro tipos de *nudges* com base na literatura. O sistema funcionava da mesma forma nas duas versões, de modo que era apresentada uma lista de 24 livros recomendados para os usuários, sendo que os livros eram sempre alternados, de maneira que sempre haveria um livro do gênero que o usuário gosta, sendo o livro seguinte de gênero diferente dos favoritos do usuário. Para a versão de tratamento, os itens diversos eram sempre indicados com um *nudge* para atrair o usuário e medir o impacto desse mecanismo.

Durante o experimento, o objetivo do usuário foi favoritar os itens da lista de filmes que mais lhe interessavam. Dessa forma, foi possível analisar as interações do usuário no aplicativo, verificando se eles interagiam mais com itens diversos ou itens de gêneros de suas preferências, bem como analisar até qual posição dos itens da lista os usuários visualizavam e o tempo de cada usuário e suas ações no sistema.

Além da análise das interações, um formulário foi enviado para cada participante ao término do experimento para que fosse possível obter mais informações sobre a percepção dos usuários com relação ao sistema proposto. Esse formulário foi desenvolvido seguindo o framework *ResQue* (PU; CHEN; HU, 2011), que é muito utilizado para avaliar a percepção do usuário em sistemas de recomendação.

Para a execução do experimento, a lista de livros foi montada a partir da escolha dos 6 gêneros dos livros mais populares da página *Amazon.com.br*, sendo selecionados alguns livros de autores mais novos, de forma a aumentar a descoberta de livros e não ficar restrito somente aos mais conhecidos.

Os participantes do experimento foram recrutados por meio de uma página de livros de rede social, assim, houve um total de 1064 usuários que participaram do experimento, sendo que

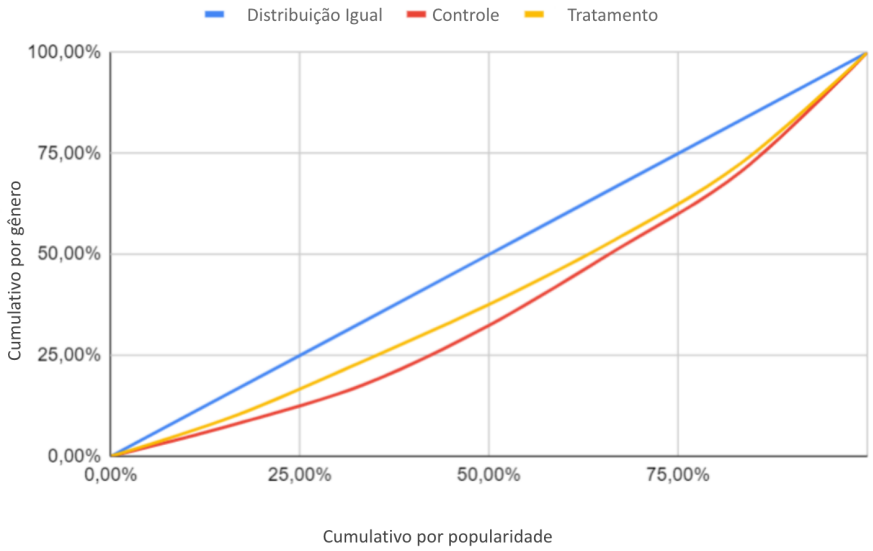
520 foram do grupo de tratamento e 544 foram do grupo de controle. O Apêndice A contém as imagens do aplicativo desenvolvido para realização do experimento.

5.3 Resultados

Uma das análises feitas sobre esse experimento foi com relação à verificação do impacto que os *nudges* tiveram para a redução do viés de popularidade no sistema. Para isso, calculamos o coeficiente de Gini referente à popularidade dos gêneros dos livros selecionados pelos usuários para os grupos de tratamento e de controle, conforme feito em (JANNACH *et al.*, 2015).

O índice de Gini é uma métrica que fica entre zero e um e indica o quão desequilibrado os dados estão distribuídos, e corresponde à razão de igualdade perfeita e desigualdade perfeita, com números mais altos indicando maior concentração. Observamos, conforme a Figura 4, que a curva para o grupo de tratamento está mais próxima da linha perfeita, indicando que as preferências de gênero estão distribuídas de forma mais uniforme no grupo de tratamento com os *nudges*.

Figura 4 – Curva de Lorenz que indica a distribuição de popularidade nas duas versões do experimento.



Fonte: Elaborada pelo autor.

O índice de Gini para o grupo de tratamento é 0,17 e o do grupo controle é 0,23, este último indicando maior concentração. Isso significa que houve uma contribuição dos *nudges* para uma melhora na distribuição dos itens, podendo ser considerada como uma redução no viés de popularidade.

Outro aspecto interessante obtido nos resultados do experimento foi com relação ao número de interações dos usuários para realizar a tarefa do experimento, que era favoritar cinco livros que lhe interessaram. Como a Tabela 5 mostra, o número de interações com itens de gêneros

diferentes das preferências dos usuários no grupo de tratamento foi maior do que no grupo de controle, fato que foi comprovado como estatisticamente significante a partir da realização do teste qui-quadrado. Isso indica que os *nudges* ajudaram a interagir com itens de gêneros diversos. Esse ponto também é verificado pela Tabela 6, que mostra que no grupo de tratamento houve um número maior de interações com ao menos um item diferente das preferências dos usuários.

Tabela 5 – Número absoluto de livros favoritados pelos usuários nos grupos de controle e tratamento.

	Controle	Tratamento	Σ
Itens do gênero preferido	1,718	1,535	**3,253**
Itens de gênero diferente das preferências do usuário	751	873	**1,624**
Σ	**2,469**	**2,408**	**4,877**

Fonte: Elaborada pelo autor.

Tabela 6 – Número de usuários que colocaram ao menos um item de gênero diferente de suas preferências na lista de favoritos.

	Controle	Tratamento	Σ
Ao menos um item diferente de suas preferências	407 (74,82 %)	440 (84,62 %)	**847**
Nenhum item diferente de suas preferências	137 (25,18 %)	80 (15,38 %)	**214**
Σ	**544**	**520**	**1,064**

Fonte: Elaborada pelo autor.

5.4 Considerações Finais

O propósito dessa abordagem foi verificar se os *nudges* podem auxiliar o usuário a interagir com itens diversos e se eles ajudam a reduzir o viés de popularidade. Para isso, foi conduzido um experimento aleatório controlado com um grupo de tratamento, que recebeu itens com *nudges*, e um grupo de controle sem essa interface. As análises estatísticas evidenciam que os *nudges* encorajam os usuários a interagirem com conteúdos fora do perfil, o que pode ajudar a reduzir o efeito do viés de popularidade.

Com relação à questão de pesquisa **RQ-1**, pode-se afirmar que os *nudges* não alteram a geração das recomendações de fato; ao invés disso, eles podem impactar visualmente as recomendações geradas, podendo ser utilizados para destacar itens específicos. No experimento, os *nudges* ajudaram a promover itens diversos, levantando mais possibilidades de utilizações desse tipo de interface em Sistemas de Recomendação.

Por sua vez, quanto a questão de pesquisa **RQ-2**, o experimento mostrou resultados promissores na interação do usuário com itens da cauda longa. Isso pode servir de motivação para experimentos futuros, apontando-se uma boa possibilidade de contribuição dos *nudges* para

a redução do viés de popularidade, já que ele pode auxiliar a deixar a interação com os itens mais equilibrada.

CAPÍTULO

6

UMA ABORDAGEM DE CALIBRAÇÃO EM PÓS-PROCESSAMENTO

A abordagem que utiliza o mecanismo de *nudges* apresentada no capítulo anterior foi interessante para validar o comportamento de usuários reais em relação a recomendações diversificadas geradas pelo sistema. Esse fato motivou a realização de um novo experimento para analisar o impacto na percepção do usuário de uma das técnicas mais conhecidas de justiça e redução de vieses, que é a calibração.

Este capítulo descreve de forma detalhada a abordagem que utiliza a calibração em etapa de pós-processamento e é organizado da seguinte forma: a Seção 6.1 apresenta a justificativa de realização dessa abordagem; a Seção 6.2 descreve a metodologia e o design do experimento; por fim, os resultados são apresentados e discutidos na Seção 6.3.

6.1 Justificativa

Uma abordagem técnica proeminente na literatura para neutralizar vieses é chamada calibração (STECK, 2018; JUGOVAC; JANNACH; LERCHE, 2017). Nessa abordagem, o objetivo é garantir que as recomendações feitas pelo sistema correspondam às preferências anteriores do usuário em termos de distribuição de determinadas características dos itens. Sendo assim, a abordagem de calibração foi adotada para a realização de um experimento com usuários com o objetivo de verificar a percepção do usuário com relação à justiça do sistema.

Ademais, um levantamento recente aponta que os estudos relacionados à justiça em Sistemas de Recomendação dependem muito de experimentos *offline*. Embora experimentos assim tragam *insights* interessantes, eles não nos informam como os usuários realmente perceberiam as recomendações otimizadas para justiça. Desse modo, foi feito um experimento com usuários (N = 500) para preencher essa lacuna.

Neste estudo, os participantes do grupo de controle receberam recomendações baseadas em um algoritmo comum de filtragem colaborativa independente de imparcialidade. Em contraste, as recomendações para os participantes em diferentes grupos de tratamento foram calibradas para aumentar a justiça em termos de duas características dos itens, ou seja, nas listas calibradas, alguns itens da lista original foram substituídos algoritmicamente por alguns "itens de justiça".

No experimento, foi explorado até que ponto o fornecimento de informações relacionadas com a equidade pode impactar a percepção das recomendações. No geral, o estudo revelou que a calibração de justiça foi amplamente eficaz no sentido de que os participantes nos grupos de tratamento selecionaram um dos itens de justiça fornecidos como a sua escolha favorita, numa medida que é principalmente proporcional ao tamanho da intervenção de calibração. Igualmente importante, a análise de um questionário pós-tarefa mostrou que o processo de calibração não impactou negativamente a percepção de qualidade das recomendações pelos usuários. Em suma, isto indica a viabilidade de aumentar a justiça sem comprometer a qualidade das recomendações, ao menos no domínio específico do estudo.

No entanto, nosso estudo também revelou desafios cruciais ao implementar uma abordagem de calibração na prática, particularmente na seleção de parâmetros adequados para o processo de calibração (KLIMASHEVSKAIA *et al.*, 2023). Além disso, e ainda mais importante, a intervenção de calibração não teve impacto na percepção de justiça dos participantes, ou seja, quando questionados, eles, em média, não consideraram as recomendações calibradas mais justas do que as não calibradas, a menos que fossem fornecidas informações explicativas adicionais. No mais, uma análise do feedback qualitativo dos participantes indicou que eles interpretam o conceito de justiça de maneiras bastante diferentes, enfatizando os desafios da investigação relacionada com a justiça em sistemas de recomendação. O experimento foi detalhado em um artigo submetido e aprovado para a conferência UMAP (*ACM Conference on User Modeling, Adaptation and Personalization*) (ALVES *et al.*, 2024).

6.2 Metodologia

A execução do experimento foi realizada por meio de um *website* criado especificamente para isso, utilizando o domínio de filmes. Depois de ler as instruções e fornecer consentimento para a realização do experimento, as preferências do usuário foram adquiridas pedindo aos participantes que fornecessem classificações para vários filmes. Com base nessas preferências, um conjunto de recomendações foi apresentado aos participantes, que foram solicitados a selecionar exatamente um filme que gostariam de assistir em seguida.

O conjunto de recomendações fornecidas variou entre diferentes grupos de participantes, onde os grupos de tratamento receberam diferentes tipos de recomendações calibradas para imparcialidade e o grupo de controle recebeu uma lista de recomendações com precisão otimizada. Após fazerem uma escolha, os participantes foram solicitados a responder perguntas, por exemplo,

sobre suas percepções subjetivas em uma série de questionários.

Durante a etapa de construção do perfil, os participantes interagiram com uma interface com múltiplas abas: uma era uma aba "geral", mostrando filmes de diferentes gêneros, e as outras abas mostravam filmes de gêneros específicos. Cada guia exibia 12 filmes. Além disso, uma funcionalidade de pesquisa estava presente. Os participantes foram instruídos a avaliar no mínimo 7 filmes de uma a cinco estrelas. Para cada filme, a aplicação forneceu informações essenciais, incluindo o título, uma imagem de pôster e o ano de lançamento. Os filmes foram selecionados a partir do conjunto de dados de classificação MovieLens-20M, que também usamos para gerar recomendações personalizadas com a ajuda de um algoritmo de filtragem colaborativa na próxima etapa.

Para nosso estudo, pré-processamos o conjunto de dados excluindo filmes sem informações de gênero. Ao coletar dados de preferência, buscamos minimizar o viés de popularidade, evitando deliberadamente os filmes convencionais em favor de filmes de média popularidade e de nicho. A metodologia de seleção foi equilibrada: 30% dos filmes foram selecionados aleatoriamente na categoria intermediária, enquanto os 70% restantes vieram da cauda longa menos popular.

A distribuição dos grupos de itens em termos de popularidade e orçamento foi feita com base na divisão descrita em (ABDOLLAHPOURI *et al.*, 2019b) e também detalhada na métrica GAP, explicada no Capítulo 2. Assim, os itens são divididos nos grupos *Head* (populares), *Mid* (diversos) e *Tail* (nicho). Essa divisão foi aplicada tanto no aspecto do orçamento do filme, quanto na popularidade dele. Assim, um orçamento de nicho é equivalente a um filme de baixo orçamento, enquanto que, um item com orçamento classificado como popular, teria um grande orçamento para produção.

No mais, esses filmes tiveram uma distribuição uniforme em termos de orçamento na tela inicial de escolha de filmes: um terço dos itens eram de alto orçamento (principal), um terço dos itens eram de médio e um terço dos itens eram de baixo orçamento. Para promover a imparcialidade, randomizamos a ordem de exibição desses 12 filmes em cada guia. Esta estratégia foi concebida para dar a cada filme, independentemente da sua popularidade e orçamento, oportunidades iguais de ser visto e avaliado pelos participantes.

O SLIM (NING; KARYPIS, 2011) foi utilizado como modelo de recomendação e treinado com o conjunto de dados MovieLens. As razões para esta escolha são que o SLIM não apenas apresenta um desempenho muito bom para este conjunto de dados (ANELLI *et al.*, 2022), mas também nos permite fazer recomendações para novos perfis de usuários de uma forma computacionalmente eficiente. Como técnica de calibração, utilizou-se o método proposto por Steck (STECK, 2018), tendo a divergência de Kullback-Leibler como métrica de calibração. Entretanto, ao invés de calibrar por gênero, o sistema foi calibrado com base na popularidade dos itens ou pelo orçamento deles. Ou seja, a implementação basicamente consistiu em combinar o modelo SLIM para gerar as recomedações e depois calibrar as recomendações geradas por meio

do método do Steck.

Os participantes foram distribuídos aleatoriamente em diferentes grupos de tratamento. Os participantes do grupo de controle do experimento receberam as recomendações de precisão otimizada fornecidas pelo SLIM. O desenho do estudo incluiu os seguintes cinco grupos de tratamento, que receberam diferentes tipos de recomendações:

- **1 - Orçamento.** As recomendações foram calibradas para incluir também filmes com um orçamento de produção relativamente baixo.
- **2 - Popularidade.** As recomendações foram calibradas para conter também filmes menos populares.
- **3 - Orçamento + Explicação.** As recomendações foram calibradas para incluir também filmes com um orçamento de produção relativamente baixo, com uma tag explicando que o item foi alterado na lista de recomendação por conta da calibração.
- **4 - Popularidade + Explicação.** As recomendações foram calibradas para incluir também filmes com um orçamento de produção relativamente baixo, com uma tag explicando que o item foi alterado na lista de recomendação por conta da calibração.
- **5 - Controle + Explicação.** Mesmas recomendações do grupo controle, mas com mensagem explicativa do motivo do item ter sido recomendado.

Portanto, enquanto os tratamentos 1 e 2 permitem avaliar os efeitos da calibração quando o comportamento justo do sistema é uma caixa preta para os usuários, os tratamentos 3 a 5 são projetados para estudar se as explicações ajudam o usuário a visualizar que aquela recomendação trouxe justiça. O tratamento 5 deverá, em particular, ajudar a desvendar os efeitos das explicações e da calibração. As mensagens explicativas foram mostradas aos usuários no topo das listas de recomendações, orientadas verticalmente.

Para os participantes dos grupos de tratamento **3 - Orçamento + Explicação** e **4 - Popularidade + Explicação**, o texto explicativo dizia: “Observe que suas recomendações também podem incluir alguns filmes com orçamentos de produção menores ou alguns filmes menos conhecidos.” Para o grupo de tratamento **5 - Controle + Explicação** a explicação dizia: “Observe que suas recomendações também podem incluir alguns filmes com orçamentos de produção menores e filmes menos conhecidos”. O termo “justiça” não foi incluído nas explicações, já que isso poderia influenciar as percepções de justiça ao final. Foi uma tentativa para compreender se os próprios usuários relacionariam questões de orçamentos de produção e popularidade com justiça. O Apêndice B apresenta as questões que os participantes tinham que responder após executar o experimento e o Apêndice C apresenta as telas do site desenvolvido para a realização do experimento.

6.3 Resultados

Inicialmente, importante rememorar que os participantes tiveram que selecionar exatamente um filme que gostariam de assistir. Ao analisar as escolhas reais dos participantes, nos grupos de tratamento cerca de 20% dos itens selecionados eram itens de justiça, ou seja, 20% dos usuários escolheram um dos itens que foram adicionados através da calibração. Isto indica que os participantes consideraram os itens de justiça relevantes.

A Tabela 7 mostra os detalhes sobre as escolhas de itens calibrados do participante. Não observamos grande diferença entre os grupos quanto à tendência de selecionar itens calibrados. Isto também indica que o fornecimento de declarações explicativas curtas (**3 - Orçamento + Explicação** e **4 - Popularidade + Explicação**) podem não ter impactado grandemente a consciência dos participantes sobre a popularidade e aspectos orçamentários quando fizeram suas escolhas.

Em termos do critério de calibração, a Tabela 7 mostra que a calibração por popularidade direciona principalmente os participantes para os itens de nicho. Em contrapartida, ao considerar o orçamento de produção, a calibração levou os participantes também a selecionar itens populares e diversos, ou seja, produções mais onerosas. Isto é esperado dadas as distribuições de classificação para itens populares, diversos e itens de nicho que observamos na fase de obtenção de preferência dos usuários.

Tabela 7 – Porcentagem dos itens selecionados a partir da calibração pelos grupos de tratamento.

Grupo de Tratamento	Selecionado da Calibração	Head	Mid	Tail	Total
1 - Orçamento.	21%	5	9	3	17
2 - Popularidade	19%	0	1	15	16
3 - Orçamento + Explicação.	18%	6	6	3	15
4 - Popularidade + Explicação.	20%	3	3	11	17

A Tabela 7 comprova que o grupo de tratamento 3, que recebeu as recomendações com base no orçamento e com a *label* de explicação, foi o que teve menor seleção de itens gerados pela calibração. Já em relação a popularidade, o grupo que recebeu a *label* de explicação teve uma seleção ligeiramente maior de itens gerados pela calibração. Além disso, para o grupo de tratamento de popularidade com explicação, houve uma seleção mais diversa de itens da calibração.

Os resultados dos questionários preenchidos pelos participantes do experimento também foram analisados. Com relação aos grupos de tratamento **1 - Orçamento** e **5 - Controle + Explicação**, foi verificado que a percepção de justiça foi a mais baixa entre todos os grupos. Lembramos que o grupo de tratamento **1 - Orçamento** recebeu recomendações calibradas pelo orçamento, onde a calibração geralmente levou a uma inclusão e seleção mais fortes de produções de orçamento mais elevado. Isto aumentou a inclusão de tais filmes sem explicação, mas não

teve impacto mensurável sobre a percepção de justiça dos usuários.

O experimento trouxe comentários deixados pelos usuários que contribuíram para os resultados da pesquisa. Um usuário trouxe uma visão interessante sobre a necessidade do sistema se adaptar e trazer recomendações personalizadas: *"Uma recomendação parece justa quando considera não apenas o gosto pessoal, mas também o humor atual do usuário. Embora eu geralmente prefira filmes de ação e aventura, há momentos em que estou mais inclinado para uma comédia."*

Há também comentários interessantes relacionados à um sistema justo: *"A imparcialidade fica comprometida se as sugestões forem limitadas a filmes de grandes estúdios ou aqueles que recebem publicidade significativa"*. Outro usuário acrescentou: *"As recomendações são justas quando se alinham com meu histórico de exibição, mas tornam-se injustas se apresentarem apenas filmes conhecidos, ignorando a riqueza de filmes menos conhecidos, mas tematicamente semelhantes, de diversas culturas, como o cinema latino ou turco."*

Já um outro usuário destacou o problema da falta de diversidade no sistema: *"As escolhas de filmes sugeridas pela Inteligência Artificial podem ser problemáticas, pois muitas vezes reforçam as preferências existentes do utilizador. Por exemplo, alguém que assiste predominantemente comédias provavelmente receberá recomendações exclusivamente cômicas, perpetuando uma exposição restrita do conteúdo."*

Os questionários também possibilitaram identificar que as abordagens de calibração não tiveram um impacto negativo nas percepções de qualidade das recomendações para os participantes. Assim, os participantes não sentiram que as recomendações eram menos precisas do que no grupo de controle, apesar de terem sido trocadas, em média, 20% dos itens por itens de justiça. Esta conclusão sugere que a calibração da justiça pode ser eficazmente aplicada sem levar a efeitos negativos na percepção da qualidade do sistema.

6.4 Considerações Finais

O objetivo dessa abordagem foi verificar como o usuário percebe a calibração de um Sistema de Recomendação. O sistema proposto foi calibrado levando em conta os aspectos de popularidade dos itens e também o orçamento. Os resultados mostram que a calibração conseguiu ser eficaz e trouxe itens relevantes para os usuários, que interagiram com os itens gerados pela calibração.

Em relação à questão de pesquisa **RQ-1**, o sistema foi calibrado de forma a trazer recomendações coerentes com as preferências dos usuários, seja pelo nível de popularidade dos itens ou pelo orçamento dos itens do domínio de filmes. Desse modo, a estratégia alterou a geração das recomendações, reclassificando a lista retornada pelo modelo SLIM e trazendo recomendações mais justas, e que continuaram relevantes para o usuário.

Como o sistema foi calibrado em termos de popularidade nos grupos de tratamento, os resultados apontam que essa calibração ajudou a lidar com o viés de popularidade pelo fato de promover itens de nicho, ajudando os usuários a interagirem com itens que passariam despercebidos. Esse ponto atende a questão de pesquisa **RQ-2**, pois a abordagem se mostrou uma forma de tratar o viés de popularidade de modo a promover itens da cauda longa. Esse ponto contribui para o desenvolvimento de outras abordagens em etapa de pós-processamento.

CAPÍTULO

7

UMA ABORDAGEM DE CALIBRAÇÃO PERSONALIZADA

A abordagem de calibração em pós-processamento apresentada no Capítulo 6 validou que a calibração de um Sistema de Recomendação não interferiu na percepção da qualidade das recomendações geradas para o usuário. Com isso, decidiu-se explorar formas de melhorar a abordagem de calibração, por meio da incorporação de diferentes aspectos de justiça: enquanto lá se usava popularidade e orçamento (sendo uma opção para cada grupo do experimento), aqui será proposto uma forma de combinar isso, mas com experimentos *offline* e utilizando termos de popularidade e gênero.

Nesse sentido, este capítulo descreve de forma detalhada a abordagem que utiliza a calibração em etapa de pós-processamento e é organizado da seguinte forma: a Seção 7.1 apresenta a justificativa de realização dessa abordagem; a Seção 7.2 descreve a metodologia e o design do experimento; por fim, os resultados são apresentados e discutidos na Seção 7.3.

7.1 Justificativa

A calibração já é um tema bem discutido na área de recomendação, o trabalho de Steck (STECK, 2018), por exemplo, apresenta uma abordagem de calibração com base nos gêneros dos itens, de modo que o usuário recebe recomendações proporcionais aos seus interesses. Assim, caso o usuário tenha uma preferência de 70% de filmes de ação e 30% de filmes de terror, as recomendações geradas obedeceriam exatamente a essa proporção.

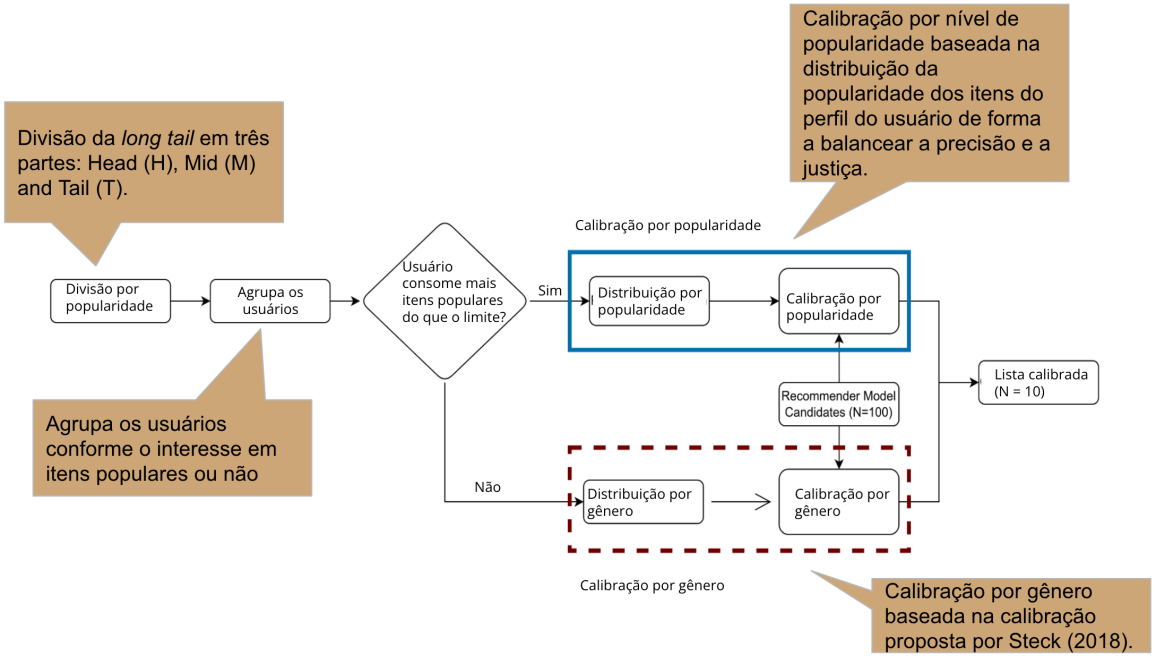
O trabalho (SILVA; MANZATO; DURÃO, 2021) também explora a calibração, de forma a adotar algumas medidas e pesos personalizados para equilibrar a relação entre precisão e calibração para que as recomendações atendam aos interesses dos usuários. Nesse sentido, há uma lacuna na área para tentar balancear a recomendação conforme o nível do interesse dos usuários pela popularidade dos itens. Assim, essa abordagem visa preencher essa lacuna,

apresentando um sistema de calibração baseado na popularidade dos itens e que tenta reduzir o impacto do viés de popularidade no sistema.

Essa abordagem foi publicada em um artigo da conferência ICEIS (SACILOTTI; SOUZA; MANZATO, 2023). Foi feita também uma submissão de um artigo que estende esse primeiro trabalho, adicionando a realização do experimento em uma base de outro domínio. Esse segundo trabalho já está aprovado e será publicado no livro da conferência ICEIS (SOUZA; MANZATO, 2024b).

7.2 Metodologia

Figura 5 – Fluxo da Abordagem de Calibração Personalizada



Fonte: Elaborada pelo autor.

O fluxo da abordagem de calibração personalizada funciona conforme a Figura 5. Em particular, a abordagem é dividida em dois métodos: **Calibração de Popularidade** (destacado em linha azul contínua na Figura 5), que estende a calibração de gênero proposta anteriormente por (STECK, 2018) (destacado na linha vermelha pontilhada na Figura 5); e **Calibração Personalizada** (Figura 5 como um todo), que usa calibração de gênero e calibração por popularidade em um modelo unificado para fornecer recomendações calibradas de acordo com popularidade e gêneros.

Para calibrar a lista de recomendações com base na popularidade dos itens consumidos pelo usuário no passado, foi feita uma divisão de popularidade para agrupar os itens com base na quantidade de interações. Dessa forma, caso o usuário consuma itens populares abaixo de um

limite estabelecido, é realizada somente uma calibração de gênero; caso contrário, é realizada uma calibração de popularidade para atender ao nível de preferência para este aspecto. A divisão de popularidade, introduzida em (SACILOTTI; SOUZA; MANZATO, 2023) e presente na Figura 5, é baseada no conceito de cauda longa dos sistemas de recomendação e é exatamente igual àquela apresentada no Capítulo 6.

Conforme indicado na Figura 5, o modelo unificado alterna entre popularidade e calibração de gênero. Para tomar essa decisão, é necessário agrupar os usuários de acordo com seus interesses em itens não populares/populares. Portanto, definimos o limite como a média de todas as proporções, que é um valor que pode ser facilmente calculado em cada conjunto de dados, conforme mostrado na Equação 7.1:

$$G_{threshold} = \frac{\sum_{u}^{U} \frac{\sum_{i}^{I_u} \mathbf{1}(i)}{|I_u|}}{|U|} \tag{7.1}$$

onde $\mathbf{1}(i)$ é uma função indicadora que retorna 1 se o item i, interagido pelo usuário u, estiver na categoria de popularidade **H**. Finalmente, assumimos que se a proporção de itens na categoria **H** for inferior a $G_{threshold}$, então devemos obter uma lista de recomendações calibrada por gênero; caso contrário, pela popularidade.

A calibração por popularidade foi uma adaptação da fórmula proposta por (STECK, 2018). Seu trabalho pressupõe que os itens podem ter mais de um gênero, o que não é válido no contexto de popularidade, onde um item possui apenas um nível de popularidade. Então, em vez disso, foram calculadas as somas dos pesos de cada tipo de popularidade sobre a soma de todos os pesos. Assim, $p(t|u)$ é a distribuição alvo baseada na popularidade dos itens com os quais o usuário interagiu no passado, e está definido como mostra a Equação 4.1, que é a mesma já apresentada no Capítulo 4.

Nesse contexto, o sistema é justo quando atende às proporções de popularidade esperadas pelo usuário. Portanto, caso o usuário consuma menos itens populares que o limite estabelecido, é realizada uma calibração baseada em gênero, pois, neste caso, o usuário não se preocupa com a popularidade dos itens. Caso o usuário consuma itens mais populares que o limite estabelecido, a calibração ocorre com base na popularidade, respeitando o nível de interesse do usuário neste aspecto.

Várias métricas avaliam a imparcialidade em sistemas de recomendação (VERMA; GAO; SHAH, 2020). Porém, nesse caso, utiliza-se a medida de divergência Kullback-Leibler pelas mesmas razões apontadas por (STECK, 2018) e exploradas por (SILVA; MANZATO; DURÃO, 2021). O Kullback-Leibler quantifica a desigualdade no intervalo $[0, \infty]$, onde 0 significa que ambas as distribuições são quase iguais e valores mais altos indicam injustiça.

Ademais, é adotada a regularização proposta por (STECK, 2018), que definiu $\alpha = 0.01$ como uma variável de regularização para evitar divisão por zero quando $q(t|u)$ vai para zero.

Embora existam outras métricas de divergência, como Hellinger e Person Qui-Square, propostas por (CHA, 2007) e explorado por (SILVA; MANZATO; DURÃO, 2021), foi utilizado apenas o Kullback-Leibler devido à sua simplicidade. Seu valor é o mesmo apresentado na Equação 4.3.

Neste experimento, a calibração é definida como o processo para encontrar o conjunto ótimo R_u^*, usando a relevância marginal máxima, conforme mostrado na Equação 7.2, onde D_{KL} é a função de justiça. Nesta formulação, quando $\lambda = 0$, focamos apenas nas pontuações de recomendação, e quando $\lambda = 1$, focamos em itens justos referentes ao perfil do usuário. A Figura 5 mostra o processo de calibração final.

$$R_u^* = \max_{CI_u} (1-\lambda) \cdot \sum_{i \in CI_u} wr_{u,i} - \lambda \cdot D_{KL}(p, q(CI_u)) \tag{7.2}$$

A execução do experimento foi feita de forma *offline* utilizando dois conjuntos de dados do domínio de filmes e um conjunto de dados do domínio de música. A Tabela 8 resume as informações dos conjuntos de dados utilizados.

- **Yahoo Movies**[1]: Esta base foi configurada da mesma maneira que no Capítulo 4.
- **MovieLens-20M**[2]: Este conjunto de dados foi configurado da mesma forma que no Capítulo 4.
- **Yahoo Songs**[3]: Este conjunto de dados é uma classificação de músicas do usuário, onde o usuário dá uma nota de um a cinco para as músicas que ouviu. Na etapa de pré-processamento, foram removidas músicas sem gênero nos metadados. Devido a limitações de hardware, foi reduzido o conjunto de dados, excluindo músicas com menos de 10 interações e usuários com menos de 10 músicas avaliadas.

Tabela 8 – Estatísticas dos conjuntos de dados após realização do pré-processamento.

Conjunto de dados	# Usuários	# Avaliações	# Itens
Yahoo Movies	7,642	211,231	11,916
MovieLens 20M	12,603	3,984,599	10,417
Yahoo Songs	2,817	680,460	22,196

O experimento foi executado três vezes, do mesmo modo que no Capítulo 4. O processo de calibração não depende do algoritmo do sistema de recomendação. Atua como uma etapa de pós-processamento onde, após o modelo prever os itens candidatos para um usuário, aplica-se a técnica de calibração descrita na Equação 7.2 para encontrar a melhor lista de itens para aquele usuário. Consequentemente, para entender o desempenho das abordagens de calibração sob

[1] https://webscope.sandbox.yahoo.com/
[2] https://grouplens.org/datasets/movielens/20m/
[3] https://webscope.sandbox.yahoo.com/

diferentes algoritmos de recomendação, foram usados quatro modelos bem conhecidos descritos abaixo, baseados nos trabalhos de (STECK, 2018) e (SILVA; MANZATO; DURÃO, 2021).

1. **SVD++**: Extensão (KOREN, 2008) para trabalhar com feedback implícito. Da mesma forma que (SILVA; MANZATO; DURÃO, 2021), foi utilizado $ne = 20$ como o número de épocas, $\gamma_u = \gamma_i = 0,005$ como a taxa de aprendizagem para usuários e itens, $\lambda_u = \lambda_i = 0,02$ como constantes de regularização e $f = 20$ fatores.

2. **NMF**: Fatoração de Matrizes Não-negativas proposta por (LUO *et al.*, 2014). Da mesma forma que (SILVA; MANZATO; DURÃO, 2021), usamos $ne = 50$, $\gamma_u = \gamma_i = 0,005$, $\lambda_u = \lambda_i = 0,06$ e $f = 15$.

3. **Item KNN**: Para implementar este algoritmo de filtragem colaborativa, foi adotada a abordagem KNNWithMeans de (HUG, 2020), usando $k = 30$ vizinhos mais próximos. Além disso, foi especificado o coeficiente de correlação de Pearson como métrica para similaridade de itens.

4. **SlopeOne**: Algoritmo de filtragem colaborativa, cuja implementação foi baseada na biblioteca *Surprise* (HUG, 2020) com valores padrão para os parâmetros.

O experimento para cada sistema de recomendação consiste em utilizar os dados de treinamento para alimentar o modelo e aprender as preferências do usuário com base nos itens interagidos no passado, representados como I_u. Após a etapa de treinamento, é feita a previsão de todas as classificações ausentes e, para cada usuário, são selecionados os 100 principais itens com a classificação prevista mais alta, representados como CI_u. O peso $w_r(u,i)$ é definido como a classificação que o algoritmo previu para o item candidato. Por fim, a lista final de recomendações R_u^* é criada com os 10 principais itens fornecidos pelo processo de calibração.

O experimento analisou separadamente o desempenho da **calibração de popularidade** e da **calibração personalizada**. Para o ***trade-off*** entre métricas de similaridade e justiça, na Equação 7.2, foram adotados os valores descritos por (STECK, 2018), variando de $\lambda \in [0, 0.1, 0.2, \ldots, 1]$. O desempenho da abordagem foi comparado com os seguintes trabalhos do estado da arte:

1. **Calibração por gêneros**: Proposto por (STECK, 2018), este método implementa uma técnica de calibração para gêneros. O método é comparado usando o mesmo conjunto de quatro algoritmos de recomendação (SVD++, NMF, ItemKNN e SlopeOne).

2. **CP**: Proposto por (ABDOLLAHPOURI *et al.*, 2021), este método implementa uma técnica de calibração de popularidade, mas usando a métrica de divergência de Jensen-Shannon para comparar as distribuições de perfil e recomendação. Este método também é comparado usando o mesmo conjunto de quatro algoritmos de recomendação.

7.3 Resultados

A Tabela 9 compara o desempenho das abordagens de calibração personalizada e de popularidade propostas com os dois métodos da literatura usando a base de dados do Yahoo Movies. Para cada combinação entre recomendação e calibração, a tabela também mostra o peso de *trade-off* selecionado λ de acordo com o melhor valor de LTC. Comparando a proposta de calibração personalizada com a calibração CP, é possível observar que ela obteve melhores valores para o LTC e F1 Score em todos os casos, indicando uma lista mais diversificada e calibrada de acordo com as preferências do usuário.

Em relação à justiça entre os três grupos de usuários, nota-se que foi possível alcançar melhor RMSE em dois dos quatro recomendadores. Quando a calibração personalizada foi combinada com os algoritmos ItemKNN e SlopeOne, produziu resultados superiores para as métricas MAP e MRR, implicando maior precisão. No entanto, o mesmo não aconteceu ao utilizar os algoritmos NMF e SVD++. Comparando a calibração personalizada com a calibração baseada em gênero, obteve-se melhores LTC e F1 para os recomendadores NMF e Item KNN, respectivamente. Ademais, a precisão do Item KNN e SVD++ foi melhorada, conforme mostrado pelas métricas MRR e MAP. Por fim, nota-se a melhor equidade entre os grupos de usuários para todos os recomendadores, exceto para SlopeOne, cujo RMSE foi menor quando a calibração por gêneros foi aplicada.

Analisando a proposta de calibração baseada em popularidade juntamente com a calibração CP, há melhor F1 em todos os casos, indicando melhor calibração de gêneros e popularidade. A calibração baseada em popularidade proposta também foi capaz de melhorar a justiça em todos os recomendadores, como mostram os valores mais baixos no RMSE. Quando comparada à calibração baseada em gênero, a baseada em popularidade foi capaz de melhorar o Item KNN em termos de pontuação F1, e o Item KNN, NMF e SVD++ em termos de MRR, MAP e RMSE.

Tabela 9 – Comparação das abordagens de calibração propostas com os outros trabalhos usando o conjunto de dados do Yahoo Movies. O símbolo ▲ significa uma melhoria estatisticamente significativa da abordagem proposta em comparação com os outros trabalhos, com um valor $p < 0,05$ usando o teste t de Student; o símbolo • não denota nenhum ganho ou perda estatisticamente significativo; e o símbolo ▼ indica que o trabalho da literatura é estatisticamente melhor que a proposta. Cada par de símbolos está relacionado aos trabalhos CP e Calibração por gêneros, respectivamente.

Algoritmo	LTC	MRMC Gêneros	MRMC Pop.	F1 Score	MRR	MAP	ΔGAP_{BB}	ΔGAP_{N}	ΔGAP_{D}	RMSE
Item KNN + CP ($\lambda = 1.0$)	0.004	0.555	0.463	0.486	0.002	0.001	-0.990	-0.897	-0.974	0.551
NMF + CP ($\lambda = 0.2$)	0.187	0.517	0.296	0.573	0.018	0.006	-0.941	-0.769	-0.862	0.498
SlopeOne + CP ($\lambda = 0.2$)	0.096	0.538	0.379	0.529	0.003	0.001	-0.969	-0.844	-0.926	0.528
SVD++ + CP ($\lambda = 1.0$)	0.045	0.464	0.135	0.637	0.105	0.039	-0.746	-0.126	-0.487	0.353
Item KNN + Calibração por gêneros ($\lambda = 1.0$)	0.007	0.248	0.684	0.445	0.002	0.001	-0.994	-0.932	-0.987	0.561
NMF + Calibração por gêneros ($\lambda = 1.0$)	0.221	0.182	0.454	0.655	0.015	0.005	-0.974	-0.748	-0.934	0.514
SlopeOne + Calibração por gêneros ($\lambda = 1.0$)	0.149	0.213	0.511	0.603	0.006	0.002	-0.982	-0.765	-0.958	0.524
SVD++ + Calibração por gêneros ($\lambda = 0.2$) (STECK, 2018)	0.053	0.143	0.289	0.777	0.054	0.019	-0.887	0.205	-0.731	0.388
Item KNN + Calibração Personalizada ($\lambda = 1.0$)	0.007 ▲ •	0.423	0.539	0.512 ▲ ▲	0.002 ▲ ▲	0.001 ▲ ▲	-0.994	-0.895	-0.978	0.552 ▼ ▲
NMF + Calibração Personalizada ($\lambda = 1.0$)	0.224 ▲ ▲	0.366	0.328	0.653 ▲ ▼	0.014 ▼ ▼	0.005 ▼ •	-0.973	-0.708	-0.873	0.494 ▲ ▲
SlopeOne + Calibração Personalizada ($\lambda = 1.0$)	0.126 ▲ ▼	0.402	0.422	0.588 ▲ ▼	0.003 ▲ ▼	0.001 ▲ ▼	-0.982	-0.837	-0.936	0.532 ▼ ▼
SVD++ + Calibração Personalizada ($\lambda = 1.0$)	0.051 ▲ ▼	0.330	0.217	0.722 ▲ ▼	0.059 ▼ •	0.023 ▼ ▲	-0.882	-0.028	-0.589	0.349 ▲ ▲
Item KNN + Calibração por Popularidade ($\lambda = 1.0$)	0.004 • ▼	0.556	0.461	0.487 ▲ ▲	0.002 ▲ ▲	0.001 ▲ ▲	-0.990	-0.895	-0.974	0.551 ▲ ▲
NMF + Calibração por Popularidade ($\lambda = 0.2$)	0.200 ▲ ▼	0.512	0.257	0.589 ▲ ▼	0.019 • ▲	0.007 ▲ ▲	-0.934	-0.723	-0.846	0.485 ▲ ▲
SlopeOne + Calibração por Popularidade ($\lambda = 1.0$)	0.096 ▲ ▼	0.540	0.377	0.529 ▲ ▼	0.003 • ▼	0.001 ▲ ▼	-0.969	-0.840	-0.926	0.527 ▲ ▼
SVD++ + Calibração por Popularidade ($\lambda = 1.0$)	0.045 • ▼	0.462	0.122	0.667 ▲ ▼	0.112 ▲ ▲	0.043 ▲ ▲	-0.720	-0.037	-0.442	0.281 ▲ ▲

A Tabela 10 compara os métodos propostos de calibração personalizada e de popularidade com os dois métodos de última geração. Ao analisar a proposta de calibração personalizada em comparação à calibração CP, observa-se que a proposta obteve melhores valores para as métricas LTC, F1 e RMSE em todos os casos, indicando uma lista de recomendações mais diversificada e justa.

Quando comparada à calibração baseada em gênero (STECK, 2018), a calibração personalizada foi superior para todos os recomendadores em termos de precisão, conforme mostrado pelas métricas MRR e MAP, e também de justiça, conforme mostrado pela métrica RMSE. Em relação à proposta de calibração baseada em popularidade, ela demonstrou ser superior à CP em LTC, F1, MRR e MAP para a maioria dos recomendadores e mais justa que a calibração baseada em gênero em termos de RMSE para a maioria dos recomendadores.

Tabela 10 – Comparação das abordagens de calibração propostas com os outros trabalhos usando o conjunto de dados do MovieLens 20M. O símbolo ▲ significa uma melhoria estatisticamente significativa da abordagem proposta em comparação com os outros trabalhos, com um valor $p < 0,05$ usando o teste t de Student; o símbolo • não denota nenhum ganho ou perda estatisticamente significativo; e o símbolo ▼ indica que o trabalho da literatura é estatisticamente melhor que a proposta. Cada par de símbolos está relacionado aos trabalhos CP e Calibração por gêneros, respectivamente.

Algoritmo	**LTC**	**MRMC Gêneros**	**MRMC Pop.**	**F1 Score**	**MRR**	**MAP**	ΔGAP_{BB}	ΔGAP_{N}	ΔGAP_{D}	**RMSE**
Item KNN + CP ($\lambda = 0.0$)	0.428	0.553	0.341	0.531	0.244	0.105	-0.666	0.501	-0.471	0.319
NMF + CP ($\lambda = 0.2$)	0.114	0.578	0.383	0.499	0.072	0.028	-0.859	-0.359	-0.698	0.392
SlopeOne + CP ($\lambda = 0.2$)	0.026	0.556	0.639	0.396	0.044	0.017	-0.939	-0.673	-0.844	0.485
SVD++ + CP ($\lambda = 0.2$)	0.016	0.575	0.152	0.565	0.189	0.084	-0.623	0.153	-0.366	0.246
Item KNN + Calibração por gêneros ($\lambda = 1.0$)	0.505	0.253	0.333	0.705	0.134	0.053	-0.797	0.207	-0.632	0.347
NMF + Calibração por gêneros ($\lambda = 1.0$)	0.175	0.267	0.373	0.676	0.076	0.03	-0.809	0.095	-0.647	0.345
SlopeOne + Calibração por gêneros ($\lambda = 1.0$)	0.039	0.268	0.49	0.602	0.111	0.046	-0.893	0.001	-0.57	0.353
SVD++ + Calibração por gêneros ($\lambda = 1.0$)	0.03	0.289	0.227	0.748	0.188	0.083	-0.667	0.762	-0.311	0.353
Item KNN + Calibração Personalizada ($\lambda = 1.0$)	0.495 ▲▼	0.401	0.219	0.678 ▲▼	0.220 ▼▲	0.101 ▼▲	-0.797	0.274	-0.385	0.310 ▲▲
NMF + Calibração Personalizada ($\lambda = 1.0$)	0.175 ▲•	0.411	0.228	0.668 ▲▼	0.145 •▲	0.063 •▲	-0.785	0.282	-0.372	0.302 ▲▲
SlopeOne + Calibração Personalizada ($\lambda = 1.0$)	0.033 ▲▼	0.418	0.299	0.640 ▲▲	0.193 •▲	0.090 •▲	-0.893	0.374	-0.295	0.340 ▲▲
SVD++ + Calibração Personalizada ($\lambda = 1.0$)	0.030 ▲▲	0.431	0.162	0.678 ▲▼	0.204 ▲▲	0.090 ▲▲	-0.627	0.287	-0.198	0.238 ▲▲
Item KNN + Calibração por Popularidade ($\lambda = 0.0$)	0.428 ▲▼	0.553	0.341	0.531 ▲▼	0.244 ▲▲	0.105 ▲▲	-0.666	0.501	-0.471	0.319 •▲
NMF + Calibração por Popularidade ($\lambda = 0.0$)	0.114 ▲▼	0.566	0.579	0.424 ▼▼	0.047 ▼▼	0.017 ▼▼	-0.905	-0.167	-0.801	0.407 ▼▼
SlopeOne + Calibração por Popularidade ($\lambda = 1.0$)	0.023 ▼▼	0.559	0.068	0.603 ▲▲	0.031 ▼▼	0.011 ▼▼	-0.996	0.460	-0.896	0.134 •▲
SVD++ + Calibração por Popularidade ($\lambda = 1.0$)	0.016 ▲▼	0.575	0.066	0.584 ▲▼	0.254 ▲▲	0.123 ▲▲	-0.063	0.300	0.040	0.102 •▲

A Tabela 11 compara o método de calibração personalizado proposto com os dois métodos de última geração. Pode-se notar que a calibração personalizada, em comparação ao CP, rendeu resultados superiores para as métricas LTC e F1, indicando calibração mais diversificada e melhor em relação a gêneros e popularidade. No entanto, a CP foi capaz de fornecer melhor justiça do que a calibração personalizada, conforme mostrado pela métrica RMSE. Quando comparada à calibração baseada em gênero, a calibração personalizada foi superior em termos de MRR e MAP, mas não conseguiu superar esta abordagem nas métricas F1 e RMSE.

Em comparação com a proposta baseada em popularidade, obteve-se resultados superiores à calibração CP para a métrica LTC, indicando maior diversidade. Além disso, quando combinados com SlopeOne e SVD++, foram obtidos valores mais elevados para os gêneros MAP, MRR e MRMC, demonstrando boa precisão e maior justiça em termos de gêneros. Finalmente, no que diz respeito à calibração baseada em gênero, alcançou-se valores mais elevados para as métricas LTC, MRMC Pop, MAP e MRR, confirmando boa precisão, diversidade e justiça em termos de popularidade.

Tabela 11 – Comparação das abordagens de calibração propostas com os outros trabalhos usando o conjunto de dados do Yahoo Songs. O símbolo ▲ significa uma melhoria estatisticamente significativa da abordagem proposta em comparação com os outros trabalhos, com um valor $p < 0,05$ usando o teste t de Student; o símbolo • não denota nenhum ganho ou perda estatisticamente significativo; e o símbolo ▼ indica que o trabalho da literatura é estatisticamente melhor que a proposta. Cada par de símbolos está relacionado aos trabalhos CP e Calibração por gêneros, respectivamente.

Algoritmo	**LTC**	**MRMC Gêneros**	**MRMC Pop.**	**F1 Score**	**MRR**	**MAP**	ΔGAP_{BB}	ΔGAP_{N}	ΔGAP_{D}	**RMSE**
Item KNN + CP ($\lambda = 1.0$)	0.105	0.354	0.036	0.774	0.026	0.01	-0.357	-0.807	-0.675	0.370
NMF + CP ($\lambda = 1.0$)	0.105	0.355	0.044	0.771	0.014	0.005	-0.592	-0.811	-0.72	0.412
SlopeOne + CP ($\lambda = 1.0$)	0.08	0.362	0.04	0.767	0.008	0.003	-0.551	-0.814	-0.736	0.409
SVD++ + CP ($\lambda = 1.0$)	0.077	0.363	0.003	0.769	0.025	0.009	-0.167	-0.803	-0.635	0.346
Item KNN + Calibração por gêneros ($\lambda = 1.0$)	0.11	0.122	0.24	0.814	0.018	0.006	-0.823	-0.630	-0.735	0.424
NMF + Calibração por gêneros ($\lambda = 1.0$)	0.108	0.121	0.255	0.806	0.007	0.002	-0.851	-0.706	-0.787	0.453
SlopeOne + Calibração por gêneros ($\lambda = 1.0$)	0.086	0.119	0.255	0.807	0.003	0.001	-0.853	-0.696	-0.798	0.453
SVD++ + Calibração por gêneros ($\lambda = 1.0$)	0.082	0.129	0.182	0.845	0.017	0.006	-0.763	-0.604	-0.639	0.388
Item KNN + Calibração Personalizada ($\lambda = 1.0$)	0.111 ▲•	0.275	0.125	0.793 ▲▼	0.024 •▲	0.009 •▲	-0.823	-0.807	-0.672	0.445 ▼▼
NMF + Calibração Personalizada ($\lambda = 1.0$)	0.110 ▲•	0.278	0.134	0.788 ▲▼	0.011 •▲	0.004 •▲	-0.853	-0.807	-0.728	0.460 ▼▼
SlopeOne + Calibração Personalizada ($\lambda = 1.0$)	0.091 ▲▲	0.283	0.132	0.785 ▲▼	0.006 •▲	0.002 •▲	-0.853	-0.814	-0.739	0.464 ▼▼
SVD++ + Calibração Personalizada ($\lambda = 1.0$)	0.080 ▲▼	0.276	0.109	0.799 ▲▼	0.020 ▼▲	0.007 ▼▲	-0.738	-0.802	-0.608	0.415 ▼▼
Item KNN + Calibração por Popularidade ($\lambda = 1.0$)	0.111 ▲•	0.354	0.034	0.774 ▲▼	0.028 ▲▲	0.010 ▲▲	-0.357	-0.807	-0.573	0.351 ▲▲
NMF + Calibração por Popularidade ($\lambda = 0.4$)	0.108 ▲▲	0.349	0.063	0.768 ▼▼	0.012 ▼▲	0.004 ▼▲	-0.556	-0.793	-0.690	0.396 ▲▲
SlopeOne + Calibração por Popularidade ($\lambda = 1.0$)	0.090 ▲▲	0.362	0.038	0.767 ▲▼	0.010 ▲▲	0.004 ▲▲	-0.551	-0.814	-0.658	0.394 ▲▲
SVD++ + Calibração por Popularidade ($\lambda = 1.0$)	0.078 •▼	0.357	0.029	0.774 ▲▼	0.028 ▲▲	0.010 ▲▲	-0.320	-0.798	-0.519	0.334 ▲▲

7.4 Considerações Finais

Essa abordagem visava gerar recomendações calibradas conforme o nível do interesse do usuários pela popularidade dos itens. Assim, caso ele consuma acima do limite, ele é um usuário que se interessa por esse aspecto e recebe a calibração com base nisso; caso contrário, ele recebe somente uma calibração com base no gênero dos itens.

O experimento feito de forma *offline* trouxe resultados interessantes e que atendem às duas questões de pesquisa. Com relação à **RQ-1**, o sistema acaba por impactar a geração das recomendações, que funcionam em etapa de pós-processamento e reclassificam a lista gerada por qualquer modelo de calibração. Isso permite que o impacto da calibração seja com base nos gêneros ou com base na popularidade, o que visa atender da melhor forma possível o interesse do usuário.

Além disso, os resultados mostram que o desempenho da abordagem apresentou bons resultados para a métrica LTC, indicando uma lista mais diversificada e uma redução do viés de popularidade, o que responde à **RQ-2**.

CAPÍTULO

8

UMA ABORDAGEM DE CALIBRAÇÃO DUPLA

Este capítulo descreve de forma detalhada a abordagem que utiliza a calibração dupla em etapa de pós-processamento e é organizado da seguinte forma: a Seção 8.1 apresenta a justificativa de realização dessa abordagem; a Seção 8.2 descreve a metodologia e o design do experimento; por fim, os resultados são apresentados e discutidos na Seção 8.3.

8.1 Justificativa

A abordagem de calibração apresentada no Capítulo 7 utiliza uma estratégia de calibração do tipo chaveamento, onde dependendo da preferência do usuário, ele pode receber uma calibração por popularidade ou uma calibração por gênero. Sabendo disso, a estratégia de calibração deste capítulo visa unificar o modelo de calibração em um novo modelo integrado, que aproveita as categorias dos itens e calibrações de popularidade dentro de uma única abordagem, definida como empilhamento, em razão de empilhar dois tipos de calibração em sequência.

Nesta metodologia proposta, primeiro é implementada uma etapa de calibração de popularidade para garantir que os itens recomendados estejam alinhados com as preferências predominantes dos usuários nesse aspecto. Posteriormente, um componente de calibração de gênero que leva em conta especificamente as preferências diferenciadas associadas às categorias de itens é integrado ao sistema. Ao adotar esta abordagem em duas etapas, pretende-se melhorar significativamente a precisão das recomendações e entregar sugestões mais personalizadas e alinhadas com os gostos e interesses de cada indivíduo.

Em poucas palavras, é levantada a hipótese de que o viés de popularidade e a injustiça podem ser reduzidos na lista final de recomendações aos usuários com a adoção de um modelo único. A proposta foi avaliada com extensos experimentos *offline* usando dois conjuntos de dados. Os resultados obtidos são promissores quando comparados com uma variedade de trabalhos do

estado da arte. A proposta também foi comparada com a abordagem do Capítulo 7 e o trabalho foi submetido e aprovado como pôster para a conferência ACM SAC (SOUZA; MANZATO, 2024a).

8.2 Metodologia

Da mesma forma que foi apresentado nos Capítulos 6 e 7, a tarefa é explorar as preferências dos usuários para gerar uma lista de recomendações calibrada que aumente a justiça da popularidade e dos gêneros. Para fazer isso, será proposta uma abordagem de calibração em dois estágios. Na prática, o método estende a calibração de gênero proposta por (STECK, 2018) para realizar uma etapa adicional de calibração de acordo com diferentes níveis de popularidade de interesse do usuário. Como resultado, os usuários recebem uma lista de recomendações próxima ao perfil do usuário em termos de popularidade e gênero. A Figura 6 apresenta a estrutura de calibração em dois estágios, cujos detalhes são descritos a seguir.

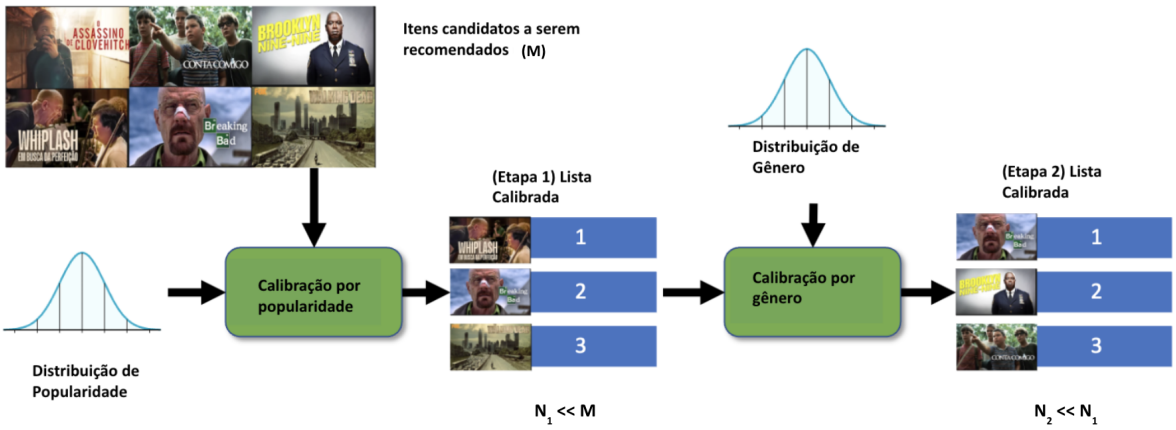


Figura 6 – Estrutura de calibração proposta. A saída da calibração de popularidade é a entrada para calibração de gênero, resultando em uma lista calibrada de recomendações de acordo com as preferências do usuário sobre popularidade e gêneros.

O primeiro passo é calcular a distribuição de popularidade, que é usada para calibrar a recomendação gerada pelo modelo de recomendação e que será a mesma definida na Equação 4.1. A divisão da popularidade dos itens também seguiu a mesma apresentada no Capítulo 6.

O processo de calibração consiste em reordenar a lista de candidatos à recomendação de acordo com a distribuição de popularidade. Para fazer isso, é necessário calcular a distribuição de recomendação $q_p(t_p|u)$, que é definida na Equação 4.2.

A segunda etapa da nossa proposta é a calibração dos gêneros, que requer o cálculo da distribuição dos gêneros. É utilizada a mesma formulação de (STECK, 2018), que calcula a distribuição de todos os gêneros $t_g \in C_g$ no perfil do usuário e na lista de recomendações.

Denota-se $p_g(t_g|u)$ como a distribuição do gênero alvo, definida como:

$$p_g(t_g|u) = \frac{\sum_{i \in I_u} w_r(u,i) \cdot p_g(t_g|i)}{\sum_{i \in I_u} w_r(u,i)} \tag{8.1}$$

onde $p_g(t_g|i)$ representa uma função indicadora definida como 1 se o gênero t_g estiver presente no item i e 0 caso contrário. Da mesma forma que a calibração de popularidade, a calibração de gênero requer a distribuição de recomendação, definida como $q_g(t_g|u)$:

$$q_g(t_g|u) = \frac{\sum_{i \in R^*_u} w_p(u,i) \cdot p_g(t_g|i)}{\sum_{i \in R^*_u} w_p(u,i)} \tag{8.2}$$

Calibração é definida como o processo para encontrar o conjunto ideal R^*_u, usando a relevância marginal máxima conforme mostrado na Equação 8.3:

$$R^*_{u,stage1} = \max_{CI_u} (1-\lambda_1) \cdot \sum_{i \in CI_u} w_r(u,i) - \lambda_1 \cdot D_{KL}(p_p, q_p(CI_u)) \tag{8.3}$$

onde λ_1 é um peso de *trade-off* que será explicado a seguir e D_{KL} é a função de justiça, definida na Equação 4.3.

Nessa abordagem, foi utilizada a medida de divergência Kullback-Leibler pelas mesmas razões apontadas por (STECK, 2018) e exploradas por (SILVA; MANZATO; DURÃO, 2021). O primeiro estágio de calibração é usado para encontrar o conjunto ideal de N_{stage1} itens de um conjunto de N itens candidatos CI_u recomendados pelo modelo, onde $N_{stage1} < N$.

Este conjunto ideal é baseado nos dados históricos de consumo de popularidade do usuário e no nível de popularidade do item. Ao final da primeira etapa, há uma lista de recomendações calibrada que respeita o nível e a proporção de popularidade que o usuário está interessado. A segunda etapa consiste em encontrar outro conjunto ideal de N_{stage2} itens do conjunto $R^*_{u,estgio1}$, onde $N_{estgio2} < N_{estgio1}$:

$$R^*_{end} = \sum_{i \in R^*_{u,stage1}} w_r(u,i) - \lambda_2 \cdot D_{KL}(p_g, q_g(R^*_{u,stage1})) \tag{8.4}$$

$$R^*_{u,stage2} = \max_{R^*_{u,stage1}} (1-\lambda_2) \cdot R^*_{end} \tag{8.5}$$

A lista de recomendações reordenada, $R^*_{u,stage2}$, de tamanho N_{stage2}, respeita as preferências do usuário tanto em popularidade quanto em gênero. O método propõe dois parâmetros λ_1 e λ_2, cujos valores podem ser $[0,1] \in \mathbb{R}$. O λ_1 controla a calibração de popularidade: se $\lambda_1 = 1$, então a recomendação no estágio 1 foca apenas nas preferências de popularidade dos usuários; se $\lambda_1 = 0$, focamos apenas nas pontuações de recomendação retornadas pelo modelo.

Tabela 12 – Estatísticas dos conjuntos de dados pré-processados.

Conjunto de dados	# Usuários	# Avaliações	# Itens
Yahoo Movies	7,642	221,367	10,825
MovieLens 20M	11,530	3,786,788	10,347

Da mesma forma, λ_2 segue a mesma lógica, mas no contexto da calibração de gênero. Assim como (SILVA; MANZATO; DURÃO, 2021), é adotada uma abordagem personalizada de pesos de compensação, ou seja, Variância Normalizada (**VAR**). A ideia é definir valores ótimos de λ_1 e λ_2 de acordo com as preferências de cada usuário para que a calibração possa ter mais ou menos efeito na lista de recomendações final. Esta técnica foi inicialmente usada na calibração de gênero (SILVA; MANZATO; DURÃO, 2021), mas foi estendida para a calibração de popularidade. A Variância Normalizada (VAR) é definida como:

$$\lambda_1(u) = 1 - \frac{\sum_{t_p \in C_p} |p_p(t_p|u) - m_p(u)|^2}{|C_p|} \tag{8.6}$$

$$\lambda_2(u) = 1 - \frac{\sum_{t_g \in C_g} |p_g(t_g|u) - m_g(u)|^2}{|C_g|} \tag{8.7}$$

onde $m_p(u)$ e $m_g(u)$ são respectivamente os valores médios de todas as distribuições de popularidade e gênero do usuário u, definidos como:

$$m_p(u) = \frac{\sum_{t_p \in C_p} p_p(t_p|u)}{|C_p|} \tag{8.8}$$

$$m_g(u) = \frac{\sum_{t_g \in C_g} p_g(t_g|u)}{|C_g|} \tag{8.9}$$

O experimento foi executado de forma *offline* em dois conjuntos de dados de filmes:

Yahoo Movies[1]: este conjunto foi configurado da mesma forma que nos Capítulos 4 e 7.

MovieLens-20M[2]: este conjunto de dados também foi configurado da mesma forma que nos Capítulos 4 e 7.

O processo de calibração não depende do algoritmo do sistema de recomendação. Atua como uma etapa de pós-processamento onde, após o modelo prever os itens candidatos para um usuário, aplica-se a calibração proposta em dois estágios para encontrar a melhor lista de itens para aquele usuário. Consequentemente, para entender o desempenho da calibração sob

[1] https://webscope.sandbox.yahoo.com/
[2] https://grouplens.org/datasets/movielens/20m/

diferentes algoritmos de recomendação, foram usados três modelos bem conhecidos descritos abaixo, baseados nos trabalhos de (STECK, 2018) e (SILVA; MANZATO; DURÃO, 2021). Para alguns modelos, foi utilizada a implementação fornecida por (HUG, 2020).

1. **SVD++**: O modelo usa a mesma configuração já apresentada no Capítulo 7.
2. **NMF**: Este modelo usa a mesma configuração já apresentada no Capítulo 7.
3. **VAE**: Autoencodificador Variacional para Filtragem Colaborativa, proposto por (LIANG *et al.*, 2018). Foi utilizada a implementação feita pela Microsoft[3].

O processo de calibração foi executado três vezes, da mesma forma que no Capítulo 7. Foram selecionados seis trabalhos do estado da arte especializados em viés de popularidade e calibração de gênero:

1. **Calibração por gênero**: o mesmo apresentado no Capítulo 7.
2. **PairWise**: este método foi apresentado no Capítulo 4.
3. **MF MACR**: proposto por (WEI *et al.*, 2021), este método atua como uma etapa de pós-processamento para redução de popularidade. Para o conjunto de dados do Yahoo Movies, foram escolhidos os parâmetros ajustados como $epoch = 100$, $batch = 1024$, $lr = 0,01$, $reg = 0,01$, $alpha = 0,001$, $beta = 0,001$ e $c = 20$. Para o conjunto de dados Movielens 20M, utilizou-se $batch = 2048$. Foi seguida a implementação dos autores[4].
4. **PDA**: proposto por (ZHANG *et al.*, 2021), este método implementa um novo paradigma de treinamento e inferência por meio de intervenção causal para redução do viés de popularidade. Para ambos os conjuntos de dados foram utilizados $epoch = 2000$, $batch = 2048$, $lr = 0,01$, $reg = 0,01$ e $pop_{exp} = 0,16$. Seguiu-se a implementação dos autores[5].
5. **CP**: este método é o mesmo já apresentado no Capítulo 7.
6. **Calibração Personalizada**: proposto por (SACILOTTI; SOUZA; MANZATO, 2023) e apresentado no Capítulo 7, este método implementa uma calibração baseada em chaveamento, onde alguns usuários recebem a calibração de gênero e outros recebem a calibração de popularidade. Seguiu-se a metodologia dos autores e explorou-se o parâmetro $\lambda \in [0,1]$.

A estrutura de calibração proposta em dois estágios usa pesos de compensação (λ_1, λ_2) para equilibrar as recomendações entre precisão e justiça. Conforme descrito anteriormente, o λ_1

[3] github.com/microsoft/recommenders/blob/main/examples/02_model_collaborative_filtering/multi_vae_deep_dive.ipynb
[4] https://github.com/weitianxin/MACR
[5] https://github.com/zyang1580/PDA

representa o *trade-off* da calibração de popularidade, referido como estágio 1, e λ_2 representa o estágio 2, calibração de gêneros. Para os modelos propostos existe uma combinação de ambos os pesos que minimiza a calibração incorreta.

Em relação aos trabalhos comparados, o λ controla a calibração de um estágio, que pode ser baseada na popularidade (CP (ABDOLLAHPOURI *et al.*, 2021), Personalizado (SACILOTTI; SOUZA; MANZATO, 2023)) ou gêneros (Steck (STECK, 2018), Personalizado (SACILOTTI; SOUZA; MANZATO, 2023)). Na análise, foram experimentados valores diferentes para λ, λ_1 e λ_2: $[0,0,0,1,...,1,0]$ e VAR. Este procedimento foi realizado para cada conjunto de dados, algoritmo e métrica. Os melhores valores foram selecionados e estão apresentados na Tabela 13.

Para isso, foram selecionadas as combinações de *trade-off* que rendeu o melhor valor para a métrica LTC, pois mede a cobertura dos itens recomendados por calibração. Da Tabela 13, é interessante notar que a abordagem proposta foi a única que gerou os melhores resultados de LTC usando a variância normalizada na segunda etapa, ou seja, calibração de gênero.

Tabela 13 – Comparação dos algoritmos implementados e dos valores de *trade-off* definidos para cada um.

	MovieLens 20M			**Yahoo Movies**		
Algoritmo	λ	λ_1	λ_2	λ	λ_1	λ_2
SVD++ + CP	0.9	-	-	0.9	-	-
SVD++ + Steck	0.8	-	-	1	-	-
SVD++ + Calibração Personalizada	0.9	-	-	0.8	-	-
SVD++ + Calibração Dupla	-	0.5	VAR	-	0.5	VAR
NMF + CP	0.1	-	-	0.1	-	-
NMF + Steck	0.1	-	-	1	-	-
NMF + Calibração Personalizada	0.9	-	-	0.8	-	-
NMF + Calibração Dupla	-	0.5	VAR	-	0.5	VAR
VAE + CP	1	-	-	0.2	-	-
VAE + Steck	0.6	-	-	1	-	-
VAE + Calibração Personalizada	0.1	-	-	1	-	-
VAE + Calibração Dupla	-	0.5	VAR	-	0.5	VAR

8.3 Resultados

A Tabela 14 apresenta os resultados obtidos para o conjunto de dados MovieLens 20M. Analisando apenas a **precisão** dos modelos pela métrica MAP, notamos que as abordagens PDA (ZHANG *et al.*, 2021) e *PairWise* (BORATTO; FENU; MARRAS, 2021) atingiram os maiores valores de MAP para todos os algoritmos. No entanto, esta conquista significa que os itens não são muito diversos entre si, como mostram os seus resultados relativos a LTC, F1 e RMSE.

Tabela 14 – Comparação do método proposto com os métodos de outros trabalhos da literatura no conjunto de dados MovieLens 20M. Os melhores valores obtidos para os algoritmos de recomendação SVD++, NMF e VAE estão em negrito. Os resultados comparando a estrutura de calibração proposta com outros métodos são estatisticamente significativos.

Algoritmo	**MAP**	**MRMC Gêneros**	**MRMC Pop.**	**F1 Score**	**LTC**	ΔGAP_{BB}	ΔGAP_{N}	ΔGAP_{D}	**RMSE**
MF MACR (WEI *et al.*, 2021)	0.168	0.58	0.29	0.528	0.26	-0.001	-0.514	-0.273	0.194
PairWise (BORATTO; FENU; MARRAS, 2021)	0.583	0.56	0.38	0.515	0.09	0.672	1.362	1.099	0.625
PDA (ZHANG *et al.*, 2021)	0.595	0.56	0.38	0.515	0.11	0.803	0.547	1.099	0.489
SVD++ + CP (ABDOLLAHPOURI *et al.*, 2021)	0.001	0.56	**0.68**	0.371	0.35	-0.991	-0.976	-0.987	0.569
SVD++ + Calibração por gêneros (STECK, 2018)	0.001	0.27	**0.68**	0.445	0.38	-0.989	-0.970	-0.984	**0.566**
SVD++ + Calibração personalizada (SACILOTTI; SOUZA; MANZATO, 2023)	0.001	0.41	0.69	0.407	**0.42**	-0.992	-0.970	-0.985	0.568
SVD++ + Calibração dupla	0.001	**0.26**	**0.68**	**0.447**	0.38	-0.990	-0.969	-0.983	**0.566**
NMF + CP (ABDOLLAHPOURI *et al.*, 2021)	**0.084**	0.51	**0.11**	0.633	0.01	-0.136	0.145	-0.118	**0.077**
NMF + Calibração por gêneros (STECK, 2018)	0.057	**0.23**	0.23	0.770	0.01	-0.652	0.238	-0.402	0.267
NMF + Calibração personalizada (SACILOTTI; SOUZA; MANZATO, 2023)	0.077	0.35	**0.11**	0.752	0.01	-0.171	0.238	-0.212	0.121
NMF + Calibração dupla	0.081	**0.23**	0.17	**0.799**	0.01	-0.379	0.046	-0.152	0.137
VAE + CP (ABDOLLAHPOURI *et al.*, 2021)	0.072	0.58	**0.13**	0.567	**0.20**	0.425	0.401	0.587	**0.276**
VAE + Calibração por gêneros (STECK, 2018)	0.058	**0.27**	0.31	0.710	0.11	0.458	1.684	1.101	0.688
VAE + Calibração personalizada (SACILOTTI; SOUZA; MANZATO, 2023)	**0.079**	0.47	0.35	0.584	0.14	0.850	1.553	1.386	0.750
VAE + Calibração dupla	0.055	**0.27**	0.28	**0.725**	0.11	0.359	1.636	0.981	0.647

Em relação à **justiça dos gêneros** através do MRMC de gêneros, a Tabela 14 indica que os três algoritmos de recomendação adotados pela proposta de calibração produziram os 3 melhores resultados, indicando que foi capaz de fornecer os mais equitativos itens em termos de gênero, para respeitar as proporções de preferência dos usuários por aquele aspecto. Outra abordagem que teve bons resultados foi a calibração de Steck (STECK, 2018), que foca na distribuição de itens por gênero.

Em relação à **justiça de popularidade**, os melhores resultados na popularidade do MRMC foram obtidos quando todas as abordagens de calibração foram combinadas com o modelo NMF. O modelo SVD++ apresentou os piores resultados, independente da calibração. Cabe ressaltar que os trabalhos com os 3 melhores resultados foram aqueles que aplicaram calibração com base no nível de interesse dos usuários em níveis de popularidade, como CP (ABDOLLAHPOURI *et al.*, 2021), Personalizado (SACILOTTI; SOUZA; MANZATO, 2023) e a proposta de calibração dupla.

Em termos de **cobertura de cauda longa**, a tabela indica que o modelo mais eficaz foi o SVD++ calibrado com abordagem de calibração personalizada. O NMF, no entanto, produziu os resultados menos favoráveis, talvez pela sua maior precisão: trabalhos com pontuações mais altas no MAP obtiveram valores mais baixos para o LTC. Além disso, é interessante notar a relação entre a popularidade do MRMC e o LTC. Embora o SVD++ tenha alcançado os melhores valores no LTC, ou seja, o algoritmo poderia recomendar mais itens desconhecidos aos usuários (a parte final da curva de cauda longa), isso impactou negativamente o erro de calibração da popularidade, provavelmente porque os usuários neste conjunto de dados têm um alto viés de popularidade.

Em relação à métrica **F1**, é possível observar que a proposta conseguiu alcançar os melhores resultados para os algoritmos SVD++, NMF e VAE, indicando que a abordagem de calibração dupla foi capaz de calibrar recomendações de acordo com gêneros e popularidade. Este aspecto é ainda validado ao analisar a métrica **RMSE**, onde o mesmo recomendador obteve

menor erro com a calibração dupla do que a de Steck, indicando que ela aborda os pontos de justiça mencionados e reduz o viés de popularidade do sistema. As combinações NMF + CP e VAE + CP obtiveram os melhores RMSE, mas com pontuação F1 inferior.

Os resultados relatados na Tabela 14 mostram que a abordagem de calibração em dois estágios foi capaz de equilibrar recomendações de acordo com gêneros e popularidade, em oposição aos outros trabalhos, que são mais adequados para um único aspecto, como precisão, gêneros ou popularidade. Embora não tenha conseguido atingir os valores mais baixos de RMSE, significando melhor equidade de acordo com diferentes grupos de usuários relacionados às preferências de popularidade, lembramos que esta métrica considera apenas a informação de popularidade, portanto combinações com os valores mais baixos de RMSE ainda são propensas a injustiças em relação aos gêneros dos itens.

Os resultados apresentados na Tabela 14 demonstram a importância de adotar métricas além da precisão na análise de algoritmos de recomendação. Reconhece-se a alta precisão do *PairWise* e do PDA, conforme indicado pela métrica MAP. No entanto, os usuários que preferem itens de nicho, diversos e impopulares são afetados por recomendações injustas e tendenciosas produzidas por essas abordagens.

Tabela 15 – Comparação da calibração proposta com os outros trabalhos da literatura no conjunto de dados do Yahoo Movies. Os melhores valores obtidos para os algoritmos de recomendação SVD++, NMF e VAE estão em negrito. Os resultados da comparação dos métodos propostos com outros métodos são estatisticamente significativos.

Algoritmo	MAP	MRMC Gêneros	MRMC Pop.	F1 Score	LTC	ΔGAP_{BB}	ΔGAP_{N}	ΔGAP_{D}	RMSE
MF MACR (WEI *et al.*, 2021)	0.010	0.35	0.40	0.624	0.13	-0.93	-0.57	-0.86	0.463
PairWise (BORATTO; FENU; MARRAS, 2021)	0.040	0.50	0.33	0.573	0.14	-0.66	1.33	-0.36	0.509
PDA (ZHANG *et al.*, 2021)	0.160	0.43	0.29	0.633	0.12	0.09	2.7	1.13	0.976
SVD++ + CP (ABDOLLAHPOURI *et al.*, 2021)	0.016	0.48	0.22	0.624	0.05	-0.768	-0.174	-0.546	**0.319**
SVD++ + Calibração por gêneros (STECK, 2018)	0.018	**0.13**	0.27	0.794	**0.06**	-0.872	0.187	-0.707	0.379
SVD++ + Calibração personalizada (SACILOTTI; SOUZA; MANZATO, 2023)	**0.028**	0.26	**0.19**	0.774	**0.06**	-0.749	-0.188	-0.611	0.329
SVD++ + Calibração dupla	0.022	**0.13**	0.25	**0.806**	**0.06**	-0.859	0.282	-0.666	0.374
NMF + CP (ABDOLLAHPOURI *et al.*, 2021)	0.004	0.51	0.31	0.574	0.21	-0.937	-0.708	-0.860	**0.485**
NMF + Calibração por gêneros (STECK, 2018)	0.004	0.16	0.44	0.672	0.26	-0.968	-0.712	-0.925	0.505
NMF + Calibração personalizada (SACILOTTI; SOUZA; MANZATO, 2023)	0.006	0.31	0.38	0.654	**0.27**	-0.927	-0.718	-0.899	0.493
NMF + Calibração dupla	**0.009**	**0.15**	**0.32**	**0.756**	**0.27**	-0.909	-0.384	-0.795	0.422
VAE + CP (ABDOLLAHPOURI *et al.*, 2021)	0.001	0.48	0.46	0.530	0.10	-0.988	-0.863	-0.972	**0.544**
VAE + Calibração por gêneros (STECK, 2018)	0.001	0.15	0.41	0.697	0.12	-0.980	-0.802	-0.955	0.529
VAE + Calibração personalizada (SACILOTTI; SOUZA; MANZATO, 2023)	0.001	0.30	0.35	0.675	0.11	-0.956	-0.803	-0.941	0.522
VAE + Calibração dupla	**0.010**	**0.13**	**0.25**	**0.806**	**0.13**	-0.845	0.548	-0.654	0.400

A Tabela 15 apresenta os resultados obtidos para o conjunto de dados do Yahoo Movies. Analisando a **precisão**, tanto o PDA (ZHANG *et al.*, 2021) quanto o PairWise (BORATTO; FENU; MARRAS, 2021) superaram as outras abordagens. No entanto, os resultados também indicam que estas abordagens devolvem recomendações injustas em termos de gênero e popularidade, e carecem de diversidade.

As abordagens de calibração combinadas com o recomendador VAE produziram os piores resultados, enquanto as abordagens combinadas com SVD++ alcançaram os melhores resultados. Em relação à **justiça dos gêneros**, a calibração em duas etapas foi capaz de alcançar os melhores resultados e também melhorar todos os modelos de recomendação. Também é interessante notar que superou a calibração por gêneros de Steck (STECK, 2018), uma abordagem de calibração

específica de gênero, para os algoritmos NMF e VAE. Esta conquista se deve à calibração da popularidade que afeta indiretamente a imparcialidade dos gêneros, já que alguns gêneros são mais populares que outros.

Em relação à **justiça de popularidade**, trabalhos que consideraram o aspecto popularidade do item, como a calibração personalizada (SACILOTTI; SOUZA; MANZATO, 2023), CP (ABDOLLAHPOURI *et al.*, 2021) e a proposta, tiveram melhor desempenho, principalmente quando combinados com SVD++. O PDA (ZHANG *et al.*, 2021) também obteve resultados promissores, mas ao custo de grande injustiça para usuários que preferem itens de nicho (ΔGAP_N).

Nos algoritmos NMF e VAE, a abordagem de dois estágios melhorou suas recomendações comparado aos outros algoritmos de calibração. Em relação à **cobertura de cauda longa**, o NMF calibrado pela abordagem Personalizada (SACILOTTI; SOUZA; MANZATO, 2023) e pela calibração dupla obteve os melhores resultados entre todas as abordagens, enquanto o SVD++ produziu os piores resultados independentemente da calibração. Além disso, PDA (ZHANG *et al.*, 2021) e *PairWise* (BORATTO; FENU; MARRAS, 2021) alcançaram valores baixos apesar de terem alta precisão. Em SVD++, NMF e VAE, a calibração dupla produziu resultados competitivos de LTC.

Em relação à **F1**, pode-se observar que a proposta obteve os melhores valores, destacando seu alto desempenho em termos de justiça nos gêneros e popularidade. Ademais, o SVD++ obteve os melhores resultados em **RMSE** independentemente da calibração, indicando que o sistema reduziu com sucesso o viés de popularidade para diferentes grupos de usuários. Embora não tenha conseguido superar o CP nesta métrica, ela se destaca por manter resultados promissores tanto para justiça nos gêneros quanto para popularidade (F1), o que não é verdade para a abordagem CP.

A Tabela 15 relata resultados semelhantes aos do conjunto de dados MovieLens, indicando que a proposta melhorou a justiça dos gêneros e a popularidade em ambos os conjuntos de dados. Embora a abordagem de calibração proposta não tenha alcançado alta precisão, obteve o menor erro de calibração de gênero e de popularidade competitiva, o que significa que o modelo fornece recomendações que respeitam o perfil do usuário tanto no gênero quanto no consumo de popularidade. De fato, o maior valor da métrica F1 foi obtido por SVD++ e VAE calibrado com a calibração dupla. Isso indica que o pipeline de calibração proposto pode reduzir a injustiça e o viés de popularidade em um modelo unificado que não depende de um algoritmo de recomendação específico.

8.4 Considerações Finais

Nesta abordagem, foi proposta uma técnica de calibração baseada na popularidade e nos gêneros dos itens para trazer recomendações mais justas e que atendam às preferências dos usuários. É possível utilizar dois pesos de *trade-off* para ajustar o sistema e obter a melhor

combinação possível para as métricas analisadas.

Em relação a questão de pesquisa **RQ-1**, a abordagem impacta a geração das recomendações reclassificando a lista gerada por qualquer modelo acoplado a ela. Assim, o intuito é trazer recomendações coerentes com as preferências dos usuários.

Ademais, os experimentos demonstraram que a abordagem apresentou uma redução no viés de popularidade dos itens recomendados, uma lista de sugestões mais justa para três grupos diferentes de usuários e recomendações mais justas em termos de gênero e popularidade, o que atende a questão de pesquisa **RQ-2**. Apesar disso, os resultados mostraram que existem trabalhos com melhores valores de acurácia, mas que sofrem de viés de popularidade.

CAPÍTULO

9

CONCLUSÃO

9.1 Contribuições

A principal contribuição desta pesquisa é a realização de um estudo de abordagens de calibração e redução do viés de popularidade em Sistemas de Recomendação. Este estudo permitiu verificar como as abordagens podem fornecer recomendações relevantes para diferentes grupos de usuários, medidos por meio de um conjunto de métricas relacionado à injustiça.

Este trabalho consistiu na realização de experimentos *online* e *offline*, tendo ao todo 5 trabalhos submetidos e aceitos. Os trabalhos contribuíram com informações relevantes como a validação de que os itens calibrados foram percebidos como relevantes para os usuários de sistemas calibrados no aspecto de popularidade. Os *nudges* também se mostraram interfaces interessantes para promover itens de nicho e aumentar a diversidade, bem como reduzir o impacto do viés de popularidade.

As abordagens em etapas de pós-processamento se mostraram interessantes em razão de ser possível combiná-las com qualquer modelo de recomendação e classificarem a lista de forma balanceada conforme o interesse do usuário, seja usando a estratégia de *switch* ou a de *stack*. Por fim, as abordagens de calibração em etapa de processamento também apresentaram bons resultados e podem ser utilizadas para lidar com o viés de popularidade, de forma a aumentar a diversidade do sistema.

9.2 Comparação entre as Abordagens

A Tabela 16 faz um resumo dos resultados obtidos na base de dados do Yahoo Movies com as abordagens propostas e estudadas neste trabalho. É possível notar que cada tipo de abordagem tem uma vantagem em algum aspecto, sendo então sua utilização compatível de acordo com a necessidade do usuário.

Para o caso de ser necessário uma maior cobertura dos itens de cauda longa, ou seja, uma maior diversidade e surpresa durante o uso do sistema, a abordagem do BPR modificado é mais interessante em razão dos valores superiores na métrica LTC. Caso a necessidade seja uma precisão maior no sistema, a abordagem de calibração personalizada juntamente com o SVD++ é a melhor escolha. Caso o objetivo seja um sistema justo em termos de popularidade e gêneros, a abordagem de calibração dupla combinada com o VAE ou com o SVD++ é a melhor escolha. Por fim, caso a redução do viés de popularidade para diferentes grupos seja o grande objetivo do sistema, a combinação da proposta de calibração personalizada com o SVD++ é a melhor opção.

Tabela 16 – Comparação de todas as abordagens propostas neste trabalho no conjunto de dados do Yahoo Movies. Os melhores valores obtidos para cada métrica estão em negrito. Os resultados da comparação são estatisticamente significativos.

Algoritmo	MAP	MRMC Gêneros	MRMC Pop.	F1 Score	LTC	ΔGAP_{BB}	ΔGAP_{N}	ΔGAP_{D}	RMSE
BPR Modificado	0.004	0.59	0.50	0.444	**0.32**	-0.934	-0.142	-0.835	0.420
SVD++ + Calibração personalizada (SACILOTTI; SOUZA; MANZATO, 2023)	**0.028**	0.26	**0.19**	0.774	0.06	-0.749	-0.188	-0.611	**0.329**
SVD++ + Calibração dupla	0.022	**0.13**	0.25	**0.806**	0.06	-0.859	0.282	-0.666	0.374
NMF + Calibração personalizada (SACILOTTI; SOUZA; MANZATO, 2023)	0.006	0.31	0.38	0.654	0.27	-0.927	-0.718	-0.899	0.493
NMF + Calibração dupla	0.009	0.15	0.32	0.756	0.27	-0.909	-0.384	-0.795	0.422
VAE + Calibração personalizada (SACILOTTI; SOUZA; MANZATO, 2023)	0.001	0.30	0.35	0.675	0.11	-0.956	-0.803	-0.941	0.522
VAE + Calibração dupla	0.010	**0.13**	0.25	**0.806**	0.13	-0.845	0.548	-0.654	0.400

A Tabela 17 faz um resumo dos resultados obtidos na base de dados do Movie Lens 20M com as abordagens propostas e estudadas neste trabalho. Assim como na base do Yahoo Movies, é possível notar que cada tipo de abordagem tem uma vantagem em alguma característica, sendo então sua utilização compatível de acordo com a necessidade do sistema.

A abordagem do BPR modificado também obteve valores superiores para métrica LTC, sendo indicada para casos de necessidade de diversidade do sistema. Caso a necessidade seja uma precisão maior no sistema, a abordagem de calibração dupla combinada com o NMF é a melhor escolha. Caso o objetivo seja um sistema justo em termos de popularidade e gêneros, a abordagem de calibração dupla combinada com o NMF é a melhor opção. Por fim, caso a redução do viés de popularidade para diferentes grupos seja o grande objetivo do sistema, a combinação da proposta de calibração personalizada com o NMF é a que irá trazer melhores resultados.

Tabela 17 – Comparação de todas as abordagens propostas neste trabalho no conjunto de dados do Movie Lens 20M. Os melhores valores obtidos para cada métrica estão em negrito. Os resultados da comparação são estatisticamente significativos.

Algoritmo	MAP	MRMC Gêneros	MRMC Pop.	F1 Score	LTC	ΔGAP_{BB}	ΔGAP_{N}	ΔGAP_{D}	RMSE
BPR Modificado	0.001	0.45	0.33	0.596	**0.46**	-0.865	-0.060	-0.693	0.370
SVD++ + Calibração personalizada (SACILOTTI; SOUZA; MANZATO, 2023)	0.001	0.41	0.69	0.407	0.42	-0.992	-0.970	-0.985	0.568
SVD++ + Calibração dupla	0.001	0.26	0.68	0.447	0.38	-0.990	-0.969	-0.983	0.566
NMF + Calibração personalizada (SACILOTTI; SOUZA; MANZATO, 2023)	0.077	0.35	**0.11**	0.752	0.01	-0.171	0.238	-0.212	**0.121**
NMF + Calibração dupla	**0.081**	**0.23**	0.17	**0.799**	0.01	-0.379	0.046	-0.152	0.137
VAE + Calibração personalizada (SACILOTTI; SOUZA; MANZATO, 2023)	0.079	0.47	0.35	0.584	0.14	0.850	1.553	1.386	0.750
VAE + Calibração dupla	0.055	0.27	0.28	0.725	0.11	0.359	1.636	0.981	0.647

9.3 Publicações

Durante o desenvolvimento desta pesquisa, 5 trabalhos foram submetidos e publicados. Um deles foi publicado na revista *UMUAI - User Modeling and User-Adapted Interaction 2022*, outro na conferência ICEIS, um deles no ICEIS Book, outro na conferência ACM SAC e outro na conferência ACM UMAP. Os trabalhos estão listados abaixo:

- **Digitally nudging users to explore off-profile recommendations: here be dragons**

  Abstract: In many application domains of recommender systems, e.g., on media streaming sites, one main goal of the provider of the recommendation service is to increase the engagement of users by helping them *discover* new types of content they like. Standard collaborative filtering algorithms by design often lead to a certain level of discovery. Nonetheless, in certain domains, it may be helpful to more actively promote content to users beyond their past preference profile ("off-profile") and thereby help users explore new content. However, when showing such off-profile content to users in combination with more familiar content, the new content items may be overlooked. In this research, we explore to what extent *digital nudging*, i.e., subtly directing user choices in a specific direction, can help to raise the attention and interest of users for off-profile content. We conducted a user study (N=1,064) on a real-world social book recommendation app. We find that users who are nudged towards recommended books of their non-preferred genres significantly more often put these off-profile books on their reading lists, thus confirming the effectiveness of digital nudging in this application. However, we also found that As a result, we find that digital nudging in recommendations, while effective in the short run, must be done with due care, keeping an eye on the overall quality perceptions by users and potentially harmful long-term effects.

  Referência: ALVES, G.; JANNACH, D.; SOUZA, R. F. de; DAMIAN, D.; MANZATO, M. G. Digitally nudging users to explore off-profile recommendations: here be dragons. **User Modeling and User-Adapted Interaction**, Springer, p. 1–41, 2023.

- **Counteracting popularity-bias and improving diversity through calibrated recommendations.**

  Abstract: Recent works have shown a connection between fairness, miscalibration, and popularity bias in recommender systems. While popularity bias can shift users towards consumption of more mainstream items, this phenomenon also affects the calibration and fairness of recommendations, where certain users' tastes are not fairly represented by the system, while other users receive recommendations consistent with their preferences. However, most state-of-art works on calibration focus only on providing fairer recommendations to users, not considering the popularity bias which can amplify the long tail effect.

To fill the research gap, in this work, we propose a calibration approach that aims to meet users' interests according to different levels of the items' popularity. The proposed system works in a post-processing step and was evaluated through metrics that analyze aspects of fairness, popularity, and accuracy through an offline experiment with two different datasets. The system's efficiency was validated and evaluated with three different recommendation algorithms, verifying which behaves better and comparing the performance with four other state-of-the-art calibration approaches. As a result, the proposed technique reduced popularity bias and increased diversity and fairness in the two datasets considered.

Referência: SACILOTTI, A.; SOUZA, R. F. d.; MANZATO, M. G. Counteracting popularity-bias and improving diversity through calibrated recommendations. In: Proceedings. Prague, Czech Republic: SciTePress, 2023.

- **Enhancing Calibration and Reducing Popularity Bias in Recommender Systems**

Abstract: The recent literature highlights that recommendation systems are significantly influenced by popularity bias. This phenomenon has far-reaching implications for the fairness and accuracy of recommendations. This bias often results in some users finding their preferences inadequately reflected in their recommendations, while others benefit from more consistent suggestions. Nevertheless, despite the current state-of-art efforts in this field that primarily aim to provide fairer recommendations, a crucial aspect has been overlooked: the impact of popularity bias on the long tail effect, which leads to a decline in the visibility of less popular items in recommendations. To address this research gap, the present study introduces a calibration approach designed to cater to the diverse interests of users across various levels of item popularity. To achieve this objective, we propose a post-processing system that is independent of any specific recommendation algorithm. Building upon the foundational idea presented by (SACILOTTI; SOUZA; MANZATO, 2023), we evaluate the efficacy of our proposed system using an additional dataset from the domain of music. The performance assessment of our system encompasses a range of metrics that consider aspects related to popularity, accuracy, and fairness. Additionally, four recommendation algorithms and two distinct baselines are employed. As a result, the proposed technique mitigates popularity bias, augmenting diversity and fairness within the considered datasets.

Referência: SOUZA, R. F. d.; MANZATO, M. G. Enhancing calibration and reducing popularity bias in recommender systems. In: ICEIS. Prague, Czech Republic: SciTePress, 2024 (to appear).

- **A Two-Stage Calibration Approach for Mitigating Bias and Fairness in Recommender Systems**

Abstract: Popularity bias and unfairness are problems caused by the lack of calibration in recommender systems. Works that intend to reduce the effect of popularity bias do not consider the distribution of item genres/categories in the users' profiles. Other studies aim to calibrate the system to generate fair recommendations according to users' profiles, but usually are still biased towards popularity. We propose a system calibration approach based on users' preferences for different levels of popularity of items and their genres. The proposed approach works in the post-processing stage and can be combined with different recommendation models. We evaluated the system with offline experiments using one state-of-the-art dataset, three recommender algorithms, six baselines, and different metrics for popularity, fairness, and accuracy. The results indicate reduced popularity bias and improved fairness.

Referência: _____. A two-stage calibration approach for mitigating bias and fairness in recommender systems. In: The 39th ACM/SIGAPP Symposium on Applied Computing (SAC '24). New York, NY, USA: ACM, 2024 (to appear).

- **User Perception of Fairness-Calibrated Recommendations**

Abstract: The research community has become increasingly aware of possible undesired effects of algorithmic biases in recommender systems. One common bias in such systems is to over-proportionally expose certain items to users, which may ultimately result in a system that is considered unfair to individual stakeholders. From a technical perspective, calibration approaches are commonly adopted in such situations to ensure that the individual user's preferences are better taken into account, thereby also leading to a more balanced exposure of items overall. Given the known limitations of today's predominant offline evaluation approaches, our work aims to contribute to a better understanding of the users' *perception* of the fairness and quality of recommendations when these are served in a calibrated way. Therefore, we conducted an online user study (N=500) in which we exposed the treatment groups with recommendations calibrated for fairness in terms of two different item characteristics. Our results show that calibration can indeed be effective in guiding the users' choices towards the "fairness items" without negatively impacting the overall quality perception of the system. We however also found that calibration did not measurably impact the users' fairness perceptions unless explanatory information is provided by the system. Finally, our study points to challenges when applying calibration approaches in practice in terms of finding appropriate parameters.

Referência: ALVES, G.; JANNACH, D.; SOUZA, R. F. de; MANZATO, M. G. User perception of fairness-calibrated recommendations. In: Proceedings of the 32st ACM Conference on User Modeling, Adaptation and Personalization. New York, NY, USA: ACM, 2024 (to appear).

9.4 Limitações

Apesar das vantagens demonstradas das propostas de calibração em pós-processamento serem compatíveis com vários modelos de recomendação, existe uma limitação a este respeito: que o sistema em que é implementada deve ter algum meio de resolver o problema do *cold start*, dependendo do modelo utilizado.

Esta limitação surge porque, ao integrar com recomendações obtidas do modelo inicial, o sistema fica sujeito à forma como o modelo inicial lida com o problema de partida fria, pois modelos como UserKNN e ItemKNN não podem gerar recomendações significativas sem informações do usuário. Este ponto representa um dos aspectos que podem ser mais explorados em pesquisas futuras relacionadas à nossa proposta.

9.5 Trabalhos Futuros

Em nossa proposta ainda há espaço para novos trabalhos, como os listados abaixo:

1. Estudar a possibilidade de incorporar etapas adicionais de calibração com base em diferentes informações secundárias, como os metadados dos itens. Assim, pode haver um aumento no desempenho do sistema.
2. Realização dos experimentos em diferentes domínios para validar a generalização das abordagens.
3. Estudar outras estratégias para combinar os módulos do sistema, incluindo recomendação e calibração, para melhorar ainda mais o desempenho geral do nosso sistema.
4. Realização de experimentos mais profundos para explorar a relação entre *nudges* e viés de popularidade.

REFERÊNCIAS

ABDOLLAHPOURI, H.; BURKE, R.; MOBASHER, B. Popularity-aware item weighting for long-tail recommendation. **arXiv preprint arXiv:1802.05382**, 2018. Citado na página 28.

____. Managing popularity bias in recommender systems with personalized re-ranking. In: **The thirty-second international flairs conference**. California, USA: AAAI Press, 2019. Citado nas páginas 19, 37 e 41.

ABDOLLAHPOURI, H.; MANSOURY, M.; BURKE, R.; MOBASHER, B. The impact of popularity bias on fairness and calibration in recommendation. **arXiv preprint arXiv:1910.05755**, 2019. Citado na página 35.

____. The unfairness of popularity bias in recommendation. **arXiv preprint arXiv:1907.13286**, 2019. Citado nas páginas 28, 36 e 63.

____. The connection between popularity bias, calibration, and fairness in recommendation. In: **Fourteenth ACM conference on recommender systems**. New York, NY, USA: Association for Computing Machinery, 2020. p. 726–731. Citado nas páginas 18 e 30.

ABDOLLAHPOURI, H.; MANSOURY, M.; BURKE, R.; MOBASHER, B.; MALTHOUSE, E. C. User-centered evaluation of popularity bias in recommender systems. In: MASTHOFF, J.; HERDER, E.; TINTAREV, N.; TKALCIC, M. (Ed.). **Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization, UMAP 2021, Utrecht, The Netherlands, June, 21-25, 2021**. ACM, 2021. p. 119–129. ISBN 978-1-4503-8366-0. Disponível em: <https://doi.org/10.1145/3450613.3456821>. Citado nas páginas 40, 42, 45, 51, 73, 82, 83, 84 e 85.

AGGARWAL, C. C. **Recommender Systems: The Textbook**. 1st. ed. New York, NY, USA: Springer Publishing Company, Incorporated, 2016. ISBN 3319296574. Citado nas páginas 24, 25, 26 e 27.

AHANGER, A. B.; AALAM, S. W.; BHAT, M. R.; ASSAD, A. Popularity bias in recommender systems-a review. In: SPRINGER. **International Conference on Emerging Technologies in Computer Engineering**. New York, NY, USA, 2022. p. 431–444. Citado na página 38.

ALVES, G.; JANNACH, D.; SOUZA, R. F. de; DAMIAN, D.; MANZATO, M. G. Digitally nudging users to explore off-profile recommendations: here be dragons. **User Modeling and User-Adapted Interaction**, Springer, p. 1–41, 2023. Citado na página 56.

ALVES, G.; JANNACH, D.; SOUZA, R. F. de; MANZATO, M. G. User perception of fairness-calibrated recommendations. In: **Proceedings of the 32st ACM Conference on User Modeling, Adaptation and Personalization**. New York, NY, USA: ACM, 2024. Citado na página 62.

ANELLI, V. W.; BELLOGÍN, A.; NOIA, T. D.; JANNACH, D.; POMO, C. Top-n recommendation algorithms: A quest for the state-of-the-art. In: **Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization**. New York, NY, USA: ACM, 2022. p. 121–131. Citado na página 63.

BALTRUNAS, L.; AMATRIAIN, X. Towards time-dependant recommendation based on implicit feedback. **Proceedings of the Third ACM Conference on Recommender Systems**, 01 2009. Citado na página 17.

BENNETT, J.; LANNING, S. *et al.* The netflix prize. In: **Proceedings of KDD cup and workshop**. New York, NY, USA: ACM, 2007. p. 35. Citado na página 23.

BEUTEL, A.; CHI, E.; GOODROW, C.; CHEN, J.; DOSHI, T.; QIAN, H.; WEI, L.; WU, Y.; HELDT, L.; ZHAO, Z.; HONG, L. Fairness in recommendation ranking through pairwise comparisons. In: **Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining**. New York, NY, USA: Association for Computing Machinery, 2019. p. 2212–2220. ISBN 978-1-4503-6201-6. Citado nas páginas 39 e 41.

BORATTO, L.; FENU, G.; MARRAS, M. Connecting user and item perspectives in popularity debiasing for collaborative recommendation. **Information Processing & Management**, Elsevier, v. 58, n. 1, p. 102387, 2021. Citado nas páginas 39, 41, 44, 51, 52, 82, 83, 84 e 85.

BORGES, R.; STEFANIDIS, K. On mitigating popularity bias in recommendations via variational autoencoders. In: **Proceedings of the 36th Annual ACM Symposium on Applied Computing**. New York, NY, USA: Association for Computing Machinery, 2021. (SAC '21), p. 1383–1389. ISBN 9781450381048. Disponível em: <https://doi.org/10.1145/3412841.3442123>. Citado nas páginas 37 e 41.

BOZDAG, E.; GAO, Q.; HOUBEN, G.-J.; WARNIER, M. Does offline political segregation affect the filter bubble? an empirical analysis of information diversity for dutch and turkish twitter users. **Computers in human behavior**, Elsevier, v. 41, p. 405–415, 2014. Citado na página 31.

BURKE, R. Hybrid web recommender systems. **The adaptive web**, Springer, p. 377–408, 2007. Citado na página 24.

CARABAN, A.; KARAPANOS, E.; GONÇALVES, D.; CAMPOS, P. 23 ways to nudge: A review of technology-mediated nudging in human-computer interaction. In: **Proceedings of the 2019 CHI conference on human factors in computing systems**. New York, NY, USA: ACM, 2019. p. 1–15. Citado na página 55.

CHA, S.-H. Comprehensive survey on distance/similarity measures between probability density functions. **City**, v. 1, n. 2, p. 1, 2007. Citado nas páginas 47 e 72.

CHEN, J.; DONG, H.; WANG, X.; FENG, F.; WANG, M.; HE, X. Bias and debias in recommender system: A survey and future directions. **ACM Transactions on Information Systems**, ACM New York, NY, New York, NY, USA, v. 41, n. 3, p. 1–39, 2023. Citado nas páginas 18 e 30.

CHEN, X.; FAN, W.; CHEN, J.; LIU, H.; LIU, Z.; ZHANG, Z.; LI, Q. Fairly adaptive negative sampling for recommendations. In: **Proceedings of the ACM Web Conference 2023**. New York, NY, USA: ACM, 2023. p. 3723–3733. Citado nas páginas 39 e 44.

CHEN, Z.; WU, J.; LI, C.; CHEN, J.; XIAO, R.; ZHAO, B. Co-training disentangled domain adaptation network for leveraging popularity bias in recommenders. In: **Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval**. New York, NY, USA: ACM, 2022. p. 60–69. Citado na página 19.

DELDJOO, Y.; JANNACH, D.; BELLOGIN, A.; DIFONZO, A.; ZANZONELLI, D. Fairness in recommender systems: research landscape and future directions. **User Modeling and User-Adapted Interaction**, Springer, p. 1–50, 2023. Citado nas páginas 17, 18 e 35.

EKSTRAND, M. D.; TIAN, M.; AZPIAZU, I. M.; EKSTRAND, J. D.; ANUYAH, O.; MCNEILL, D.; PERA, M. S. All the cool kids, how do they fit in?: Popularity and demographic biases in recommender evaluation and effectiveness. **Proceedings of Machine Learning Research**, PMLR, v. 81, p. 172–186, 23–24 Feb 2018. Citado na página 30.

ELSWEILER, D.; TRATTNER, C.; HARVEY, M. Exploiting food choice biases for healthier recipe recommendation. In: **Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval**. New York, NY, USA: Association for Computing Machinery, 2017. (SIGIR '17), p. 575–584. ISBN 9781450350228. Disponível em: <https://doi.org/10.1145/3077136.3080826>. Citado na página 30.

GELFERT, A. Fake news: A definition. **Informal logic**, Informal Logic, v. 38, n. 1, p. 84–117, 2018. Citado na página 31.

GEYIK, S. C.; AMBLER, S.; KENTHAPADI, K. Fairness-aware ranking in search recommendation systems with application to linkedin talent search. In: **Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery Data Mining**. New York, NY, USA: Association for Computing Machinery, 2019. (KDD '19), p. 2221–2231. ISBN 9781450362016. Disponível em: <https://doi.org/10.1145/3292500.3330691>. Citado nas páginas 19, 40 e 41.

GHARAHIGHEHI, A.; VENS, C.; PLIAKOS, K. Fair multi-stakeholder news recommender system with hypergraph ranking. **Information Processing & Management**, Elsevier, v. 58, n. 5, p. 102663, 2021. Citado nas páginas 37 e 41.

HELBERGER, N.; KARPPINEN, K.; D'ACUNTO, L. Exposure diversity as a design principle for recommender systems. **Information, Communication & Society**, Taylor & Francis, v. 21, n. 2, p. 191–207, 2018. Citado na página 31.

HUG, N. Surprise: A python library for recommender systems. **Journal of Open Source Software**, v. 5, n. 52, p. 2174, 2020. Citado nas páginas 73 e 81.

INGESSON, E. **Algorithmic vs. Perceived Fairness in Music Recommender Systems: An Investigation of Popularity Bias from a User Perspective**. Dissertação (Mestrado), 2022. Citado nas páginas 19 e 36.

JANNACH, D.; LERCHE, L.; KAMEHKHOSH, I.; JUGOVAC, M. What recommenders recommend: an analysis of recommendation biases and possible countermeasures. **User Modeling and User-Adapted Interaction**, Springer, v. 25, n. 5, p. 427–491, 2015. Citado na página 57.

JESSE, M.; JANNACH, D. Digital nudging with recommender systems: Survey and future directions. **Computers in Human Behavior Reports**, Elsevier, v. 3, p. 100052, 2021. Citado na página 32.

JUGOVAC, M.; JANNACH, D.; LERCHE, L. Efficient optimization of multiple recommendation quality factors according to individual user tendencies. **Expert Systems with Applications**, Elsevier, v. 81, p. 321–331, 2017. Citado na página 61.

KAYA, M.; BRIDGE, D. A comparison of calibrated and intent-aware recommendations. In: **Proceedings of the 13th ACM Conference on Recommender Systems**. New York, NY, USA: Association for Computing Machinery, 2019. (RecSys '19), p. 151–159. ISBN 9781450362436. Disponível em: <https://doi.org/10.1145/3298689.3347045>. Citado nas páginas 32, 40 e 41.

KHENISSI, S.; NASRAOUI, O. Modeling and counteracting exposure bias in recommender systems. **arXiv preprint arXiv:2001.04832**, 2020. Citado na página 18.

KLIMASHEVSKAIA, A.; ELAHI, M.; JANNACH, D.; SKJÆRVEN, L.; TESSEM, A.; TRATTNER, C. Evaluating the effects of calibrated popularity bias mitigation: A field study. In: **Proceedings of the 17th ACM Conference on Recommender Systems**. New York, NY, USA: ACM, 2023. p. 1084–1089. Citado na página 62.

KONSTAN, J. A.; RIEDL, J. Recommender systems: from algorithms to user experience. **User modeling and user-adapted interaction**, Springer, v. 22, n. 1, p. 101–123, 2012. Citado na página 31.

KOREN, Y. Factorization meets the neighborhood: a multifaceted collaborative filtering model. In: **Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining**. New York, NY, USA: ACM, 2008. p. 426–434. Citado na página 73.

KOWALD, D.; SCHEDL, M.; LEX, E. The unfairness of popularity bias in music recommendation: A reproducibility study. In: **Advances in Information Retrieval: 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14–17, 2020, Proceedings, Part II**. Berlin, Heidelberg: Springer-Verlag, 2020. p. 35–42. ISBN 978-3-030-45441-8. Disponível em: <https://doi.org/10.1007/978-3-030-45442-5_5>. Citado nas páginas 17 e 36.

LEONARD, T. C. **Richard H. Thaler, Cass R. Sunstein, Nudge: Improving decisions about health, wealth, and happiness: Yale University Press, New Haven, CT, 2008, 293 pp,** 26.00.*NewYork*, *NY*, *USA* : *Springer*, 2008. *Citadonapgina*55.

LESOTA, O.; MELCHIORRE, A.; REKABSAZ, N.; BRANDL, S.; KOWALD, D.; LEX, E.; SCHEDL, M. Analyzing item popularity bias of music recommender systems: are different genders equally affected? In: **Proceedings of the 15th ACM Conference on Recommender Systems**. New York, NY, USA: ACM, 2021. p. 601–606. Citado na página 19.

LIANG, D.; KRISHNAN, R. G.; HOFFMAN, M. D.; JEBARA, T. **Variational Autoencoders for Collaborative Filtering**. arXiv, 2018. Disponível em: <https://arxiv.org/abs/1802.05814>. Citado na página 81.

LIN, A.; WANG, J.; ZHU, Z.; CAVERLEE, J. Quantifying and mitigating popularity bias in conversational recommender systems. In: **Proceedings of the 31st ACM International Conference on Information & Knowledge Management**. New York, NY, USA: ACM, 2022. p. 1238–1247. Citado na página 19.

LIU, H.; ZHAO, N.; ZHANG, X.; LIN, H.; YANG, L.; XU, B.; LIN, Y.; FAN, W. Dual constraints and adversarial learning for fair recommenders. **Knowledge-Based Systems**, Elsevier, v. 239, p. 108058, 2022. Citado nas páginas 39 e 44.

LUNARDI, G. M.; MACHADO, G. M.; MARAN, V.; OLIVEIRA, J. P. M. de. A metric for filter bubble measurement in recommender algorithms considering the news domain. **Applied Soft Computing**, Elsevier, v. 97, p. 106771, 2020. Citado na página 31.

LUO, X.; ZHOU, M.; XIA, Y.; ZHU, Q. An efficient non-negative matrix-factorization-based approach to collaborative filtering for recommender systems. **IEEE Transactions on Industrial Informatics**, IEEE, v. 10, n. 2, p. 1273–1284, 2014. Citado na página 73.

MANSOURY, M.; ABDOLLAHPOURI, H.; PECHENIZKIY, M.; MOBASHER, B.; BURKE, R. Feedback loop and bias amplification in recommender systems. In: **Proceedings of the 29th ACM International Conference on Information Knowledge Management**. New York, NY, USA: Association for Computing Machinery, 2020. (CIKM '20), p. 2145–2148. ISBN 9781450368599. Disponível em: <https://doi.org/10.1145/3340531.3412152>. Citado na página 36.

MUNSON, S. A.; RESNICK, P. Presenting diverse political opinions: How and how much. In: **Proceedings of the SIGCHI Conference on Human Factors in Computing Systems**. New York, NY, USA: Association for Computing Machinery, 2010. (CHI '10), p. 1457–1466. ISBN 9781605589299. Disponível em: <https://doi.org/10.1145/1753326.1753543>. Citado na página 31.

NAGHIAEI, M.; RAHMANI, H. A.; DEHGHAN, M. The unfairness of popularity bias in book recommendation. In: SPRINGER. **International Workshop on Algorithmic Bias in Search and Recommendation**. New York, NY, USA, 2022. p. 69–81. Citado na página 19.

NAGULENDRA, S.; VASSILEVA, J. Understanding and controlling the filter bubble through interactive visualization: A user study. In: **Proceedings of the 25th ACM Conference on Hypertext and Social Media**. New York, NY, USA: Association for Computing Machinery, 2014. (HT '14), p. 107–115. ISBN 9781450329545. Disponível em: <https://doi.org/10.1145/2631775.2631811>. Citado na página 31.

NGUYEN, T. T.; HUI, P.-M.; HARPER, F. M.; TERVEEN, L.; KONSTAN, J. A. Exploring the filter bubble: The effect of using recommender systems on content diversity. In: **Proceedings of the 23rd International Conference on World Wide Web**. New York, NY, USA: Association for Computing Machinery, 2014. (WWW '14), p. 677–686. ISBN 9781450327442. Disponível em: <https://doi.org/10.1145/2566486.2568012>. Citado na página 30.

NING, X.; KARYPIS, G. Slim: Sparse linear methods for top-n recommender systems. In: IEEE. **2011 IEEE 11th international conference on data mining**. New York, NY, USA, 2011. p. 497–506. Citado na página 63.

PARISER, E. **The Filter Bubble: What the Internet Is Hiding from You**. London, UK: Penguin Group , The, 2011. ISBN 1594203008. Citado na página 31.

PARRA, D.; SAHEBI, S. Recommender systems: Sources of knowledge and evaluation metrics. In: **Advanced techniques in web intelligence-2**. New York, NY, USA: Springer, 2013. p. 149–175. Citado na página 27.

PASSE, J.; DRAKE, C.; MAYGER, L. Homophily, echo chambers, & selective exposure in social networks: What should civic educators do? **The Journal of Social Studies Research**, Elsevier, v. 42, n. 3, p. 261–271, 2018. Citado na página 31.

PITOURA, E.; STEFANIDIS, K.; KOUTRIKA, G. Fairness in rankings and recommendations: an overview. **The VLDB Journal**, Springer, p. 1–28, 2022. Citado nas páginas 38, 39 e 43.

PU, P.; CHEN, L.; HU, R. A user-centric evaluation framework for recommender systems. In: **Proceedings of the Fifth ACM Conference on Recommender Systems**. New York, NY, USA: Association for Computing Machinery, 2011. (RecSys '11), p. 157–164. ISBN 9781450306836. Disponível em: <https://doi.org/10.1145/2043932.2043962>. Citado na página 56.

QIN, Y. **A historical survey of music recommendation systems: Towards evaluation**. Dissertação (Mestrado) — McGill University Libraries, 2013. Citado na página 27.

RENDLE, S.; FREUDENTHALER, C.; GANTNER, Z.; SCHMIDT-THIEME, L. Bpr: Bayesian personalized ranking from implicit feedback. **arXiv preprint arXiv:1205.2618**, 2012. Citado nas páginas 21, 38, 43, 44, 47, 48 e 51.

RICCI, F.; ROKACH, L.; SHAPIRA, B. Introduction to recommender systems handbook. In: RICCI, F.; ROKACH, L.; SHAPIRA, B.; KANTOR, P. B. (Ed.). **Recommender Systems Handbook**. Boston, MA: Springer US, 2011. p. 1–35. ISBN 978-0-387-85820-3. Disponível em: <https://doi.org/10.1007/978-0-387-85820-3_1>. Citado nas páginas 23 e 24.

____. Recommender systems: Introduction and challenges. In: ____. **Recommender Systems Handbook**. Boston, MA: Springer US, 2015. p. 1–34. ISBN 978-1-4899-7637-6. Disponível em: <https://doi.org/10.1007/978-1-4899-7637-6_1>. Citado nas páginas 23, 26 e 31.

SACILOTTI, A.; SOUZA, R. F. d.; MANZATO, M. G. Counteracting popularity-bias and improving diversity through calibrated recommendations. In: **In Proceedings of the 25th International Conference on Enterprise Information Systems**. Prague, Czech Republic: Scitepress, 2023. v. 1. Citado nas páginas 70, 71, 81, 82, 83, 84, 85, 88 e 90.

SEYMEN, S.; ABDOLLAHPOURI, H.; MALTHOUSE, E. C. A constrained optimization approach for calibrated recommendations. In: **Fifteenth ACM Conference on Recommender Systems**. New York, NY, USA: Association for Computing Machinery, 2021. p. 607–612. ISBN 9781450384582. Disponível em: <https://doi.org/10.1145/3460231.3478857>. Citado nas páginas 40 e 42.

SILVA, D. C. da; MANZATO, M. G.; DURÃO, F. A. Exploiting personalized calibration and metrics for fairness recommendation. **Expert Systems with Applications**, Elsevier, v. 181, p. 115112, 2021. Citado nas páginas 19, 27, 39, 41, 44, 46, 47, 51, 69, 71, 72, 73, 79, 80 e 81.

SOUZA, R.; MANZATO, M. A two-stage calibration approach for mitigating bias and fairness in recommender systems. In: **Proceedings of the 39th ACM/SIGAPP Symposium on Applied Computing**. New York, NY, USA: ACM, 2024. p. 1659–1661. Citado na página 78.

SOUZA, R. F. d.; MANZATO, M. G. Enhancing calibration and reducing popularity bias in recommender systems. In: **Proceedings of the 26th International Conference on Enterprise Information Systems**. Prague, Czech Republic: SciTePress, 2024. Citado na página 70.

STECK, H. Calibrated recommendations. In: **Proceedings of the 12th ACM Conference on Recommender Systems**. New York, NY, USA: Association for Computing Machinery, 2018. (RecSys '18), p. 154–162. ISBN 9781450359016. Disponível em: <https://doi.org/10.1145/3240323.3240372>. Citado nas páginas 18, 19, 32, 39, 40, 41, 44, 45, 46, 50, 61, 63, 69, 70, 71, 73, 74, 75, 78, 79, 81, 82, 83 e 84.

VERMA, S.; GAO, R.; SHAH, C. Facets of fairness in search and recommendation. In: SPRINGER. **Bias and Social Aspects in Search and Recommendation: First International Workshop, BIAS 2020, Lisbon, Portugal, April 14, Proceedings 1**. New York, NY, USA, 2020. p. 1–11. Citado nas páginas 46 e 71.

WANG, C.; WANG, K.; BIAN, A.; ISLAM, R.; KEYA, K. N.; FOULDS, J.; PAN, S. Do humans prefer debiased ai algorithms? a case study in career recommendation. In: **27th International Conference on Intelligent User Interfaces**. New York, NY, USA: Association for Computing Machinery, 2022. (IUI '22), p. 134–147. ISBN 9781450391443. Disponível em: <https://doi.org/10.1145/3490099.3511108>. Citado na página 36.

WANG, W.; FENG, F.; HE, X.; WANG, X.; CHUA, T.-S. Deconfounded recommendation for alleviating bias amplification. In: **Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining**. New York, NY, USA: Association for Computing Machinery, 2021. p. 1717–1725. Citado nas páginas 37 e 41.

WEI, T.; FENG, F.; CHEN, J.; WU, Z.; YI, J.; HE, X. Model-agnostic counterfactual reasoning for eliminating popularity bias in recommender system. In: **Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery Data Mining**. New York, NY, USA: Association for Computing Machinery, 2021. (KDD '21), p. 1791–1800. ISBN 9781450383325. Disponível em: <https://doi.org/10.1145/3447548.3467289>. Citado nas páginas 38, 41, 81, 83 e 84.

WEINMANN, M.; SCHNEIDER, C.; BROCKE, J. v. Digital nudging. **Business & Information Systems Engineering**, Springer, v. 58, p. 433–436, 2016. Citado na página 55.

YALCIN, E. Blockbuster: A new perspective on popularity-bias in recommender systems. In: IEEE. **2021 6th International Conference on Computer Science and Engineering (UBMK)**. New York, NY, USA, 2021. p. 107–112. Citado na página 19.

YALCIN, E.; BILGE, A. Investigating and counteracting popularity bias in group recommendations. **Information Processing & Management**, Elsevier, v. 58, n. 5, p. 102608, 2021. Citado nas páginas 37 e 41.

____. Treating adverse effects of blockbuster bias on beyond-accuracy quality of personalized recommendations. **Engineering Science and Technology, an International Journal**, Elsevier, v. 33, p. 101083, 2022. Citado na página 19.

YOO, K.-H.; GRETZEL, U.; ZANKER, M. **Persuasive recommender systems: conceptual background and implications**. New York, NY, USA: Springer Science & Business Media, 2012. Citado na página 55.

ZHANG, Y.; FENG, F.; HE, X.; WEI, T.; SONG, C.; LING, G.; ZHANG, Y. Causal intervention for leveraging popularity bias in recommendation. In: **Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval**. New York, NY, USA: Association for Computing Machinery, 2021. (SIGIR '21), p. 11–20. ISBN 9781450380379. Disponível em: <https://doi.org/10.1145/3404835.3462875>. Citado nas páginas 37, 41, 81, 82, 83, 84 e 85.

ZHU, Z.; HE, Y.; ZHAO, X.; ZHANG, Y.; WANG, J.; CAVERLEE, J. Popularity-opportunity bias in collaborative filtering. In: **Proceedings of the 14th ACM International Conference on Web Search and Data Mining**. New York, NY, USA: Association for Computing Machinery, 2021. p. 85–93. Citado nas páginas 37 e 41.

ZHU, Z.; WANG, J.; CAVERLEE, J. Measuring and mitigating item under-recommendation bias in personalized ranking systems. In: **Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval**. New York, NY, USA:

Association for Computing Machinery, 2020. p. 449–458. ISBN 9781450380164. Disponível em: <https://doi.org/10.1145/3397271.3401177>. Citado nas páginas 39 e 41.

APÊNDICE

A

IMAGENS DAS TELAS DO EXPERIMENTO COM USUÁRIOS DA ABORDAGEM *NUDGE*

A Figura 7 apresenta a tela de escolha de gêneros preferidos do experimento com usuários feito no Capítulo 5. A Figura 8 apresenta a tela de livros recomendados para o usuário com o *nudge*. A Figura 9 apresenta a versão de livros recomendados para o usuário sem *nudge*. Por fim, a Figura 10 apresenta os detalhes do livro para o usuário.

Figura 7 – Tela de seleção de gêneros preferidos do usuário.

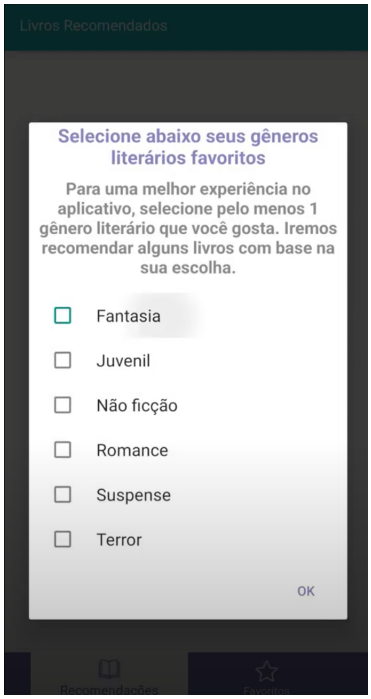


Figura 8 – Tela de livros recomendados para o usuário com *nudge*.

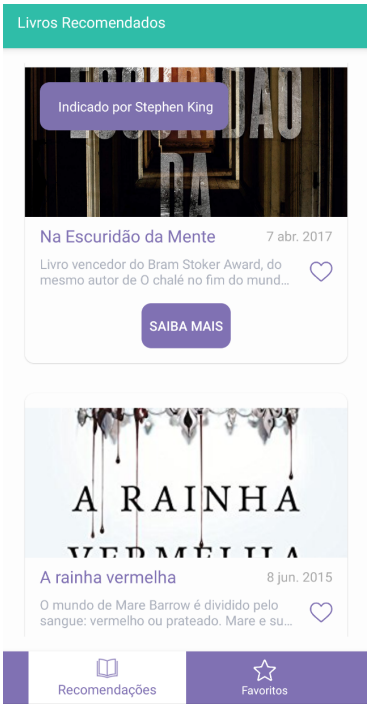


Figura 9 – Tela de livros recomendados para o usuário sem *nudge*.

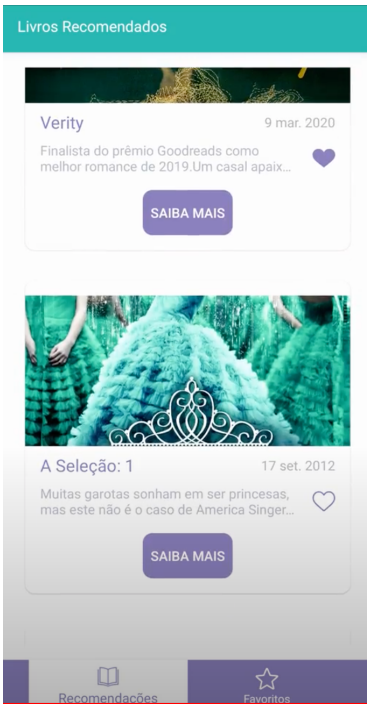


Figura 10 – Tela de detalhes do livro.

APÊNDICE

B

QUESTÕES DE PESQUISA DO EXPERIMENTO DE CALIBRAÇÃO EM PÓS-PROCESSAMENTO

A Tabela 18 apresenta as questões de pesquisa relacionadas ao experimento com usuários feito no Capítulo 6.

Tabela 18 – Itens do Questionário Pós-Tarefa: Percepções de Qualidade e Justiça

	Questões de Pesquisa
Q1	Este sistema de recomendação me deu boas sugestões.
Q2	O sistema entende meus gostos e preferências de filmes.
Q3	Escolher apenas um filme para assistir depois foi difícil para mim.
Q4	Eu entendi por que os itens foram recomendados para mim.
Q5	Acredito que alguns filmes tiveram uma chance maior de serem recomendados do que outros. *(Justiça)*
Q6	Já conheço muitos dos filmes recomendados.
Q7	Estou confiante de que fiz uma boa escolha.
Q8	Havia várias boas opções nas recomendações.
Q9	Por favor, selecione "concordo parcialmente"para mostrar que você está prestando atenção a esta pergunta. (Verificação de atenção)
Q10	Os filmes recomendados são bem conhecidos e populares. *(Percepção de popularidade)*
Q11	Os filmes recomendados eram produções de alto orçamento. *(Percepção de orçamento)*
Q12	No geral, estou satisfeito com este sistema de recomendação.
Q13	As recomendações feitas pelo sistema foram geralmente justas. *(Percepção de justiça)*
Q14	Eu usaria um sistema como este se precisasse de ajuda para encontrar um novo filme para assistir no futuro.(Intenção de uso)

APÊNDICE

C

IMAGENS DAS TELAS DO EXPERIMENTO DE CALIBRAÇÃO EM PÓS-PROCESSAMENTO

A Figura 11 apresenta a tela de avaliação de filmes do experimento com usuários feito no Capítulo 6. O usuário tinha que escolher no mínimo 7 filmes de seu interesse e avaliá-los para que o sistema pudesse construir o seu perfil. Já a Figura 12, apresenta a tela de filmes recomendados para o usuário. O sistema recomendava dez filmes e o usuário deveria escolher qual deles achou mais interessante. Dessa forma era possível avaliar se o filme selecionado veio ou não da calibração.

Figura 11 – Tela de seleção de filmes para avaliação e construção do perfil do usuário.

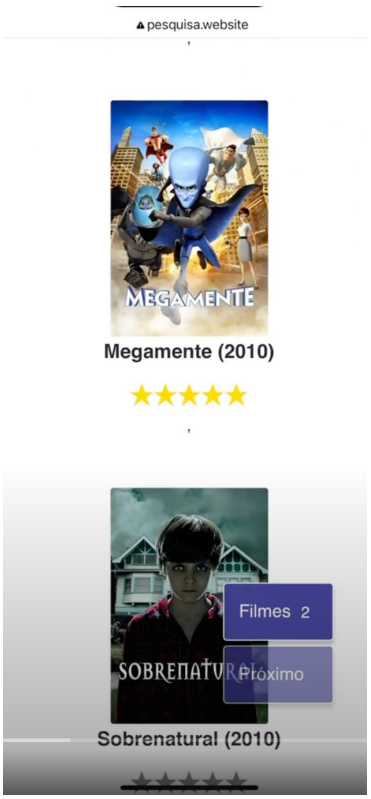


Figura 12 – Tela de filmes recomendados para o usuário.


ICMC
USP
SÃO CARLOS