PROJETO TCC - BCC	ANO/SEMESTRE:	2020.2

SISTEMAS DE RECOMENDAÇÃO DE PRODUTOS EM E-COMMERCE

Giulio Giovanella

Prof. Aurélio Faustino Hoppe – Orientador

1 INTRODUÇÃO

De acordo com Singh e Singh (2012), grande parte dos dados mundiais, cerca de 90%, foram gerados nos últimos anos. Parte das ações dos usuários, que são geradas ao navegar na internet, são capturadas, gerando dados cada vez mais volumosos ~~dados~~. Redes sociais, dispositivos móveis, e tecnologias como Internet das Coisas, geraram uma explosão de dados, criando o cenário chamado de Big Data (FAN; BIFET, 2012).

Com a enorme quantidade de opções nos modernos website, o usuário possuí mais possibilidades do que pode lidar (SCHWARTZ, 2011; RICCI *et al*., 2010). Diante deste contexto, os Sistemas de Recomendação (SsR) ganham a função primordial de poupar o usuário de horas gastas em escolha, bem como aumentar seu engajamento à plataforma que está acessando, com recomendações que o atraem. Sua principal dependência está nos dados coletados de cada usuário, para compreender seu perfil e suas preferências, para então fazer recomendações personalizadas.

O sucesso de um comércio eletrônico está diretamente ligado à sua capacidade de transformar novos usuários em usuários recorrentes (RAMOS, 2015 *apud* SILVA, 2018). Porém, em conformidade com uma pesquisa da Experian Hitwise, somente 1,65 usuários entre 100 que entram em um e-commerce acabam adquirindo algum produto, sendo considerada uma taxa de conversão baixa (REIS, 2016 *apud* SILVA, 2018). Portanto, é crucial para os empreendimentos em comércio eletrônico que os novos usuários se sintam atraídos pelos produtos apresentados, caso contrário os novos usuários tendem a parar de acessar a página. Diante deste contexto, Silva (2018) ressalta que um dos grandes desafios é como atrair os novos usuários com recomendações pertinentes a ele com pouca~~,~~ ou nenhuma~~,~~ informação do cliente, tal problema é conhecido como *cold-start*.

Comentado [AS1]: Quando usar apud deve ser inserido a página. Rever isso em todo o texto.

Formatado: Realce

Silva (2018) também afirma que uma das soluções tradicionais para tentar atrair novos usuários, é a utilização de Sistemas de Recomendação não-personalizados, ou seja, recomendar os mesmos produtos para todos os usuários, independentemente de seu perfil. Esses Sistemas de Recomendação não-personalizados costumam utilizar dados como, popularidade, avaliações obtidas, e o período de consumo. O problema dessa estratégia é que ela parte da premissa que itens populares, com boas avaliações e que estão em período de consumo, tem o potencial de serem do agrado da maioria dos usuários. Porém, mesmo considerando que esta premissa seja válida para a maioria dos casos, a diversidade com que a internet está imersa nos dias atuais contempla homens e mulheres, jovens, adultos, idosos, sem distinção de opção religiosa, classe social, faixa etária, ou qualquer outra (HAMMOND *et al*., 2017 *apud* SILVA, 2018). Silva (2018) conclui que nem sempre os produtos que agradam uma grande parcela da população serão capazes de satisfazer todas as preferências dos distintos perfis de usuário que acessam o e-commerce.

Segundo Singh e Singh (2012), as modernas soluções de Big Data e Aprendizado de Máquina (AM) são muito eficazes em aprender padrões, mas para isso precisam de dados em grande volume para alcançarem altas taxas de acerto. Prando (2016) propôs a utilização de dados de redes sociais, extrair as preferências do usuário e fazer recomendações personalizadas com base em seu perfil, mitigando o *cold-start*. Nicolas (2018) propôs as técnicas de *Max-Coverage* e *Niche-Coverage*, para recomendações não personalizadas para novos usuários. Almeida (2016), optou pela técnica GPClerk, que utiliza Programação Genética para encontrar produtos similares utilizando os atributos dos produtos a fim de efetuar comparações.

Diante desse cenário, este trabalho propõe o desenvolvimento de um sistema de recomendação que minimiza o *cold-start*, utilizando da Ciência dos Dados e AM para recomendações para usuários com dados de navegação e compra disponíveis. Além disso, também avaliará a influência da aplicação de descontos no comportamento de compra dos consumidores.

1.1 OBJETIVOS

O objetivo deste trabalho é utilizar um sistema de recomendação para e-commerce que maximize a conversão dos produtos recomendados em vendas concretizadas, minimizando o problema de *cold-start*.

Os objetivos específicos são:

a) utilizar dados externos ao e-commerce para minimizar o *cold-start*;
b) recomendar produtos correlatos usando regras de associação;
c) avaliar a influência de descontos em produtos na compra dos consumidores.

2 TRABALHOS CORRELATOS

Neste capítulo são apresentados trabalhos que possuem semelhanças a proposta deste trabalho. A seção 2.1 apresenta um novo método para obter produtos similares a outros, proposto por Almeida (2016), chamado de GPClerk. A seção 2.2 apresenta as duas abordagens sugeridas por Silva (2018), para transformar novos usuários em usuários recorrentes. A seção 2.3 traz uma abordagem para extrair as preferências dos usuários baseando-se em redes sociais, proposta por Prando (2016).

2.1 LEARNING TO RECOMMEND SIMILAR ALTERNATIVE PRODUCTS IN E-COMMERCE CATALOGS

Almeida (2016) propôs o método GPClerk para lidar com o problema de *cold-start*. O GPClerk utiliza Programação Genética para medir a similaridade entre produtos, combinando sua estratégia para gerar exemplos de treino a fim de torná-lo viável para um cenário real de *e-commerce*. Além disso, utiliza-se uma função de comparação que analisa cada um dos atributos do par de produtos selecionado, gerando uma pontuação de similaridade de cada atributo.

O autor formula que, considerando $V_{i,j}$ como sendo os produtos visitados por um usuário U_i em uma sessão S_j sobre o e-commerce, temos o conjunto de todos os produtos visitados por algum usuário (*UV*), que pode ser obtido através dos *logs* capturados pelo e-commerce, em um determinado período de tempo (ALMEIDA, 2016). Sua afirmação é que existe uma intersecção não vazia entre o conjunto de alternativas similares a um produto, e o conjunto *UV*, visto que uma das principais motivações de um usuário navegar por produtos diferentes em uma mesma sessão, é que eles são alternativas similares.

Comentado [AS2]: Rever redação.

Almeida (2016) verificou, com o auxílio de usuários, que o GPClerk junto de sua técnica para gerar conjuntos de dados, em quase 70% dos casos os produtos recomendados eram alternativas similares ou correlatas ao produto em destaque. O autor também constatou que o GPClerk foi capaz de encontrar pares de alternativas similares, mesmo que essas não tenham sido visualizadas juntas pelos usuários. Por fim, o autor conclui que o método é um complemento adequado aos métodos tradicionais baseados em técnicas de filtragem colaborativas, especialmente para lidar com o problema de *cold-start*.

Para trabalhos futuros, Almeida (2016) sugere que sejam feitos outros tipos de relações entre produtos além da similaridade, e planeja experimentar essa abordagem para encontrar produtos complementares. Como o GPClerk necessita da especificação dos atributos dos produtos estruturadas, o autor acredita ser possível adicionar ao processo de mineração de informação métodos para extrair dados de descrições textuais não estruturadas do catálogo dos produtos.

2.2 SISTEMAS DE RECOMENDAÇÃO NÃO-PERSONALIZADOS PARA ATRAIR USUÁRIOS NOVOS

A capacidade dos SsR recomendar produtos úteis à usuários sem nenhuma informação vinculada, problema conhecido como *Pure Cold-Start*, foi a motivação de Silva (2018) para propor duas novas abordagens para atenuar o problema: *Max-Coverage* e *Niche-Coverage*. Com a premissa de que diversificar os itens recomendados para novos usuários também seria útil, elaborou-se três hipóteses para avaliar sua premissa: (1) significativa parte dos usuários não está interessada apenas em itens populares; (2) SR não-personalizados que exploram outras métricas além dos itens populares são capazes de mitigar o *cold-start*; (3) páginas de produtos compostas por SR não-personalizados complementares satisfazem o interesse de um grande número de novos usuários distintos (SILVA, 2018).

A estratégia da abordagem de Max-Coverage de Silva (2018) consiste na aplicação da tradicional estratégia de *Maximum k-Coverage* de Hochbaum e Pathria (1998) no contexto de encontrar produtos para serem recomendados para novos usuários. Dado U_i como sendo os usuários cadastrados no sistema, e I_i sendo os itens disponíveis no catálogo de produtos, modela-se um conjunto $F = \{S_1, ..., S_n\}$ em que cada subconjunto S_i é relativo a um item existente (SILVA, 2018). Dessa forma, o subconjunto *Si* é composto pelos usuários que compraram o item *i*, tornando S_i um subconjunto de usuários (SILVA, 2018). Então, o autor define como objetivo encontrar *k* subconjuntos S_i que possuam o maior número de usuários distintos, constituindo o conjunto $F^* = \{S_{i1}, ..., S_{ik}\}$, ou seja, os produtos que o foram consumidos pelo maior número de usuários diferentes.

Silva (2018) destaca que esse problema pertence à classe de problemas NP-Difícil, o qual não se conhece uma solução em tempo polinomial. No entanto, o autor argumenta que o custo computacional do problema cairia com uma heurística gulosa simples, já que cada iteração maximiza o número de usuários não cobertos, e possui uma razão de aproximadamente 63% da solução ótima, como demonstrado por Chvatal (1979).

Comentado [AS3]: Usar linguagem formal

A abordagem de *Niche-Coverage*, proposta por Silva (2018), consiste em explorar os itens característicos de distintos nichos de usuários, sendo os itens característicos definidos como os itens com maior probabilidade de satisfazer a preferência individual da maioria dos usuários do nicho. Silva (2018) define que um usuário *u* gosta

de um item *i* se, *u* adquiriu *i* e *u* forneceu uma avaliação para *i* maior que sua média pessoal. O autor ainda argumenta que na prática esta abordagem é implementada como um algoritmo guloso. Dado $U = \{u_1, ..., u_m\}$, como sendo o conjunto de usuários, primeiramente o algoritmo classifica esses usuários em *x* nichos distintos, por meio de algoritmos de clusterização, que levam em conta as relações usuários-itens. Então, $F = \{S_{i1}, ..., S_{ik}\}$ é modelado como uma família de subconjuntos, onde S^*_i contém apenas os usuários que consumiram e gostaram do item *i*, sendo o objetivo encontrar uma subfamília $F^* = \{S_{i1}, ..., S_{ik}\}$ contendo os *k* itens que satisfazem todos os nichos (SILVA, 2018).

Silva (2018) analisa que: (1) as estratégias tradicionais são muito semelhantes por recomendarem itens idênticos; (2) *Max-Coverage* e *Niche-Coverage* são complementares as estratégias tradicionais, visto que recomendam itens distintos; (3) as recomendações feitas pelas novas estratégias são de itens menos populares que os apresentados pelas tradicionais. As experimentações feitas por Silva (2018) foram feitas usando as bases da CiaoDVD, relativa à venda de DVDs; Amazon, referente a venda de produtos relacionados a vídeo games; e a ML-1M e ML-10M, sobre o consumo de filmes. O autor avaliou certa superioridade de suas abordagens em relação aos demais SsR não-personalizados, pertencentes ao estado-da-arte, testados: *Max-Coverage* apresentou aumento de 5% de acurácia para o ML-10M e Amazon, e 47% para CiaoDVD, enquanto *Niche-Coverage* obteve aumento de 55% de acurácia para CiaoDVD.

O autor conclui que proprietários de aplicações de entretenimento ou e-commerce revejam suas estratégias utilizadas para a primeira interação com usuários novos no sistema, dados os resultados superiores de suas novas abordagens (SILVA, 2018). No entanto, o autor pretende avaliar em trabalhos futuros o desempenho dessas estratégias em sistemas reais. Silva (2018) também argumenta que os usuários existentes possuem o atual comportamento por terem iniciado sua interação com filmes populares, e alterar a primeira interação desses usuários abre questões como: (1) Como essas recomendações iniciais impactam no comportamento dos usuários? (2) Qual é o impacto dessa alteração na popularidade dos itens do domínio? e (3) Como serão a performance das estratégias atualmente consideradas estado-da-arte sobre esse novo cenário?

2.3 UM SISTEMA DE RECOMENDAÇÃO PARA E-COMMERCE UTILIZANDO REDES SOCIAIS PARA SOLUÇÃO DE COLD-START

Prando (2016) propõe um SR personalizado, como alternativa para amenizar o *cold-start*, baseando-se em dados minerados de redes sociais. O autor argumenta que existem poucos trabalhos que comprovam o desempenho desta abordagem, sendo sua principal contribuição a avaliação do uso de dados de redes sociais para sanar o *cold-start*.

A proposta de Prando (2016) é utilizar a interação de usuários em redes sociais para determinar suas preferências, fazendo um relacionamento entre três elementos obtidos das redes sociais: (1) postagens diretas, (2) curtidas em conteúdo, (3) curtidas em páginas. A partir desses dados o autor argumenta ser possível selecionar melhor os produtos, mesmo sem informações diretas sobre a sua avaliação. No entanto, os dados não-estruturados obtidos através do Facebook e Twitter, possuem vários formatos, mas o predominante é o tipo textual. Portanto, é necessário o uso do técnicas de mineração de texto e AM para relacionar as postagens dos usuários com produtos do e-commerce, o que caracteriza uma abordagem de recomendação baseada em conteúdo (PRANDO, 2016).

O autor formula o processo do SR da seguinte forma: (1) Inserir usuário, onde ocorre a inserção do usuário no banco de dados do SR pelo *e-commerce*, contendo a chave de segurança para acesso ao Facebook e Twitter; (2) Processar dados sociais, em que o SR extrai os dados sociais, realizando consultas a API do Facebook e/ou Twitter, usando a chave de segurança do usuário; (3) Treinar base de produtos aplicando AM, em que o SR aplica o algoritmo de AM para classificar os produtos em classes que possuem características semelhantes; (4) Processar a recomendação, onde o SR, utilizando o algoritmo de AM, classifica os dados sociais do usuário dentro das classes de produtos treinada; (5) Listar recomendação, onde o SR lista os produtos recomendados separados por categorias (PRANDO, 2016).

Prando (2016) avaliou o desempenho do seu SR com o auxílio de 98 participantes. Desse montante, 16 participantes não possuíam dados suficientes em suas redes sociais, e/ou o classificador não conseguiu encontrar a classe dos dados obtidos; 10 participantes tiveram problemas técnicos para avançar a etapa de avaliação das recomendações, por incompatibilidade do formulário com o navegador *web*; 72 participantes finalizaram todo o processo. Ao todo, foram 718 recomendações geradas pelo SR e avaliadas pelos participantes, obtendo o resultado de 1.71 pelo *Root Mean Square Error*, mostrando-se uma alternativa razoável para o *cold-start*.

3 PROPOSTA

A seguir é apresentada a justificativa para o desenvolvimento desse trabalho, os principais requisitos e a metodologia de desenvolvimento que será utilizada. Também são relacionados os assuntos e as fontes bibliográficas que irão fundamentar o estudo proposto.

3.1 JUSTIFICATIVA

É apresentado no Quadro 1 um comparativo entre os trabalhos correlatos. Nas colunas estão os trabalhos correlatos, e nas linhas suas características.

Quadro 1 – Comparativos entre os trabalhos correlatos.

Características \ Trabalhos	Almeida (2016)	Silva (2018)	Prando (2016)
Problema a ser resolvido	*Cold-start*	*Cold-start*	*Cold-start*
Objetivo	Verificar a similaridade de produtos	Recomendar itens abrangentes	Recomendar correlatos
Tipo de recomendação	Personalizada	Não-personalizada	Personalizada
Técnicas utilizadas	Programação Genética	*Max-Coverage* e *Niche-Coverage*	Algoritmos de Aprendizado de Máquina
Fonte de dados	Logs e catálogo de produtos do *e-commerce*	Logs e catálogo de produtos do *e-commerce*	Catálogo de produtos do *e-commerce* e Redes Sociais
Custo computacional	Médio	Médio	Alto

Fonte: elaborado pelo autor.

Conforme observado no Quadro 1, os trabalhos de Almeida (2016) e Prando (2016) abordam SsR personalizados para sanar o problema de *cold-start*, enquanto Silva (2018) propôs duas abordagens usando SsR não-personalizados. Silva (2018) argumenta que sua abordagem tem o intuito de complementar as recomendações feitas para novos usuários, aplicando em conjunto uma abordagem personalizada. O SsR de Prando (2016) torna-se custoso, visto que a obtenção dos dados sociais dos usuários ocorre logo após seu cadastro, necessitando de uma infraestrutura robusta para gerar as recomendações de novos usuários em um tempo plausível, atendendo as necessidades do *e-commerce*. As abordagens de Silva (2018) e Almeida (2016) adotam alternativas como o pré-processamento das recomendações, portanto, economizam recursos computacionais para performar as recomendações em tempo hábil ao contato com o usuário.

Enquanto o objetivo de Almeida (2016) foi o de propor uma abordagem para verificar produtos que são alternativas similares ao que está sendo visualizado pelo usuário, Prando (2016) tinha como objetivo encontrar produtos correlatos com base no perfil do usuário. O trabalho de Silva (2018) por sua vez, tinha o objetivo de recomendar produtos que fossem abrangentes ao maior número de usuários possíveis, com um SR não-personalizado não baseado apenas nos tradicionais produtos mais populares.

O grande diferencial de Prando (2016) é o uso de dados externos aos *logs* de atividades desempenhadas pelos usuários no *e-commerce*. Enquanto Silva (2018) e Almeida (2016) utilizam apenas os dados contidos no banco de dados do *e-commerce*, contendo histórico de visualização e venda dos produtos, catálogo de produtos, entre outros dados. Prando (2016) busca expandir sua base de dados disponíveis extraindo dados sociais dos clientes em suas respectivas redes sociais.

Almeida (2016) utiliza Programação Genética para sua abordagem *GPClerk* performar a comparação de similaridade entre produtos. Silva (2018) aplica a tradicional estratégia de *Maximum k-Coverage* e suas variantes, a *Max-Coverage* e a *Niche-Coverage*. Prando (2016) empregou algoritmos de AM supervisionados: Naive Bayes, Árvores de Decisão e Máquina de Vetores de Suporte.

Diante deste cenário, o trabalho proposto se difere dos demais pois irá avaliar a influência dos descontos dos produtos nos hábitos de consumo dos usuários, verificando se há o aumento de vendas, assim como, se a utilização de SsR efetivamente fazem com que as vendas também aumentem no e-commerce. Além disso, outra diferença, talvez a mais importante, é a integração com a API da Suaview, possibilitando validar o desempenho do SsR em um ambiente real com grande volume de dados e, ao mesmo tempo, agregando e correlacionando dados disponíveis em redes sociais. Destaca-se também que este trabalho deve contribuir com o aprimoramento de técnicas que mitiguem o *cold-start*, resultando no auxílio ao usuário nas suas escolhas, e na maior performance de vendas em aplicações de comércio eletrônico. Além disso, aumenta~~r~~ a satisfação do usuário, agiliza~~r~~ o processo de busca por produtos de seu interesse~~,~~ e os comprados com mais frequência, reduzindo o tempo necessário de navegação pelo site para efetivar uma compra.

3.2 REQUISITOS PRINCIPAIS DO PROBLEMA A SER TRABALHADO

O sistema de recomendação a ser desenvolvido deverá:

a) identificar, quando possível, o usuário que acessou o *e-commerce* (Requisito Funcional - RF);

b) acessar o endereço do perfil dos usuários nas redes sociais, por meio da chave fornecida pelas plataformas ou através de um *webcrawler* (RF);
c) realizar a redução, remoção de dados incompletos, ruídos e registros redundantes oriundos das redes sociais ou da base de dados da SuaView (RF);
d) calcular a similaridade de palavras-chave, amigos em comum, grau de valor e perfil do usuário (RF);
e) enquadrar o perfil do usuário em uma classe que remeta suas preferências (RF);
f) recomendar itens utilizando regras de associação/similaridade (RF);
g) gerar relatórios estatísticos a partir dos *logs* de acessos aos produtos e das efetivações de compra (RF);
h) ser implementado na linguagem Python (Requisito Não Funcional - RNF);
i) realizar a busca e extração dos dados de redes sociais em um tempo máximo de 30 segundos (RNF);
j) integrar com a API Suaview (RNF).

3.3 METODOLOGIA

O trabalho será desenvolvido observando as seguintes etapas:

a) levantamento bibliográfico: pesquisar trabalhos correlatos e estudar sobre sistemas de recomendação e *cold-start*;
b) elicitação de requisitos: baseando-se no levantamento bibliográfico, refinar os requisitos propostos para o sistema de recomendação proposto;
c) definição das informações e redes sociais: analisar quais informações são relevantes e que devem ser obtidas das redes sociais. Também se definirá quais redes serão acessadas;
d) obtenção dos dados das redes sociais: desenvolvimento de um *webcrawler* na linguagem Python, ou integração com a API da rede social, que busque as informações a partir da etapa (c);
e) integração com API Suaview: implementar a estrutura que irá se comunicar com a API Suaview para a obtenção dos dados;
f) preparação dos dados: realizar a limpeza e normalização dos dados coletados nas etapas (d) e (e);
g) definição das regras de associação: pesquisar e escolher modelos/técnicas/algoritmos que serão utilizados para correlacionar acessos, produtos e perfil do usuário;
h) desenvolvimento do sistema de recomendação: implementar o sistema de recomendação considerando as etapas anteriores, utilizando a linguagem Pyhton como base;
i) testes: avaliar a eficiência, a integração e o desempenho do sistema de recomendação com o auxílio de usuários voluntários considerando o *cold-start*.

As etapas serão realizadas nos períodos relacionados no Quadro 2.

Quadro 2 – Cronograma de atividades a serem realizadas

	2021									
	fev.		mar.		abr.		maio		jun.	
etapas / quinzenas	1	2	1	2	1	2	1	2	1	2
levantamento bibliográfico										
elicitação de requisitos										
definição das informações e redes sociais										
obtenção dos dados das redes sociais										
integração com API Suaview										
preparação dos dados										
definição das regras de associação										
desenvolvimento do sistema de recomendação										
testes										

Fonte: elaborado pelo autor.

4 REVISÃO BIBLIOGRÁFICA

Este capítulo está dividido em duas seções. A seção 4.1 aborda Sistemas de Recomendação. Já a seção 4.2 discorre sobre o problema conhecido como *cold-start*.

4.1 SISTEMAS DE RECOMENDAÇÃO

Sistemas de Recomendação (SsR) é uma área importante de pesquisa desde o surgimento dos primeiros trabalhos sobre filtragem colaborativa na década de 90, envolvendo áreas da ciência cognitiva, teoria da aproximação, recuperação da informação, mineração de dados, aprendizado de máquina, e tendo influências das ciências de *marketing* e administração (PRANDO, 2016). De acordo com Ricci *et al.* (2010), o aumento da

informação trouxe consigo uma ampla variedade de produtos e serviços, de diferentes nichos e níveis de qualidade, trazendo certa dificuldade aos consumidores na hora da escolha sobre qual produto comprar~~,~~ ou serviço adquirir. Para mitigar essa dificuldade, surgem os SsR como uma ferramenta de auxílio aos usuários no processo de escolha, fornecendo sugestões que mais tem probabilidade de satisfazê-los (PRANDO, 2016).

Segundo Silva (2018), o problema de recomendação pode ser definido como o desafio de estimar uma pontuação, que representa a utilidade de um item para os usuários, para itens que ainda não foram consumidos pelos usuários. Para Bobadilla *et al.* (2013), o processo para resolver esse problema consiste em criar predições sobre a utilidade dos itens, com base em fatores relevantes, como o histórico de navegação do usuário. Conforme Silva (2018), o objetivo é que a pontuação de utilidade do item, gerada pela predição do SR, seja o mais próximo possível da pontuação dada pelo usuário. Depois disso, o SR seleciona um grupo de itens com as melhores pontuações de utilidade para determinado usuário, gerando assim a recomendação.

De acordo com Akshita e Smita (2013 *apud* SILVA, 2018), os SsR possuem, de maneira geral, duas categorias principais: personalizadas~~;~~ e não-personalizadas. As abordagens personalizadas são aquelas que geram recomendações específicas para cada perfil de usuário, com base em informações que modelam seu perfil, tendo o intuito de recomendar itens que melhor se adequem a ele. Já as técnicas não-personalizadas tem a característica de não levar em consideração informações sobre o usuário alvo da recomendação, mas sim das informações de todos os usuários. Segundo Silva (2018) ambas as técnicas são relevantes para comércios eletrônicos, argumentando que não pode ser comparada a aplicabilidade entre elas, já que cada classe é aplicada em cenários distintos, e podem ser complementares a outra.

Formatado: Realce

Dentro da classe de SsR personalizados, existem três subclasses: Filtragem Baseada em Conteúdo; Filtragem Colaborativa, e Métodos Híbridos (BOBADILLA *et al*., 2013; BEEL *et al*., 2016). A Filtragem Colaborativa consiste na tentativa de correlacionar usuários com preferências em comum para gerar as recomendações, enquanto a Filtragem Baseada em Conteúdo busca gerar recomendações a partir das características dos itens consumidos, e os Métodos Híbridos buscam combinar ambas as técnicas (SCHAFER *et al*., 2007).

A técnica de Filtragem Baseada em Conteúdo busca estimar um *rating* que um usuário atribuiria a um produto, encontrando itens semelhantes a ele com base nas características, atributos, dos itens (VAN METEREN; VAN SOMEREN, 2000 *apud* SILVA, 2018). A fundamentação dessa abordagem parte de que itens com atributos similares são avaliados de maneira similar, já que cada usuários possui uma preferência sistemática correlacionada com os atributos dos itens (PAZZANI, 1999; SCHAFER *et al*., 2007; RICCI *et al*., 2010). O processo de recomendação pode ser dividido em três passos, conforme Bobadilla *et at.* (2013): (1) ~~E~~extração dos atributos relacionados aos itens do domínio, onde busca-se por dados capazes de descrever os itens, como gênero, lançamento, atores, no caso de filmes; (2) c~~C~~omparação dos atributos dos itens com a preferência do usuário alvo, em que procura-se por itens que encaixam no perfil do usuário com base nas características dos produtos consumidos por ele; (3) ~~R~~recomendação dos itens, onde são apresentados os itens com atributos que atraem o usuário alvo da recomendação.

Formatado: Realce

A classe de Filtragem Colaborativa é caracterizada pela tentativa de predizer a utilidade de um item baseando-se no *feedback* que o usuário atribuiu a itens semelhantes, ou o *feedback* de usuários semelhantes a demais itens (RICCI *et al*., 2010). Segundo Yang *et al.* (2014), essa abordagem pode ser dividida em dois grupos: *memory-based* e *model-based*, sendo que cada um dos grupos pode ser orientado, ou modelado, pelo usuário ou item. Técnicas *memory-based* orientadas pelo usuário, chamadas de *user-based*, buscam agregar os *ratings* dos produtos, avaliados por usuários semelhantes ao usuário alvo da recomendação, enquanto as técnicas *item-based*, orientadas pelos itens, procuram agregar os *ratings* recebidos aos produtos mais semelhantes ao produto alvo da recomendação (YANG *et al*., 2014). Por sua vez, a abordagem *model-based* caracteriza-se por usar os *ratings* para aprender a fazer predições, inspirando-se em algoritmos de ~~aprendizado~~ Aprendizado de ~~máquina~~Máquina, e sob as mesmas divisões de modelagens baseadas nos itens ou usuários que a abordagem *memory-based* (SILVA, 2018; BILLSUS; PAZZANI, 1998).

4.2 COLD-START

Um dos grandes problemas relacionados ao desenvolvimento de SsR, é o problema denominado *cold-start* (SCHAFER *et al*., 2007; ADOMAVICIUS; TUZHILIN, 2005). Uma vez que SsR costumam basear-se em dados do usuário visitando a aplicação *web* para gerar recomendações, o que ocorre quando o usuário acaba de se cadastrar no *site*, seu primeiro contato com o mesmo, é a falta de dados do usuário para gerar recomendações pertinentes, caracterizando o problema de *cold-start*. Este problema pode ser dividido em três cenários: (1) recomendações para usuários pouco participativos; (2) recomendações de itens pouco consumidos; (3) recomendações para usuários pouco participativos de itens pouco consumidos (LIKA *et al*., 2014). Conforme Silva

Comentado [AS4]: Confuso. Rever redação.

(2018), usuários pouco participativos não recebem recomendações pertinentes as suas preferências, causando a impressão de que o SsR não fornece o esperado.

Silva (2018) destaca que o problema de *Cold-Start* é muitas vezes confundido com o problema de *Pure Cold-Start*. Inicialmente, *Pure Cold-Start* fazia referência ao problema de gerar recomendações quando não existem dados suficientes sobre os usuários, e/ou itens (SCHEIN *et al*., 2002; BURKE, 2002). No entanto, trabalhos recentes relacionam *Cold-Start* ao problema de esparsidade dos dados, *Pure Cold-Start* passou a referenciar o cenário inicial do SsR, no qual não existe nenhuma informação disponível (SHAH; SAHU, 2014 *apud* SILVA, 2018; SHI, 2016). Sendo assim, *Pure Cold-Start*, é um subproblema de *Cold-Start*, e pode ser caracterizado de três formas: (1) gerar recomendações para usuários novos, que não possuem informações; (2) recomendar itens novos, que nunca foram consumidos; e (3) recomendar um item novo para um usuário novo (NGUYEN *et al*., 2007). Conforme Silva (2018), produtos recém adicionados no catálogo, bem como usuários sem histórico no sistema, os SsR não são capazes de lidar, além de tornar quase improvável recomendar um item novo a um novo usuário.

As abordagens de recomendação clássicas não são capazes de lidar com o problema de *Pure Cold-Start*, portanto, em geral para mitigar este problema a literatura indica três principais classes de SsR: (1) SsR Interativos; (2) SsR Híbridos; e (3) SsR não-personalizados (SILVA, 2018). Silva (2018) argumenta que de forma geral, SsR Interativos e SsR Híbridos utilizam dados demográficos, ou dados do usuário como sexo e faixa etária por exemplo, que são agregados a estratégias de Filtragem Colaborativa. Sistemas de Recomendação não-personalizados, em conformidade com Silva (2018), procuram utilizar informações globais do sistema para gerar as recomendações para usuários novos. Essa última categoria se distancia muito das demais por recomendarem os meus itens para todos os usuários do mesmo sistema, não considerando o perfil do usuário alvo.

A abordagem de Recomendação Interativa para construção de SsR consiste em criar um perfil provisório para o usuário com base em um *feedback* explícito fornecido pelo usuário, e então gerar recomendações personalizadas para aquele perfil (MAHMOOD; RICCI, 2007). Os sistemas da Netflix usam essa abordagem para tratar de usuários novos (SILVA, 2018). Conforme Silva (2018), o primeiro contato do novo usuário se dá pela coleta de 3 filmes/séries que o usuário elege como sendo atrativos a ele, então o sistema gera um perfil provisório para esse usuário, e a partir dessas informações geram-se as primeiras recomendações.

Recomendadores Híbridos se referem ao uso de abordagens clássicas de recomendação, porém com o acréscimo de informações externas ao sistema (BURKE, 2002). Essas informações externas têm o intuito de ajudar o SR a identificar as preferências do usuário, podendo ser informações sociais, demográficas, sexo, idade e outras (SILVA, 2018). Uma das estratégias praticadas, como demonstrado por Prando (2016), é o uso da conta do usuário nas redes sociais como forma de *login* para o *e-commerce*, assim obtendo os dados sociais dos usuários para modelar seu perfil e suas preferências. Em conformidade com Silva (2018), apesar de possuir várias vantagens para mitigar o *Pure Cold-Start*, essa abordagem limita-se à existência de dados externos. No caso das redes sociais, também ocorre que muitos usuários não querem vincular suas redes sociais com o *e-commerce*, além de existirem usuários sem conta em redes sociais, ou com poucos dados disponíveis nas mesmas (SILVA, 2018).

Sistemas de Recomendação não-personalizados, já mencionados no item 4.1, também são uma das abordagens para desenvolver SsR que mitiguem o problema de *Pure Cold-Start*. Silva (2018) faz analogia da estratégia não-personalizada com uma vitrine de uma loja física, onde o lojista escolhea seus melhores itens para expor, não visando agradar apenas uma pessoa em específico, mas sim ao maior número possível. Sendo essa uma estratégia baseada em dados globais do *e-commerce*, e com menor custo computacional em relação a abordagens personalizadas, acaba sendo muito usada pelos comércios eletrônicos como um mecanismo de lidar com o *Pure Cold-Start* (SILVA, 2018).

REFERÊNCIAS

ADOMAVICIUS, Gediminas.; TUZHILIN, Alexandre. Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. **IEEE Transactions On on Knowledge And and Data Engineering**, [S.l.], v. 17, n. 6, p. 734-749, jun. 2005.

ALMEIDA, Urique H. S. **Learning to Recommend Similar Alternative Products in e-Commerce Catalogs**. 2016. 67 f. Dissertação (Mestrado em Computação) – Programa de Pós-Graduação em Informática, Universidade Federal do Amazonas, Manaus.

BEEL, Joeran *et al*. Research-paper recommender systems: a literature survey. **International Journal On on Digital Libraries**, [S.l.], v. 17, n. 4, p. 305-338, 26 jul. 2015.

BILLSUS, Daniel.; PAZZANI, Michael. J. Learning collaborative information filters. **ICML**, [S.l.], v. 98, p. 46-54, 1998.

BOBADILLA, Jesús. *et al.* Recommender systems survey. **Knowledge-Based Systems**, Madrid, v. 46, p. 109-132, mar. 2013.

BURKE, Robin. Hybrid Recommender Systems: Survey and Experiments. **User Modeling ~~And~~ and User-Adapted Interaction**, [S.l.], v. 12, n. 4, p. 331-370, nov. 2002.

CHVATAL, Václav. A Greedy Heuristic for the Set-Covering Problem. **Mathematics ~~Of~~ of Operations Research**, [S.l.], v. 4, n. 3, p. 233-235, ago. 1979.

FAN, Wei; BIFET, Albert. Mining big data. **~~Acm~~ ACM Sigkdd Explorations Newsletter**, [S.l.], v. 14, n. 2, p. 1-5, abr. 2013.

HOCHBAUM, Dorit S.; PATHRIA, Anu. Analysis of the greedy approach in problems of maximum k-coverage. **Naval Research Logistics**, [S.l.], v. 45, n. 6, p. 615-627, mar. 1998.

LIKA, Blerina; KOLOMVATSOS, Kostas; HADJIEFTHYMIADES, Stathes. Facing the cold start problem in recommender systems. **Expert Systems with Applications**, [S.l.], v. 41, n. 4, p. 2065-2073, mar. 2014.

MAHMOOD, Tariq; RICCI, Francesco. Learning and adaptivity in interactive recommender systems. **Proceedings of the ninth international conference on Electronic commerce**, Minneapolis, p. 75-84, 2007.

NGUYEN, An-Te; DENOS, Nathalie; BERRUT, Catherine. Improving new user recommendations with rule-based induction on cold user data. **Proceedings of the 2007 ACM conference on Recommender systems**, Minneapolis, p. 121—128, out. 2007.

PRANDO, Alan. V. **Um Sistema de Recomendação para E-commerce Utilizando Redes Sociais para solução de cold-start**. 2016. 121 p. Dissertação (Mestrado em Engenharia de Computação) – Instituto de Pesquisas Tecnológicas do Estado de São Paulo, São Paulo.

PAZZANI, Michael. J. A framework for collaborative, content-based and demographic filtering. **Artificial intelligence review**, [S.l.], v 13, n. 5, p. 393-408, dez. 1999.

RICCI, Francesco; ROKACH, Lior; SHAPIRA, Bracha. Introduction to Recommender Systems Handbook. **Recommender Systems Handbook**, [S.l.], p. 1-35, out. 2010.

SCHAFER, Bem J. *et al*. Collaborative Filtering Recommender Systems. **The Adaptive Web**, [S.l.], v. 4321, p. 291-324, 2007.

SCHEIN, Andrew I. *et al*. Methods and metrics for cold-start recommendations. In: PROCEEDINGS OF THE 25TH ANNUAL INTERNATIONAL ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 2002, Tampere. **Proceedings...** Tampere: Association for Computing Machinery, 2002. p. 253-260.

SCHWARTZ, Barry. **The Paradox of Choice: Why More Is Less**. Nova Iorque: HarperCollins, 2011.

SHI, ShengBo. Real-time job recommendation engine based on college graduates' persona. **Journal of Residuals Science & Technology**, [S.l.], v. 13, 2016.

SILVA, Nícollas. C. **Sistemas de recomendação não-personalizados para atrair usuários novos**. 2018. 96 f. Dissertação (Mestrado em Computação) – Programa de Pós-Graduação em Ciência da Computação, Universidade Federal de Minas Gerais, Belo Horizonte.

SINGH, Sachchidanand; SINGH, Nirmala. Big data analytics. In: INTERNATIONAL CONFERENCE ON COMMUNICATION, INFORMATION & COMPUTING TECHONOLGY, 2012. Mumbai. **Proceedings…** Mumbai: IEEE, 2012. p. 1-4.

Comentado [AS5]: Não usar caixa alta.

Comentado [AS6]: Não usar caixa alta.

ASSINATURAS

(Atenção: todas as folhas devem estar rubricadas)

Assinatura do(a) Aluno(a): _____

Assinatura do(a) Orientador(a): _____

Assinatura do(a) Coorientador(a) (se houver): _____

Observações do orientador em relação a itens não atendidos do pré-projeto (se houver):

FORMULÁRIO DE AVALIAÇÃO – PROFESSOR TCC I

Acadêmico(a): Giulio Giovanella _____

Avaliador(a): Andreza Sartori_____

	ASPECTOS AVALIADOS[1]	atende	atende parcialmente	não atende
ASPECTOS TÉCNICOS	1. INTRODUÇÃO O tema de pesquisa está devidamente contextualizado/delimitado?	X		
	O problema está claramente formulado?	X		
	2. OBJETIVOS O objetivo principal está claramente definido e é passível de ser alcançado?	X		
	Os objetivos específicos são coerentes com o objetivo principal?	X		
	3. JUSTIFICATIVA São apresentados argumentos científicos, técnicos ou metodológicos que justificam a proposta?	X		
	São apresentadas as contribuições teóricas, práticas ou sociais que justificam a proposta?	X		
	4. METODOLOGIA Foram relacionadas todas as etapas necessárias para o desenvolvimento do TCC?	X		
	Os métodos, recursos e o cronograma estão devidamente apresentados?	X		
	5. REVISÃO BIBLIOGRÁFICA (atenção para a diferença de conteúdo entre projeto e pré-projeto) Os assuntos apresentados são suficientes e têm relação com o tema do TCC?	X		
ASPECTOS METODOLÓGICOS	6. LINGUAGEM USADA (redação) O texto completo é coerente e redigido corretamente em língua portuguesa, usando linguagem formal/científica?	X		
	A exposição do assunto é ordenada (as ideias estão bem encadeadas e a linguagem utilizada é clara)?	X		
	7. ORGANIZAÇÃO E APRESENTAÇÃO GRÁFICA DO TEXTO A organização e apresentação dos capítulos, seções, subseções e parágrafos estão de acordo com o modelo estabelecido?	X		
	8. ILUSTRAÇÕES (figuras, quadros, tabelas) As ilustrações são legíveis e obedecem às normas da ABNT?	X		
	9. REFERÊNCIAS E CITAÇÕES As referências obedecem às normas da ABNT?	X		
	As citações obedecem às normas da ABNT?		X	
	Todos os documentos citados foram referenciados e vice-versa, isto é, as citações e referências são consistentes?	X		

PARECER – PROFESSOR DE TCC I OU COORDENADOR DE TCC
(PREENCHER APENAS NO PROJETO):

O projeto de TCC será reprovado se:
- qualquer um dos itens tiver resposta NÃO ATENDE;
- pelo menos **4 (quatro)** itens dos **ASPECTOS TÉCNICOS** tiverem resposta ATENDE PARCIALMENTE; ou
- pelo menos **4 (quatro)** itens dos **ASPECTOS METODOLÓGICOS** tiverem resposta ATENDE PARCIALMENTE.

PARECER: (x) APROVADO ( ) REPROVADO

Assinatura: _____ Data: 09/12/2020 _____

[1] Quando o avaliador marcar algum item como atende parcialmente ou não atende, deve obrigatoriamente indicar os motivos no texto, para que o aluno saiba o porquê da avaliação.