

CURSO DE CIÊNCIA DA COMPUTAÇÃO – TCC		
( X ) PRÉ-PROJETO	( ) PROJETO	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 (2012), a grande maior parte dos dados, cerca de 90%, foram gerados nos últimos anos. Parte das ações dos usuários navegando na internet são capturadas, gerando cada vez mais volumosos dados. Redes sociais, dispositivos móveis, e tecnologias como Internet das Coisas, geram uma explosão de dados, criando o cenário chamado de Big Data (FAN; BIFET, 2012).

**Comentado [ARV1]:** De que dados mesmo estamos falando?

Com a enorme quantidade de opções nos modernos Web Sites, o usuário possui mais possibilidades do que pode lidar (SCHWARTZ, 2009; RICCI et al., 2015). Diante deste contexto, os Sistemas de Recomendação (SR) ganham uma função primordial de poupar o usuário de horas gastas em sua escolha, bem como aumentar seu engajamento à plataforma que este 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 o e-commerce que os novos usuários se sintam atraídos pelos produtos apresentados, caso contrário tendem a parar de acessar a página. Diante deste contexto, Silva (2018), ressalta que em dos grandes desafios é como atrair os novos usuários com recomendações pertinentes a ele sem nenhum desses clientes, tal problema é conhecido como *cold-start*.

Silva (2018) também afirma que uma das soluções tradicionais para tentar atrair novos usuários, é 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 com essa 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). 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 (2012), as modernas soluções de Big Data e *Machine Learning* 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 para minimizar, para 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), propõem a técnica GPCLerk, que utiliza Programação Genética para encontrar produtos similares utilizando os atributos dos produtos para efetuar comparações.

**Comentado [ARV2]:** o que é "cold-star" mesmo?

Diante de cenário, este trabalho propõe o desenvolvimento de um sistema de recomendação que minimiza o *cold-start*, utilizando da Ciência dos Dados e Aprendizado de Máquina para recomendações para usuários com dados de navegação e compra disponíveis, além de avaliar a influência da aplicação de descontos no comportamento de consumo 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:

- utilizar dados externos ao e-commerce para minimizar o *cold-start*;
- recomendar produtos correlatos usando regras de associação;
- avaliar a influência de descontos em produtos na compra dos consumidores;
- integrar o SR com a API Suaview.

## 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*,  $UV = S \cap \bigcup_{i,j} V_{i,j}$  é o conjunto de todos os produtos acessados pelos usuários, e podem ser obtidos através dos dados capturados pelo *e-commerce* (ALMEIDA, 2016). Sua afirmação é que existe uma interseção  $S \cap UV$  não vazia, visto que uma das principais motivações de um usuário navegar por produtos diferentes em uma mesma sessão, é que eles são alternativas similares.

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 de Sistemas de Recomendação 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  $U = \{u_1, \dots, u_m\}$  sendo os usuários cadastrados no sistema, e  $I = \{I_1, \dots, I_n\}$  sendo os itens disponíveis no catálogo de produtos, modela-se um conjunto  $F = \{S_1, \dots, S_n\}$  em que cada subconjunto  $S_i$  é relativo a um item existente (SILVA, 2018). Dessa forma, o subconjunto  $S_i$  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  $F^* = \{S_{i_1}, \dots, S_{i_k}\}$ , 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. Porém, o autor argumenta que com uma heurística gulosa simples, que a cada iteração maximiza o número de usuários não cobertos, possui uma razão de aproximadamente 63% da solução ótima, como demonstrado por Chvatal (1979).

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  $U = \{u_1, \dots, u_m\}$ , 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,  $F = \{S_{i1}, \dots, S_{ik}\}$  é modelado como uma família de subconjuntos, onde  $S^*_{i1}$  contém apenas os usuários que consumiram e gostaram do item  $i$ , sendo o objetivo encontrar uma subfamília  $F^* = \{S_{i1}, \dots, S_{ik}\}$  contendo os  $k$  itens que satisfazem todos os nichos ( $N$ ) (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, relativa a venda de produtos relacionados a vídeo games, ML-1M e ML-10M, relativas ao consumo de filmes. O autor avaliou certa superioridade de suas abordagens em relação aos demais SR não-personalizados estado-da-arte testados: *Max-Coverage* apresenta ganhos de 5% de acurácia para o ML-10M e Amazon, e 47% para CiaoDVD, enquanto *Niche-Coverage* obteve ganhos 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 Prando argumenta ser possível melhor selecionar produtos, mesmo sem informações diretas sobre a avaliação dos produtos. 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 de técnicas de mineração de texto e Aprendizado de Máquina (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 em sua pesquisa. 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.

Trabalhos Características	Almeida (2016)	Silva (2018)	Prando (2016)
Problema a ser resolvido	<i>Cold-start</i>	<i>Cold-start</i>	<i>Cold-start</i>
Objetivo	Verificar a similaridade de produtos	Recomendar itens abrangentes	Recomendar correlatos
Tipo de recomendação	Personalizada	Não-personalizada	Personalizada
Técnicas utilizadas	Programação Genética	<i>Max-Coverage</i> e <i>Niche-Coverage</i>	Algoritmos de aprendizado de Máquina
Fonte de dados	Logs e catálogo de produtos do <i>e-commerce</i>	Logs e catálogo de produtos do <i>e-commerce</i>	Catálogo de produtos do <i>e-commerce</i> 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 SR personalizados para sanar o problema de *cold-start*, enquanto Silva (2018) propôs duas abordagens para sanar o *cold-start* usando SR não-personalizados. Silva (2018) argumenta que sua abordagem tem o intuito de complementar as recomendações feitas para novos usuários, cabendo outra abordagem personalizada sendo aplicada em conjunto. O SR de Prando (2016) torna-se custoso, visto que é a obtenção dos dados sociais dos usuários ocorre logo após seu cadastro, necessitando 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, economizando recursos computacionais para performar as recomendações em tempo hábil ao contato com o usuário.

Enquanto o objetivo de Almeida (2016) foi 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. 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 Aprendizado de Máquina supervisionados: Naive Bayes, Árvores de Decisão e *Support Vector Machine*.

Diante deste cenário, o trabalho proposto se difere dos demais pois irá avaliar a influência dos descontos nos hábitos de consumo dos usuários, verificando se há o aumento de vendas, assim como, se a utilização de SR efetivamente fazem com que as vendas também aumentem no *e-commerce*. Além disso, outra diferença, talvez a mais importante, é a integração com o *e-commerce* da Suaview, possibilitando validar o desempenho do SR em um ambiente real com grande volume de dados e, ao mesmo tempo, agregando e correlacionando dados disponíveis em redes sociais.

### 3.2 REQUISITOS PRINCIPAIS DO PROBLEMA A SER TRABALHADO

O sistema de recomendação a ser desenvolvido deverá:

- identificar, quando possível, o usuário que acessou o *e-commerce* (Requisito Funcional - RF);
- utilizar um *webcrawler* para extrair dados de redes sociais (RF);
- recomendar itens regras de associação (RF);
- gerar relatórios estatísticos a partir dos *logs* de acessos aos produtos e das efetivações de compra (RF);
- ser implementado na linguagem Python (Requisito Não Funcional - RNF);
- realizar a busca e extração dos dados de redes sociais em um tempo máximo de 30 segundos (RNF);
- integrar com a API Suaview (RNF).

**Comentado [ARV3]:** utilizar ou desenvolver? Se desenvolver, quais os requisitos dele?

**Comentado [ARV4]:** e a interação com redes sociais? não é parte do desenvolvimento?

### 3.3 METODOLOGIA

O trabalho será desenvolvido observando as seguintes etapas:

- levantamento bibliográfico: pesquisar trabalhos correlatos e estudar sobre sistemas de recomendação e *cold-start*;
- elicitação de requisitos: baseando-se no levantamento bibliográfico, refinar os requisitos propostos para o sistema de recomendação proposto;
- 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;
- desenvolvimento do *webcrawler*: implementar um *script* na linguagem Python que busque as informações a partir da etapa (c);
- integração com API Suaview: implementar a estrutura que irá se comunicar com a API Suaview para a obtenção dos dados;
- preparação dos dados: realizar a limpeza e normalização dos dados coletados nas etapas (d) e (e);
- definição das regras de associação: pesquisar e escolher modelos/técnicas/algoritmos que serão utilizados para correlacionar acessos, produtos e perfil do usuário;
- desenvolvimento do sistema de recomendação: implementar o sistema de recomendação considerando as etapas anteriores, utilizando a linguagem Python como base;
- 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

etapas / quinzenas	2021									
	fev.		mar.		abr.		maio		jun.	
	1	2	1	2	1	2	1	2	1	2
levantamento bibliográfico										
elicitação de requisitos										
definição das informações e redes sociais										
desenvolvimento do <i>webcrawler</i>										
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 descreve brevemente sobre os assuntos que fundamentarão o estudo a ser realizado: sistemas de recomendação e o problema conhecido como *cold-start*.

Sistemas de Recomendação (SR) 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). 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 (RICCI *et al.* 2011). Para mitigar

essa dificuldade, surgem os SR como uma ferramenta de auxílio aos usuários no processo de escolha, fornecendo sugestões que mais tem probabilidade de satisfazê-los (PRANDO, 2016).

Um dos grandes problemas relacionados ao desenvolvimento de SR, 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).

#### REFERÊNCIAS

- ADOMAVICIUS, G.; TUZHILIN, A. Toward the next generation of recommender systems: **A survey of the state-of-the-art and possible extensions**. IEEE transactions on knowledge and data engineering, 2005. 17(6):734-749.
- ALMEIDA, U. 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.
- CHVATAL, V. A greedy heuristic for the set-covering problem. Mathematics of operations research, 1979. 4(3):233--235.
- FAN, W.; BIFET, A. Mining big data: **current status, and forecast to the future**. ACM SIGKDD Explorations Newsletter, V.14, n.2, p.1-5, Dec. 2012.
- HAMMOND, M. I et al. **Multiple user interfaces of an artificial intelligence system to accommodate different types of users solving different types of problems with artificial intelligence**. 2017. US Patent App. 15/417,033.
- HOCHBAUM, D. S.; PATHRIA, A. (1998). Analysis of the greedy approach in problems of maximum k-coverage. Naval Research Logistics, 45(6):615--627.
- LIKA, B. et al. **Facing the cold start problem in recommender systems**. Expert Systems with Applications, 2014. 41(4):2065--2073.
- PRANDO, A. 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.
- RICCI, F. et al. Introduction to Recommender Systems Handbook. In: RICCI, F.; ROKACH, L.; SHAPIRA, B. Recommender Systems Handbook. Springer, 2011. p.1-35.
- SCHAFER, J. B.; FRANKOWSKI, D.; HERLOCKER, J. & SEN, S. (2007). Collaborative filtering recommender systems. Em The adaptive web, pp. 291--324. Springer. ramos
- SCHWARTZ, B. (2009). The Paradox of Choice: Why More Is Less, Revised Edition. HarperCollins. ISBN 9780061748998.
- SILVA, N. 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, S. Big data analytics. International Conference on Communication, Information & Computing Technology (ICICT), p.1-4, Oct. 2012.
- .

**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 AVALIADOR

Acadêmico(a): \_\_\_\_\_

Avaliador(a): \_\_\_\_\_

ASPECTOS AVALIADOS <sup>1</sup>		atende	atende parcialmente	não atende
ASPECTOS TÉCNICOS	1. INTRODUÇÃO O tema de pesquisa está devidamente contextualizado/delimitado?	x		
	O problema está claramente formulado?	x		
	1. OBJETIVOS O objetivo principal está claramente definido e é passível de ser alcançado?	x		
	Os objetivos específicos são coerentes com o objetivo principal?	x		
	2. TRABALHOS CORRELATOS São apresentados trabalhos correlatos, bem como descritas as principais funcionalidades e os pontos fortes e fracos?	x		
	3. JUSTIFICATIVA Foi apresentado e discutido um quadro relacionando os trabalhos correlatos e suas principais funcionalidades com a proposta apresentada?	x		
	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. REQUISITOS PRINCIPAIS DO PROBLEMA A SER TRABALHADO Os requisitos funcionais e não funcionais foram claramente descritos?			x
	5. METODOLOGIA Foram relacionadas todas as etapas necessárias para o desenvolvimento do TCC?		x	
	Os métodos, recursos e o cronograma estão devidamente apresentados e são compatíveis com a metodologia proposta?	x		
	6. 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	As referências contemplam adequadamente os assuntos abordados (são indicadas obras atualizadas e as mais importantes da área)?	x		
	7. 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		

### PARECER – PROFESSOR AVALIADOR: (PREENCHER APENAS NO PROJETO)

O projeto de TCC ser deverá ser revisado, isto é, necessita de complementação, se:

- qualquer um dos itens tiver resposta NÃO ATENDE;
- pelo menos 5 (cinco) tiverem resposta ATENDE PARCIALMENTE.

**PARECER:** ( ) APROVADO ( x ) REPROVADO

Assinatura: \_\_\_\_\_ Data: \_\_\_\_\_

<sup>1</sup> 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.



CURSO DE CIÊNCIA DA COMPUTAÇÃO – TCC		
( X ) PRÉ-PROJETO	( ) PROJETO	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 (2012), a grande ~~maior~~ parte dos dados, cerca de 90%, foram gerados nos últimos anos. Parte das ações dos usuários navegando na internet são capturadas, gerando cada vez mais volumosos dados. Redes sociais, dispositivos móveis, e tecnologias como Internet das Coisas, gera ~~ram~~ uma explosão de dados, criando o cenário chamado de Big Data (FAN; BIFET, 2012).

Com a enorme quantidade de opções nos modernos Web Sites, o usuário possui mais possibilidades do que pode lidar (SCHWARTZ, 2009; RICCI et al., 2015). Diante deste contexto, os Sistemas de Recomendação (SR) ganham ~~uma~~ função primordial de poupar o usuário de horas gastas em ~~sua~~ escolhas, bem como aumentar seu engajamento à plataforma que ~~este~~ 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 o e-commerce que os novos usuários se sintam atraídos pelos produtos apresentados, caso contrário ~~tendem a parar~~ de acessar a página. Diante deste contexto, Silva (2018) ressalta que ~~em um~~ dos grandes desafios é ~~eomo~~ atrair os novos usuários com recomendações pertinentes a ele sem nenhum desses clientes, tal problema é conhecido como *cold-start*.

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 ~~com~~ 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). 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 (2012), as modernas soluções de Big Data e *Machine Learning* 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 para minimizar, ~~para~~ 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), propõem a técnica GPClerk, que utiliza Programação Genética para encontrar produtos similares utilizando os atributos dos produtos ~~para a fim de~~ efetuar comparações.

Diante de cenário, este trabalho propõe o desenvolvimento de um sistema de recomendação que minimiza o *cold-start*, utilizando da Ciência dos Dados e Aprendizado de Máquina para recomendações para usuários com dados de navegação e compra disponíveis, além de avaliar a influência da aplicação de descontos no comportamento de consumo 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:

- utilizar dados externos ao e-commerce para minimizar o *cold-start*;
- recomendar produtos correlatos usando regras de associação;
- avaliar a influência de descontos em produtos na compra dos consumidores;
- integrar o SR com a API Suaview.

**Comentado [AS1]:** Frase confusa. Rever.

**Comentado [AS2]:** website

**Comentado [AS3]:** itálico. Verificar em todo o texto.

**Comentado [AS4]:** itálico. Verificar em todo o texto.

**Comentado [AS5]:** As plataformas de e-commerce?

**Comentado [AS6]:** Quem?

**Comentado [AS7]:** não existe vírgula entre sujeito e verbo

**Comentado [AS8]:** confuso. Rever frase.

**Comentado [AS9]:** Podes colocar o termo em português: Aprendizado de Máquina

**Comentado [AS10]:** Frase longa e confusa. Rever. Podes dividir em 2 para ficar mais claro.

**Comentado [AS11]:** Isso é requisito.

## 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 selecionados, 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*,  $UV = S \forall i,j \forall 2 i,j$  é o conjunto de todos os produtos acessados pelos usuários, e podem ser obtidos através dos dados capturados pelo *e-commerce* (ALMEIDA, 2016). Sua afirmação é que existe uma interseção  $S \cap UV$  não vazia, visto que uma das principais motivações de um usuário navegar por produtos diferentes em uma mesma sessão, é que eles são alternativas similares.

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 de Sistemas de Recomendação SR 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  $U = \{u_1, \dots, u_m\}$  sendo os usuários cadastrados no sistema, e  $I = \{I_1, \dots, I_n\}$  sendo os itens disponíveis no catálogo de produtos, modela-se um conjunto  $F = \{S_1, \dots, S_n\}$  em que cada subconjunto  $S_i$  é relativo a um item existente (SILVA, 2018). Dessa forma, o subconjunto  $S_i$  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  $F^* = \{S_{i_1}, \dots, S_{i_k}\}$ , 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. Porém, o autor argumenta que com uma heurística gulosa simples, que a cada iteração maximiza o número de usuários não cobertos, possui uma razão de aproximadamente 63% da solução ótima, como demonstrado por Chvatal (1979).

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

**Comentado [AS12]:** Sem a fórmula fica difícil de entender. Acredito que seria melhor explicar de forma descritiva a resolução.

**Comentado [AS13]:** Mesma questão do comentário anterior.

**Comentado [AS14]:** Frase confusa. Rever.

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  $U = \{u_1, \dots, u_m\}$ , 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,  $F = \{S_{i1}, \dots, S_{ik}\}$  é modelado como uma família de subconjuntos, onde  $S^*_{i1}$  contém apenas os usuários que consumiram e gostaram do item  $i$ , sendo o objetivo encontrar uma subfamília  $F^* = \{S_{i1}, \dots, S_{ik}\}$  contendo os  $k$  itens que satisfazem todos os nichos ( $N$ ) (SILVA, 2018).

Silva (2018) analisa que: (1) as estratégias tradicionais são muito semelhantes por recomendar ~~em~~ 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, ~~relativa~~ a venda de produtos relacionados a vídeo games; e a ML-1M e ML-10M, ~~relativas~~ ao consumo de filmes. O autor avaliou certa superioridade de suas abordagens em relação aos demais SR não-personalizados estado-da-arte testados. *Max-Coverage* apresenta ganhos de 5% de acurácia para o ML-10M e Amazon, e 47% para CiaoDVD, enquanto *Niche-Coverage* obteve ganhos 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 Prando-o autor argumenta ser possível ~~melhor~~ selecionar ~~melhor os~~ produtos, mesmo sem informações diretas sobre a ~~sua~~ avaliação-dos-produtos. 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 de técnicas de mineração de texto e ~~Aprendizado de Máquina~~ (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 ~~em sua pesquisa~~. 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*.

Comentado [AS15]: Ver comentário do correlato 1

Comentado [AS16]: Procure não repetir as palavras várias vezes na mesma frase.

Formatado: Realce

Formatado: Realce

Formatado: Realce

Comentado [AS17]: Rever.

Comentado [AS18]: Ganhos baseados em que?

Comentado [AS19]: Não é a primeira vez que você coloca este termo. Coloque a definição de sigla lá no início do texto.

### 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.

Trabalhos Características	Almeida (2016)	Silva (2018)	Prando (2016)
Problema a ser resolvido	<i>Cold-start</i>	<i>Cold-start</i>	<i>Cold-start</i>
Objetivo	Verificar a similaridade de produtos	Recomendar itens abrangentes	Recomendar correlatos
Tipo de recomendação	Personalizada	Não-personalizada	Personalizada
Técnicas utilizadas	Programação Genética	<i>Max-Coverage</i> e <i>Niche-Coverage</i>	Algoritmos de aprendizado de Máquina
Fonte de dados	Logs e catálogo de produtos do <i>e-commerce</i>	Logs e catálogo de produtos do <i>e-commerce</i>	Catálogo de produtos do <i>e-commerce</i> 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 SR personalizados para sanar o problema de *cold-start*, enquanto Silva (2018) propôs duas abordagens para sanar o *cold-start* usando SR não-personalizados. Silva (2018) argumenta que sua abordagem tem o intuito de complementar as recomendações feitas para novos usuários, cabendo outra abordagem personalizada sendo aplicada em conjunto. O SR de Prando (2016) torna-se custoso, visto que é a obtenção dos dados sociais dos usuários ocorre logo após seu cadastro, necessitando 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, economizando 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. 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 Aprendizado de Máquina supervisionados: Naive Bayes, Árvores de Decisão e *Support Vector Machine*.

Diante deste cenário, o trabalho proposto se difere dos demais pois irá avaliar a influência dos descontos nos hábitos de consumo dos usuários, verificando se há o aumento de vendas, assim como, se a utilização de SR efetivamente faz com que as vendas também aumentem no *e-commerce*. Além disso, outra diferença, talvez a mais importante, é a integração com o *e-commerce* da Suaview, possibilitando validar o desempenho do SR em um ambiente real com grande volume de dados e, ao mesmo tempo, agregando e correlacionando dados disponíveis em redes sociais.

Comentado [AS20]: Coloque o recurso de referência cruzada para os quadros.

Comentado [AS21]: O trabalho de Silva

Comentado [AS22]: Coloque o nome em português

Comentado [AS23]: Descontos de que?

Comentado [AS24]: contribuições teóricas, práticas ou sociais que justificam a proposta?

### 3.2 REQUISITOS PRINCIPAIS DO PROBLEMA A SER TRABALHADO

O sistema de recomendação a ser desenvolvido deverá:

- identificar, quando possível, o usuário que acessou o *e-commerce* (Requisito Funcional - RF);
- utilizar um *webcrawler* para extrair dados de redes sociais (RF);
- recomendar itens regras de associação (RF);
- gerar relatórios estatísticos a partir dos *logs* de acessos aos produtos e das efetivações de compra (RF);
- ser implementado na linguagem Python (Requisito Não Funcional - RNF);
- realizar a busca e extração dos dados de redes sociais em um tempo máximo de 30 segundos (RNF);
- integrar com a API Suaview (RNF).

### 3.3 METODOLOGIA

O trabalho será desenvolvido observando as seguintes etapas:

- levantamento bibliográfico: pesquisar trabalhos correlatos e estudar sobre sistemas de recomendação e *cold-start*;
- elicitação de requisitos: baseando-se no levantamento bibliográfico, refinar os requisitos propostos para o sistema de recomendação proposto;
- 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;
- desenvolvimento do *webcrawler*: implementar um *script* na linguagem Python que busque as informações a partir da etapa (c);
- integração com API Suaview: implementar a estrutura que irá se comunicar com a API Suaview para a obtenção dos dados;
- preparação dos dados: realizar a limpeza e normalização dos dados coletados nas etapas (d) e (e);
- definição das regras de associação: pesquisar e escolher modelos/técnicas/algoritmos que serão utilizados para correlacionar acessos, produtos e perfil do usuário;
- desenvolvimento do sistema de recomendação: implementar o sistema de recomendação considerando as etapas anteriores, utilizando a linguagem Python como base;
- 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](#).

**Comentado [AS25]:** Coloque o recurso de referência cruzada para os quadros.

Quadro 2 – Cronograma de atividades a serem realizadas

etapas / quinzenas	2021									
	fev.		mar.		abr.		maio		jun.	
	1	2	1	2	1	2	1	2	1	2
levantamento bibliográfico										
elicitação de requisitos										
definição das informações e redes sociais										
desenvolvimento do <i>webcrawler</i>										
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 descreve brevemente sobre os assuntos que fundamentarão o estudo a ser realizado: **sistemas** Sistemas de **recomendação** Recomendação e o problema conhecido como *cold-start*.

Sistemas de Recomendação (SR) 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). 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 (RICCI *et al.* 2011). Para mitigar

essa dificuldade, surgem os SR como uma ferramenta de auxílio aos usuários no processo de escolha, fornecendo sugestões que mais tem probabilidade de satisfazê-los (PRANDO, 2016).

Um dos grandes problemas relacionados ao desenvolvimento de SR, 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).

#### REFERÊNCIAS

- ADOMAVICIUS, G.; TUZHILIN, A. Toward the next generation of recommender systems: **A survey of the state-of-the-art and possible extensions**. IEEE transactions on knowledge and data engineering, 2005. 17(6):734-749.
- ALMEIDA, U. 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.
- CHVATAL, V. A greedy heuristic for the set-covering problem. Mathematics of operations research, 1979. 4(3):233--235.
- FAN, W.; BIFET, A. Mining big data: **current status, and forecast to the future**. ACM SIGKDD Explorations Newsletter, V.14, n.2, p.1-5, Dec. 2012.
- HAMMOND, M. I et al. **Multiple user interfaces of an artificial intelligence system to accommodate different types of users solving different types of problems with artificial intelligence**. 2017. US Patent App. 15/417,033.
- HOCHBAUM, D. S.; PATHRIA, A. (1998). Analysis of the greedy approach in problems of maximum k-coverage. Naval Research Logistics, 45(6):615--627.
- LIKA, B. et al. **Facing the cold start problem in recommender systems**. Expert Systems with Applications, 2014. 41(4):2065--2073.
- PRANDO, A. 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.
- RICCI, F. et al. Introduction to Recommender Systems Handbook. In: RICCI, F.; ROKACH, L.; SHAPIRA, B. Recommender Systems Handbook. Springer, 2011. p.1-35.
- SCHAFER, J. B.; FRANKOWSKI, D.; HERLOCKER, J. & SEN, S. (2007). Collaborative filtering recommender systems. Em The adaptive web, pp. 291--324. Springer. ramos
- SCHWARTZ, B. (2009). The Paradox of Choice: Why More Is Less, Revised Edition. HarperCollins. ISBN 9780061748998.
- SILVA, N. 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, S. Big data analytics. International Conference on Communication, Information & Computing Technology (ICCICT), p.1-4, Oct. 2012.

Comentado [AS26]: Frase longa e confusa. Rever.

Comentado [AS27]: negrito deve ser no título do documento.

**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 <sup>1</sup>		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
ASPECTOS METODOLÓGICOS	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		
	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:** ( ) APROVADO ( ) REPROVADO

Assinatura: \_\_\_\_\_ Data: 23/10/2020 \_\_\_\_\_

<sup>1</sup> 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.