PROPOSIÇÃO DE UM SISTEMA DE RECOMENDAÇÃO PARA UM *E-COMMERCE* DO RAMO LITERÁRIO

Lucas de Oliveira Leite - leite.lucas.95@gmail.com
Marcelo Nogueira Cortimiglia - cortimiglia@gmail.com
Rodrigo Dalla Vecchia - rodrigovecchia@gmail.com

RESUMO

A experiência de consumo nas plataformas vinculadas ao comércio eletrônico está sendo comprometida devido à alta complexidade da oferta de produtos e serviços. A implementação de um sistema de recomendação dentro desse contexto se torna interessante na medida em que ele pode auxiliar na melhora da experiência de consumo dos clientes, reduzindo os riscos deles realizarem decisões equivocadas diante de um significativo volume de dados e informações. Tendo isso em vista, este trabalho tem como objetivo desenvolver um sistema de recomendação adaptado ao contexto da loja virtual operada pela empresa TAG Livros, avaliando posteriormente se a sua performance, em termos de qualidade, é suficientemente atraente para uma potencial implementação. Os resultados apresentam o desenvolvimento de dois tipos de sistemas (baseado em filtragem de conteúdo e filtragem colaborativa), sendo o último mais apropriado para a implementação devido a sua qualidade superior e a facilidade em conseguir dados para abastecer o sistema.

1. INTRODUÇÃO

O comércio eletrônico apresenta um contexto de alta complexidade no que diz respeito à oferta de produtos e serviços, comprometendo a experiência de consumo nessas plataformas (ALJUKHADAR et al., 2012). Contribuem para esse cenário o crescimento acelerado do volume de dados e da variedade de informações disponíveis na internet, aliado ao vasto número de serviços disponibilizados nos sites de compras online, tais como: comparação de produtos, leilão de ofertas e avaliação de produtos (SHAH et al., 2017). Como consequência, surgiu a necessidade de desenvolver sistemas computacionais que providenciassem recomendações automáticas a partir de filtros aplicados em todas as opções de compra da

plataforma, reduzindo os riscos dos usuários realizarem decisões equivocadas (SHAH et al., 2017).

Resumidamente, os Sistemas de Recomendação (SRs) são recursos e técnicas desenvolvidos em softwares que servem para reter e filtrar informações no intuito de prover sugestões significativas para os seus usuários, por meio das suas interações com a plataforma (RICCI et al., 2011). Os algoritmos relacionados aos SRs frequentemente operam em ambientes desafiadores, onde os resultados devem ser exibidos em tempo real (não demorando mais que meio segundo) e as informações dos usuários são geralmente limitadas e voláteis - cada interação com a plataforma fornece dados valiosos que precisam ser processados pelos algoritmos e transformados em informações imediatamente (LINDEN et al., 2003). Por conseguinte, esses sistemas podem ser aplicados em diversos processos de tomada de decisão, como a escolha de uma notícia para ler, uma música para escutar, um filme para assistir ou um produto para comprar (SHAH et al., 2017).

No campo das recomendações para compras online de produtos, um dos *cases* de maior reconhecimento é o da empresa Amazon.com. Com o objetivo de oferecer uma experiência de consumo personalizada para cada cliente, sua loja virtual muda radicalmente com base nos interesses identificados, mostrando livros de programação para engenheiros de computação e brinquedos de bebê para gestantes (LINDEN et al., 2003). Como resultado, o SR da empresa apresenta desempenho muito superior em dois dos principais indicadores utilizados para medir a eficiência de publicidades baseadas na *web* (taxa de conversão e taxa de cliques), quando comparado a outros recursos de propaganda como banner e lista de mais vendidos (LINDEN et al., 2003).

Assim como o método da Amazon.com, outros mais populares possuem, cada um, vantagens e desvantagens particulares quanto às suas implementações e resultados, provendo margem para a aplicação de melhorias (SHAH et al., 2017). Estes aperfeiçoamentos incluem o desenvolvimento de métodos menos intrusivos e mais flexíveis, que melhor representam os comportamentos de usuários e informações sobre itens, incorporação de mais informações contextuais no processo de recomendação e utilização de avaliações multicriteriais (ADOMAVICIUS; TUZHILIN, 2005). Tendo isso em vista, o objetivo deste trabalho é

propor um sistema de recomendação para a loja virtual recém criada pela empresa Tag Livros, de forma a oferecer uma experiência de consumo personalizada para os seus clientes.

Após a introdução, este artigo também apresenta uma segunda seção destinada à revisão da literatura sobre os principais tópicos que abrangem os sistemas de recomendação. A seção 3 descreve o método empregado neste trabalho e a seção 4 os resultados obtidos. Ao final, na seção 5, os resultados são resumidos e apresentados junto à conclusão do artigo.

2. REFERENCIAL TEÓRICO

Esta seção do artigo apresenta os conceitos básicos que permeiam os sistemas de recomendação, partindo de um *framework* comum para as suas implementações. Este *framework* é derivado da pesquisa feita por Bobadilla et al. (2013), e sua utilização neste trabalho se justifica pela sua característica generalista, abrangendo temas que se aplicam a qualquer tipo de SR. Os principais algoritmos de recomendação são explorados em uma subseção própria, em que são destacados fatores como características, vantagens e desvantagens.

2.1. *FRAMEWORK* PARA SISTEMAS DE RECOMENDAÇÃO

Bobadilla et al. (2013) apresentam um *framework* completo a respeito das considerações necessárias a serem feitas durante o processo de implementação de qualquer sistema de recomendação. O primeiro item mencionado pelos autores é a identificação dos tipos de dados disponíveis no banco de dados que será utilizado pelo SR, que podem ser, por exemplo, avaliações, informações de registro dos usuários, características de produtos, relações sociais entre usuários, dentre outros. Após essa identificação, torna-se então possível definir o algoritmo de filtragem que gerará as recomendações a partir dos dados disponíveis, assim como o método de consulta dos dados, podendo ser baseado, por exemplo, no uso direto de todo o banco (*memory-based*) ou no uso de modelos gerados a partir do banco (*model-based*) (BOBADILLA et al., 2013). A escolha do método de consulta é fortemente influenciada pela performance esperada do sistema com base no consumo de tempo e memória (SCHAFER et al., 2007).

Por fim, Bobadilla et al. (2013) também citam a necessidade de definir as técnicas de recomendação (relacionadas ao *machine learning*) empregadas nos algoritmos de filtragem, o

nível desejado de esparsidade e escalabilidade do banco de dados, e o tipo de resultado de recomendação desejado junto com a sua qualidade esperada (os resultados mais comuns são do tipo *Top N*, referente aos n principais resultados, enquanto que a qualidade geralmente é avaliada em termos de novidade, cobertura e precisão). A Figura 1 ilustra este *framework* de forma esquemática.

Figura 1 - Esquematização do *framework* proposto por Bobadilla et al. (2013).

2.2. CLASSIFICAÇÕES PARA BANCOS DE DADOS

Com relação à forma como a literatura trata os bancos de dados utilizados pelos sistemas de recomendação, Wei et al. (2007) resumem os mais variados tipos de dados em duas categorias: dados do usuário e dados de produção, sendo que a primeira ainda pode ser dividida em mais quatro subcategorias (dados demográficos, de avaliação, de padrões comportamentais e de transação). Ainda segundo os autores, os dados demográficos apresentam informações referentes à identidade do usuário, enquanto que os de avaliação podem incluir pontuações atribuídas pelo usuário de forma discreta, contínua, ou através de comentários. Já os dados baseados em padrões comportamentais registram os comportamentos do usuário enquanto ele está navegando em algum *website*, e os dados de transação informam sobre o histórico de compras do usuário (WEI et al., 2007).

A categoria dados de produção, por sua vez, baseia-se nos atributos de produtos utilizados no processo de recomendação (WEI et al., 2007). Segundo Mooney et al. (2000), os principais atributos para livros são título, autores, sinopses, críticas publicadas, comentários

de leitores, autores relacionados e títulos relacionados. Todas essas categorias estão apresentadas na Tabela 1.

Classificação	Exemplos
Dados Demográficos	Nome, idade, gênero, data de nascimento, profissão, *hobbies*
Dados de Avaliação	Pontuação discreta (escala de cinco estrelas), pontuação contínua (barra de satisfação) e comentários (bom, ruim, ótimo, pior)
Dados de Padrões Comportamentais	Tempo na página, número de cliques, *links* acessados, pesquisas realizadas, *download* de conteúdos
Dados de Transação	Data da compra, quantidade comprada, preço, desconto adquirido
Dados de Produção	Título, autor, sinopse, gênero, palavras-chave

Tabela 1 - Taxonomia do banco de dados, adaptado de Wei et al. (2007).

Além da questão taxonômica, Bobadilla et al. (2013) evidenciam a relevância que bancos de dados disponíveis publicamente possuem para a pesquisa e desenvolvimento de sistemas de recomendação, pois facilitam experimentações de técnicas, métodos e algoritmos a fim de validá-las e melhorá-las. Segundos os autores, o banco de dados público relacionado a livros mais referenciado na literatura é do *Book-Crossing*, que contém mais de um milhão de avaliações de quase duzentos e oitenta mil usuários para aproximadamente duzentos e setenta mil itens.

2.3. PROBLEMAS DOS SISTEMAS DE RECOMENDAÇÃO

No que diz respeito aos problemas enfrentados durante o processamento de sistemas de recomendação, o principal denomina-se *cold-start* e está relacionado à impossibilidade de realizar recomendações confiáveis devido ao baixo volume ou completa ausência de dados do usuário durante o início da operação do sistema (BOBADILLA et al., 2013). O *cold-start* pode ser separado em três categorias: nova comunidade, novo ítem e novo usuário

(SCHAFER et al., 2007), sendo o último o mais desafiador para os sistemas de recomendação já em operação (BOBADILLA et al., 2013).

O problema de nova comunidade refere-se à dificuldade de obter dados suficientes para iniciar um sistema de recomendação, o que pode ser mitigado encorajando os usuários a fazerem avaliações de itens por meio de incentivos (SCHAFER et al., 2007). O problema de novos itens, por sua vez, surge quando novos itens são inseridos no banco de dados sem avaliações prévias ou histórico de compras, dificultando a utilização desses itens pelo processo de recomendação e consequentemente deixando-os de fora indefinidamente (SCHAFER et al., 2007). Este caso apresenta menor impacto em SRs onde os itens podem ser descobertos por outros meios e ainda podem ser mitigados através de usuários motivados que são responsáveis por avaliar cada novo item inserido no sistema (BOBADILLA et al., 2013). Já o problema de novos usuários está calcado na impossibilidade de realizar predições através de algoritmos baseados em filtragem colaborativa por conta da inexistência de avaliações ou histórico de compras deste novo usuário (SCHAFER et al., 2007). A estratégia comum utilizada para superar esta situação consiste em agregar outras informações disponíveis do usuário ao processo de recomendação (BOBADILLA et al., 2013).

De maneira geral, no intuito de amenizar os impactos causados pelos diferentes problemas do tipo *cold-start* na qualidade de predição dos SRs, técnicas de clusterização podem ser aplicados nos algoritmos (BOBADILLA et al., 2013). A forma mais comum de clusterização neste caso ocorre somente com itens, mas também há outras abordagens que incluem usuários, denominados de biclusterização (BOBADILLA et al., 2013).

2.4. MÉTODOS DE AVALIAÇÃO PARA SRs

Segundo Herlocker et al. (2004), as avaliações das predições e recomendações se tornaram importantes desde o começo das pesquisas na área de SRs, e estão atreladas a métricas que apuram a qualidade das recomendações. Os autores também ligam a importância dessas avaliações com a facilitação da comparação de diversas soluções que atuam em um mesmo problema, selecionando aquela que apresenta melhores resultados.

As principais métricas utilizadas para recomendação são Precisão e *Recall*, sendo necessário que ambas sejam analisadas em conjunto (BOBADILLA et al., 2013). O primeiro indicador busca saber, de todos os produtos recomendados, quantos o usuário de fato gostou ou comprou (BOBADILLA et al., 2013). Logo, se cinco produtos são recomendados para o

usuário, sendo que desse montante quatro foram efetivamente comprados, então a Precisão será de 80%. O *Recall*, por sua vez, mede quantos produtos comprados pelo usuário foram recomendados (BOBADILLA et al., 2013). Portanto, se um usuário comprou cinco produtos, sendo que a recomendação continha três deles, então o *Recall* será de 60%. Em suma, o objetivo do sistema é otimizar ambas as métricas para que fiquem o mais próximo possível de 100%, não sendo interessante, por exemplo, ter um Recall de 100% em uma situação onde são recomendados todos os produtos disponíveis da loja, pois a Precisão nesse caso terá uma alta probabilidade de possui um valor muito baixo (BOBADILLA et al., 2013).

Bobadilla et al. (2013), entretanto, ressaltam que os *frameworks* de avaliação desenvolvidos a partir dos métodos apresentam deficiências. A primeira delas refere-se à falta de formalização de muitos detalhes da implementação dos métodos, propiciando a geração de diferentes resultados em experimentos semelhantes. Já a segunda deficiência está relacionada à ausência de padronização das medidas utilizadas durante a avaliação da qualidade dos modelos (BOBADILLA et al., 2013).

2.5. PRINCIPAIS ALGORITMOS DE RECOMENDAÇÃO

De acordo com Portugal et al. (2018), os principais tipos de algoritmos de recomendação são os baseados em filtragem de conteúdo, filtragem colaborativa e em abordagens híbridas. No entanto, os dois primeiros algoritmos citados ainda apresentam um papel predominante em diversos tipos de aplicação (LU et al., 2015).

Weng (2008) descreve os algoritmos baseados em filtragem de conteúdo como uma abordagem que recomenda itens com características similares aos de outros itens já consumidos pelo usuário, podendo utilizar dois métodos: baseado em classificador ou em vizinhança. O primeiro método associa os usuários a perfis, e um classificador então decide se um novo item apresentado deve ser recomendado ou não levando em consideração suas características (WENG, 2008). O segundo método, por sua vez, armazena os itens que o usuário já visualizou ou avaliou e os compara com uma rede subjacente que contém outros itens para descobrir quais deles o usuário teria interesse baseado na similaridade de características (WENG, 2008). Mooney e Roy (2000) propõem a implementação de um sistema de recomendação de livros por meio de um algoritmo baseado em filtragem de conteúdo, utilizando técnicas de extração de informações e recursos de *machine-learning* para a categorização de textos. A partir de dados obtidos na internet, as características dos livros e

perfis de usuários são definidas e os melhores livros são recomendados (MOONEY; ROY, 2000).

Com relação às desvantagens desse tipo de algoritmo, Burke (2002) destaca que o problema *cold-start* prejudica a qualidade da recomendação no momento em que não há avaliações suficientes para gerar um classificador confiável. O algoritmo também é limitado pelas características dos itens, acarretando em recomendações com baixa diversidade de gênero uma vez que ele será sempre semelhante ao dos itens avaliados ou visualizados pelo usuário (BURKE, 2002). Por outro lado, podem ser citados como vantagens o caráter adaptativo do algoritmo (a qualidade melhora ao longo do tempo, na medida em que mais dados são coletados) e o fato de ser suficiente para a geração de recomendações apenas dados implícitos (que não requer uma ação direta do usuário), como é o caso das características de um item (BURKE, 2002).

No que se refere aos algoritmos baseados em filtragem colaborativa, a recomendação ocorre a partir de uma quantidade suficiente de avaliações ou histórico de compras de itens feitas por usuários considerados semelhantes à pessoa interessada (SCHAFER et al., 2007). As avaliações também podem ser coletadas de forma implícita, ou seja, sem ação direta dos usuários, como ocorre por exemplo na contagem de vezes que uma música é escutada e na identificação das informações pesquisadas pelo usuário analisado (BOBADILLA et al., 2013). Este tipo de algoritmo possui métodos para consulta de dados que podem ser agrupados em duas classes: baseado em memória (*memory-based*) e baseado em modelo (*model-based*) (BREESE et al., 1998). O método baseado em memória se trata de uma heurística que busca encontrar, em todo o banco de dados, usuários que apresentam similaridades com o usuário ativo (aquele que se quer realizar a recomendação), no intuito de usar suas preferências como base para a predição das preferências do usuário ativo (BREESE et al., 1998). Já o método baseado em modelo usa as informações do banco de dados para gerar um modelo que servirá como referência para as próximas recomendações (BOBADILLA et al., 2013).

Burke (2002) destaca como exemplo de desvantagem deste algoritmo a dificuldade em gerar recomendações quando a quantidade de avaliações ou o histórico de compras é escasso, acarretando em poucas avaliações para os mesmos itens ou poucas vendas cruzadas (quantidade de vezes em que um determinado item foi comprado junto a outros). Ele funciona melhor quando a densidade de avaliações ou compras é alta em um pequeno e estático universo de itens, pois caso a lista de itens mude rapidamente, informações antigas terão

pouca significância para novos usuários, que não poderão ter suas próprias avaliações e compras comparadas ao dos usuários já existentes (BURKE, 2002). Ao mesmo tempo, se o número de itens for muito grande e as avaliações e compras forem esparsas, as probabilidades de sobreposição das avaliações e vendas cruzadas serão pequenas (BURKE, 2002). Por outro lado, são citados como vantagens o fato do algoritmo depender apenas das avaliações ou histórico de compras dos usuários, podendo recomendar itens sem precisar consultar suas características, e a capacidade de fazer recomendações "fora da caixa", ou seja, sem limitação de gênero (BURKE, 2002). O resumo das características dos dois algoritmos pode ser visualizado na Tabela 2.

Algoritmo	Vantagens	Desvantagens
Filtragem de Conteúdo	Qualidade melhora conforme maior agregação de dados; Depende apenas de dados implícitos.	Resultados com baixa diversidade de gênero.
Filtragem Colaborativa	Funciona apenas com avaliações/compras de usuários; Resultados com maior diversidade de gênero.	Ineficiente para bancos com baixa quantidade de avaliações/histórico de compras.

Tabela 2 - Comparação das características de dois algoritmos de recomendação.

3. METODOLOGIA

Nesta seção será discutida a metodologia aplicada no trabalho, começando pela contextualização do ambiente onde ele será realizado, seguido pela caracterização da pesquisa quanto à sua natureza, abordagem e objetivo. Por fim, serão exploradas todas as etapas que envolvem o processo de desenvolvimento de um sistema de recomendação para uma plataforma de vendas online.

3.1. DESCRIÇÃO DO CENÁRIO

O cenário deste trabalho é delimitado na loja virtual criada pela Tag Livros no início de 2018. A empresa em si foi fundada em 2014 na cidade de Porto Alegre, quando sua única fonte de receita era o clube de assinatura voltado para leitores que estavam em busca de novas experiências literárias. O modelo de negócio se baseia no envio mensal de um kit contendo

livro, revista personalizada, marcador de página e um brinde, sendo o conteúdo de cada kit desconhecido pelos assinantes até o momento em que eles o recebem em casa. Dentro desse contexto, a TAG oferece duas opções de assinatura, denominadas de Curadoria e Inéditos, com cada uma apresentando temáticas diferentes para a montagem dos kits.

A loja virtual surgiu, portanto, da necessidade da empresa expandir o seu nicho de atuação e, ao mesmo tempo, vender o estoque de kits não utilizados em períodos anteriores. Qualquer pessoa, incluindo os atuais assinantes da empresa, pode adquirir através do site diferentes produtos relacionados à literatura. São eles kits passados, livros avulsos, itens de decoração, acessórios, brindes, papelaria e vestuário. Os assinantes são inclusive incentivados a adquirir produtos na loja por meio de descontos exclusivos.

A loja virtual já apresenta um sistema de recomendação próprio, porém ele é incipiente e seu *output* é gerado manualmente pelos funcionários responsáveis pelo gerenciamento da plataforma, impossibilitando a flexibilização dos resultados de acordo com o perfil do cliente. Este sistema engloba listas de produtos mais vendidos, com menores preços, e lançamentos (expostas na página inicial do site), além de produtos relacionados a um item específico (expostos na página do item), conforme mostram as Figuras 2 e 3.

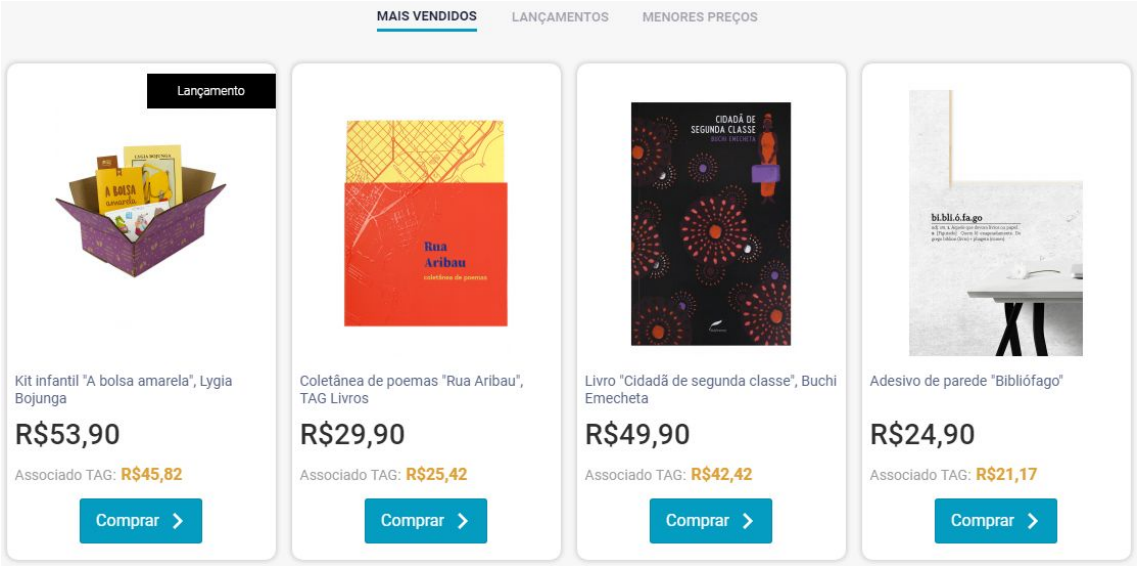

Figura 2 - Recomendações expostas na página inicial da loja virtual.

Figura 3 - Recomendações expostas na página de um item vendido na loja virtual.

3.2. CARACTERIZAÇÃO DO MÉTODO DE PESQUISA

No que se refere às classificações da pesquisa, a mesma pode ser definida como de natureza aplicada, pois prediz uma aplicação dos conhecimentos para encontrar soluções de propósitos práticos. A abordagem do trabalho, por sua vez, é de natureza tanto quantitativa quanto qualitativa, pois, ao mesmo tempo em que a avaliação do SR é fundamentada em métricas numéricas, algumas etapas da sua construção envolvem conhecimentos qualitativos dos gestores da empresa sobre os seus clientes e os produtos que vendem. Por fim, o objetivo de pesquisa pode ser classificado como descritivo, em razão da apresentação e análise dos fatores que influenciam na performance do processo de recomendação.

3.3. MÉTODO DE TRABALHO

O método aplicado neste trabalho é dividido em quatro etapas, abrangendo os seis fatores apresentados no *framework* proposto por Bobadilla et al. (2013). A primeira etapa refere-se à manipulação do banco de dados da Tag Livros, e contém como subatividades a coleta de dados junto à empresa, limpeza dos dados (removendo ruídos e inconsistências), integração dos dados em um formato padrão e, por fim, seleção daqueles que tendem a ser mais relevantes para o processo de recomendação, com base em uma avaliação qualitativa dos atributos (ressaltando que o foco da recomendação é apenas para a categoria de livros). Caso necessário, também será aplicado o processo de *feature engineering*, no qual variáveis

qualitativas são transformadas em uma ou mais variáveis numéricas, possíveis de serem interpretadas pelo algoritmo aplicado.

Após a conclusão desta etapa, torna-se possível a definição dos tipos de algoritmo de recomendação que serão testados, assim como os métodos utilizados na consulta do banco de dados, as técnicas de *machine learning* empregadas nos algoritmos e o tipo de resultado (*output*) que se deseja expor para o usuário final. A escolha dos algoritmos de recomendação depende exclusivamente dos tipos de dados coletados, enquanto que os métodos de consulta e as técnicas serão definidos de acordo com o que é recomendado pela literatura para cenários semelhantes a deste trabalho. O tipo de resultado exposto, por sua vez, será determinado pelo representante da Tag Livros responsável pela gestão da loja virtual.

Em seguida, o sistema de recomendação será desenvolvido em linguagem Python a partir do que foi definido anteriormente, tendo em vista a facilidade para programar e acessar uma ampla biblioteca de códigos (relacionados a SRs) disponível gratuitamente na internet. Finalmente, os sistemas desenvolvidos serão testados utilizando o banco de dados da empresa e os resultados dos testes serão submetidos a métodos de avaliação da qualidade, que servirão como referência para a escolha do melhor sistema.

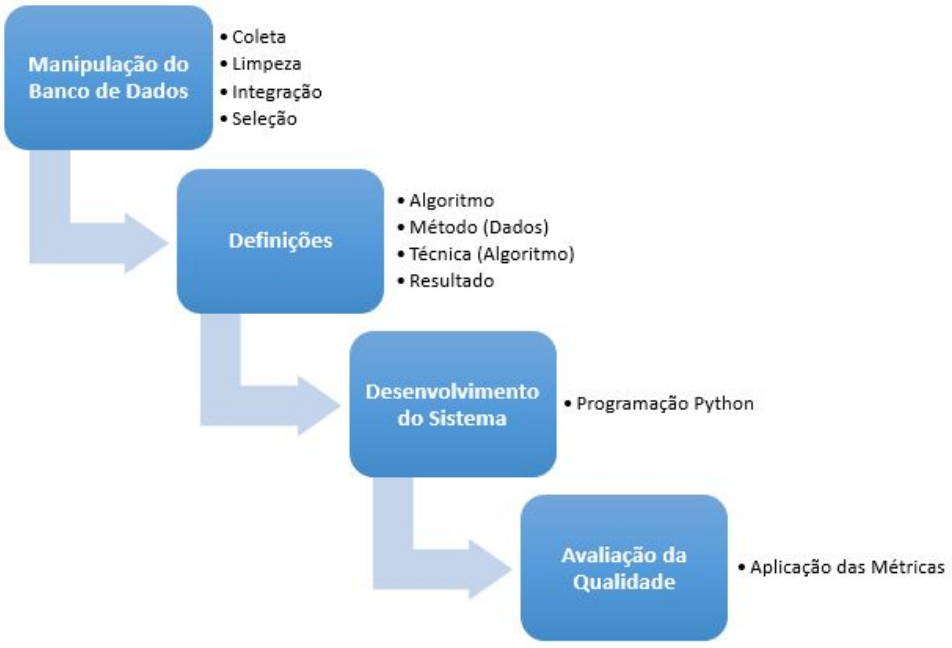

Figura 4 - Passo a passo do método de trabalho.

4. RESULTADOS

Nesta seção, os resultados obtidos por meio da aplicação do método de trabalho serão apresentados na ordem cronológica em que ocorreram as etapas de desenvolvimento do sistema de recomendação sugerido.

4.1. MANIPULAÇÃO DO BANCO DE DADOS

Durante o momento de coleta dos dados, foi disponibilizado pela empresa um banco de dados que continha quatro tabelas. As tabelas e seus respectivos conteúdos estão descritos na Tabela 3.

Tabela	Conteúdo
Pedidos	ID do pedido, data de compra, total da compra (em reais), forma de pagamento e ID do cliente.
Clientes	ID do cliente, grupo (não associado, associado a 1 clube, associado a 2 clubes), estado e data (para não associados, a data se refere à primeira compra no site, enquanto que para associados se refere à data de associação).
Produtos	ID do produto, categoria, subcategoria, nome do produto (para o caso de livros, o atributo continha título da obra e autor), preço de venda para não associados, preço de venda para associados a 1 clube e preço de venda para associados a 2 clubes.
Itens	ID do pedido, ID dos produtos comprados e quantidade de produtos iguais comprados.

Tabela 3 - Banco de dados disponibilizado pela TAG Livros.

Além do banco fornecido pela TAG, também foram acrescentados dois novos atributos na tabela Produtos, com o objetivo de enriquecer os dados e auxiliar em melhores resultados de recomendação. Os novos atributos são gênero e sinopse (ambos para livros), e, a partir desse novo banco de dados, iniciou-se o processo de limpeza. Para tanto, foi realizado a filtragem dos produtos no intuito de deletar todos os eventos que não fossem relacionados a livros. Além disso, as letras maiúsculas foram trocadas por minúsculas, os cedilhas foram trocados por "c" e todos os acentos foram removidos, garantindo, desta forma, que não houvesse problemas durante a leitura dos dados pelos códigos em Python. Em seguida, foi definido por meio de uma avaliação qualitativa que os atributos mais relevantes para o desenvolvimento de SRs são título, autor, gênero, sinopse, ID do cliente e ID do produto. Os dois últimos são oriundos do histórico de compra da loja virtual, com um evento único para cada produto comprado pelo cliente. Os atributos que não constam nesta lista, por sua vez, foram removidos do banco de dados.

Posteriormente, foi realizada a integração dos dados em um formato padrão. Tendo em vista que o sistema de recomendação analisa cada palavra para identificar semelhanças entre os produtos, concatenou-se o nome e sobrenome dos autores em uma só palavra, evitando que obras escritas por autores de mesmo nome tenham seus níveis de semelhança aumentados apenas por conta dessa coincidência. Além disso, foram extraídas das sinopses as principais palavras contidas nos textos, garantindo que os artigos, preposições, conjunções, pronomes, advérbios, numerais e verbos mais utilizados na língua portuguesa fossem ignorados, uma vez que não agregam valor às características dos produtos. Por fim, todos os atributos dos produtos foram organizados em uma só coluna, denominada Vetor, tendo o título como índice, para que cada obra fosse reconhecida como um vetor pelo sistema. Em suma, obteve-se a partir dessas ações duas tabelas: uma relacionada aos produtos (livros) e suas características (dados de produção), e outra relacionada ao histórico de compras do site (dados de transação). As Tabelas 4 e 5 mostram o resultado desta etapa.

	Vetor
Título	
flores partidas	karinslaughter crime ficcao literatura estra...
esposa perfeita	karinslaughter crime ficcao literatura estra...
a intuicionista	colsonwhitehead literatura estrangeira ficcao...
amor insensato	junichirotanizaki literatura estrangeira proje...
uma questao pessoal	kenzaburooe literatura estrangeira autismo re...

Tabela 4 - Tratamento dos dados de produção.

ID Cliente	ID Produto
692	322
6374	266
6374	313
3466	313
3466	215

Tabela 5 - Tratamento dos dados de transação.

4.2. CARACTERÍSTICAS DOS SISTEMAS DE RECOMENDAÇÃO

Após analisar as características das variáveis disponíveis para *inputs* do sistema de recomendação, chegou-se à conclusão que tanto o algoritmo baseado em conteúdo quanto o baseado em filtragem colaborativa poderiam ser desenvolvidos. Para o primeiro tipo, foi utilizado como *input* a tabela contendo os dados de produção, enquanto que o algoritmo baseado em filtragem colaborativa utilizou a tabela contendo os dados de transação.

Tendo em vista o volume moderado de dados de produção e de transação (154 livros e histórico de um semestre de compras), também foi decidido que os métodos de consulta do banco de dados aplicados nos algoritmos baseados em conteúdo e filtragem colaborativa seriam, respectivamente, baseados em vizinhança e memória. Portanto, no primeiro caso, os produtos que o cliente já comprou são comparados com todos os outros produtos existentes, a fim de descobrir quais deles possuem maior similaridade de características. Já no segundo

caso, todo o histórico de compras é utilizado para identificar os produtos que possuem maiores chances de serem compradas pelo cliente.

Finalizando a definição das características dos sistemas, para ambos os tipos de algoritmo decidiu-se aplicar a técnica de *machine-learning* denominada *cosine*, uma vez que se trata de uma ferramenta capaz de calcular a similaridade entre produtos por meio da distância/ângulo entre os seus respectivos vetores, podendo variar entre 1 (ângulo igual a zero graus, representando similaridade total) e 0 (ângulo igual a noventa graus, representando nenhuma similaridade). O resultado da aplicação dessa técnica gera uma matriz que compara os valores de similaridade calculados para cada par de produtos, conforme pode ser observado na Figura 5. Por fim, o tipo de *output* escolhido para exibir os resultados da recomendação aos clientes da loja virtual foi o *Top N*, no qual será exposto uma lista de dez produtos seguindo a ordem dos que possuem maiores possibilidades de compra, conforme pode ser observado na Figura 6.

```
array([[ 1.        ,  0.19862652,  0.081683  , ...,  0.06873217,
         0.18184824,  0.1977887 ],
       [ 0.19862652,  1.        ,  0.08939982, ...,  0.08425255,
         0.13931955,  0.15153203],
       [ 0.081683  ,  0.08939982,  1.        , ...,  0.08661987,
         0.06875239,  0.10545769],
       ...,
       [ 0.06873217,  0.08425255,  0.08661987, ...,  1.        ,
         0.12148858,  0.10164464],
       [ 0.18184824,  0.13931955,  0.06875239, ...,  0.12148858,
         1.        ,  0.14940358],
       [ 0.1977887 ,  0.15153203,  0.10545769, ...,  0.10164464,
         0.14940358,  1.        ]])
```

Figura 5 - Matriz de similaridades oriundo da técnica *cosine*.

ID cliente	Produto	score	rank
692	284	0.15815615654	1
692	337	0.152537743251	2
692	319	0.147287944953	3
692	324	0.140632768472	4
692	346	0.123104314009	5
692	267	0.122875809669	6
692	323	0.120600978533	7
692	336	0.118552943071	8
692	266	0.115030964216	9
692	326	0.113199571768	10
6374	318	0.267524790764	1
6374	310	0.263566923141	2
6374	267	0.253871870041	3
6374	276	0.23514572382	4
6374	312	0.220768117905	5
6374	274	0.22014952898	6
6374	337	0.219813370705	7

Figura 6 - Lista de recomendação *Top N*.

4.3. AVALIAÇÃO DA QUALIDADE

Após concluída a programação dos dois tipos de sistemas de recomendação em Python (baseado em conteúdo e filtragem colaborativa), a última etapa de todo o processo metodológico buscou avaliar a qualidade dos resultados obtidos em cada sistema. Visualizando os valores contidos na Tabela 6, é possível perceber que o sistema de recomendação baseado em filtragem colaborativa apresentou desempenho superior nas duas métricas utilizadas. Em termos de *Recall*, aproximadamente 45% dos produtos comprados por cada cliente estariam na lista de recomendação, ou seja, se um cliente comprou 5 produtos, 2 estariam na lista de recomendação. Já em termos de Precisão, 6,4% dos produtos contidos na lista de recomendação seriam comprados pelo cliente. Portanto, ao recomendar 10 produtos, o cliente estará próximo de comprar ao menos 1 produto.

Sistema	*Recall*	Precisão
Filtragem Colaborativa	45,3%	6,4%
Baseado em Conteúdo	13,9%	4%

Tabela 6 - Resultados das métricas de qualidade para recomendação.

Vale ressaltar a comparação entre os resultados apresentados por ambos os sistemas e o valor mínimo estatístico, que representa nesse caso a probabilidade de se recomendar

produtos ao acaso e obter a mesma precisão. Considerando que o sistema baseado em filtragem colaborativa garante uma precisão de um produto comprado a cada 10 recomendados, o valor mínimo estatístico fica aproximadamente 0,65%, equivalente à escolha de um livro dentre uma amostra de 154.

5. CONCLUSÕES

Esse estudo buscou desenvolver um sistema de recomendação adaptado ao contexto da loja virtual da TAG Livros, focando apenas em obras literárias, com o objetivo de melhorar a experiência de consumo dos seus clientes e aumentar o volume de vendas nessa plataforma. Portanto, uma vez reconhecidos os resultados de qualidade de cada opção oferecida, cabe aos representantes da empresa decidir se será vantajoso ou não implementar um dos sistemas no site. No caso de uma decisão favorável, sugere-se ainda que haja um esforço por parte da empresa em manter o banco de dados atualizado, evitando que a qualidade dos resultados seja prejudicada.

Com relação aos resultados expostos na seção anterior, nota-se que, com os dados disponibilizados pela TAG Livros, é possível desenvolver os dois principais tipos de sistemas de recomendação (baseado em filtragem de conteúdo e em filtragem colaborativa), visto que há tanto dados de transação quanto de produção. No entanto, o banco original de dados de produção carece de atributos relevantes, fazendo-se necessário buscar em outras fontes dados que complementem o conteúdo dos produtos, como foi o caso do gênero e da sinopse. Portanto, se considerarmos apenas o banco fornecido pela empresa, pode-se concluir que os dados de transação são mais completos e confiáveis para serem usados como *inputs* do sistema, além de serem mais fáceis de serem coletados.

Quanto à avaliação da qualidade dos dois sistemas desenvolvidos, percebe-se que o sistema baseado em filtragem colaborativa apresenta desempenho superior nas duas métricas utilizadas, com, aproximadamente, um produto comprado em cada dez recomendados, ao mesmo tempo em que quase metade dos produtos comprados por cada cliente consta na lista de recomendação. Além disso, a comparação desses resultados com o valor mínimo estatístico evidencia a vantagem em utilizar o sistema de recomendação, uma vez que a probabilidade de se alcançar o mesmo nível de precisão de forma aleatória é de apenas 0,65%. Levando em consideração essas informações, em conjunto com as constatações feitas previamente a

respeito do banco de dados, sugere-se que a TAG Livros implemente em sua loja virtual o sistema de recomendação baseado em filtragem colaborativa.

Por fim, no intuito de aprimorar os índices de qualidade do sistema a ser implementado, sugere-se que em trabalhos futuros seja utilizado um banco de dados mais robusto, contendo mais dados de produção dos livros, assim como uma maior série histórica de dados de transação. Posteriormente, as métricas de qualidade devem ser reaplicadas para cada tipo de SR desenvolvido, de forma a avaliar os ganhos obtidos.

REFERÊNCIAS

ADOMAVICIUS, Gediminas; TUZHILIN, Alexander. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. **IEEE Transactions on Knowledge & Data Engineering**, n. 6, p. 734-749, 2005.

ALJUKHADAR, Muhammad; SENECAL, Sylvain; DAOUST, Charles-Etienne. Using recommendation agents to cope with information overload. **International Journal of Electronic Commerce**, v. 17, n. 2, p. 41-70, 2012.

BOBADILLA, Jesús et al. Recommender systems survey. **Knowledge-based systems**, v. 46, p. 109-132, 2013.

BREESE, John S.; HECKERMAN, David; KADIE, Carl. Empirical analysis of predictive algorithms for collaborative filtering. In: **Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence**. Morgan Kaufmann Publishers Inc., 1998. p. 43-52.

BURKE, Robin. Hybrid recommender systems: Survey and experiments. **User modeling and user-adapted interaction**, v. 12, n. 4, p. 331-370, 2002.

HERLOCKER, Jonathan L. et al. Evaluating collaborative filtering recommender systems. **ACM Transactions on Information Systems (TOIS)**, v. 22, n. 1, p. 5-53, 2004.

LINDEN, Greg; SMITH, Brent; YORK, Jeremy. Amazon. com recommendations: Item-to-item collaborative filtering. **IEEE Internet computing**, n. 1, p. 76-80, 2003.

LU, Jie et al. Recommender system application developments: a survey. **Decision Support Systems**, v. 74, p. 12-32, 2015.

MOONEY, Raymond J.; ROY, Loriene. Content-based book recommending using learning for text categorization. In: **Proceedings of the fifth ACM conference on Digital libraries**. ACM, 2000. p. 195-204.

PORTUGAL, Ivens; ALENCAR, Paulo; COWAN, Donald. The use of machine learning algorithms in recommender systems: a systematic review. **Expert Systems with Applications**, 2017.

RICCI, Francesco.; ROKACH, Lior.; SHAPIRA, Bracha. Introduction to Recommender Systems Handbook, Recommender Systems Handbook. 2011.

SCHAFER, J. Ben et al. Collaborative filtering recommender systems. In: **The adaptive web**. Springer, Berlin, Heidelberg, 2007. p. 291-324.

SHAH, Kunal et al. Recommender systems: An overview of different approaches to recommendations. In: **Innovations in Information, Embedded and Communication Systems (ICIIECS), 2017 International Conference on**. IEEE, 2017. p. 1-4.

TAG LIVROS. O Surgimento do Clube. Disponível em: <https://www.taglivros.com/blog/surgimento-do-clube-tag>. Acesso em: 9 nov. 2018.

WEI, Kangning; HUANG, Jinghua; FU, Shaohong. A survey of e-commerce recommender systems. In: **Service systems and service management, 2007 international conference on**. IEEE, 2007. p. 1-5.

WENG, Li-Tung. **Information enrichment for quality recommender systems**. 2008. Tese de Doutorado. Queensland University of Technology.