PROJETO TCC - BCC	ANO/SEMESTRE:	2020.2

USO DE REDES COMPLEXAS PARA ANÁLISE DE CONSUMO EM ECOMMERCE

Diogo Warmeling

Aurélio Faustino Hoppe – orientador

1 INTRODUÇÃO

Com o crescimento dos acessos à internet e a disseminação de dispositivos móveis, os comércios eletrônicos (e-commerce) vêm crescendo de forma exponencial nos últimos anos e, com a pandemia do COVID-19, o crescimento foi ainda maior, tornando-se o método mais comum de compra para diversos setores da economia brasileira (TERRA, 2020). Segundo os dados da 42ª edição do Webshoppers, um estudo sobre e-commerces do Brasil elaborado semestralmente pela Ebit e Elo, o crescimento nas vendas do primeiro semestre de 2020 foi de 47%, caracterizando-se como dobro dos registrados nos últimos anos. Isso, segundo Lunardi (2018), aconteceu devido a entrada de novos adeptos a esse meio de compra, resultando em um crescimento de 40%, chegando a 41 milhões de pessoas.

Brito (2020) destaca que esse crescimento não ficou restrito somente aos *marketplaces*, mas também atingiu muitos comércios eletrônicos de pequeno e médio porte, dos quais, a maior parte não aproveita o grande potencial que a análise dos dados de acessos de seus clientes fornece. Tais análises possibilitam entender o comportamento dos usuários para mantê-los mais tempo em seus *sites*, e conseguir recomendar o produto desejado antes da desistência da compra, assim como auxiliar os clientes a encontrarem as melhores opções disponíveis, aumentando a taxa de efetivação da compra por parte dos usuários. No entanto, a grande maioria dos *sites* apenas disponibilizam listas de produtos semelhantes e associados ao acesso do usuário, desconsiderando o seu perfil, que pode estar seguindo um padrão de acessos.

Segundo Santos (2019) existem diversas pesquisas sobre recomendações em plataformas de comércio eletrônico, que em sua maioria fazem cálculo de similaridade entre entidades (compradores e produtos) para recomendar novos produtos a um potencial comprador. Dentre elas, encontram-se uma grande quantidade de técnicas e algoritmos disponíveis pensadas para análise de dados tais como sistemas de recomendações, algoritmos de aprendizado de máquina e redes complexas, que permitem encontrar relações e agrupar itens, analisar o fluxo de navegação ou o comportamento dos usuários. Tais estruturas são baseadas em grafos, sendo composto por um conjunto de nós, ligados por arestas e que possui uma estrutura topológica não trivial. Elas podem representar diversos aspectos do mundo, como modelar pessoas e suas conexões nas redes sociais, páginas da *web* e seu fluxo de acesso, entre outros inúmeras possibilidades (BARABÁSI, 2003).

Zhong *et al.* (2014) afirmam que um sistema de recomendação baseado em grafos é mais eficiente do que um modelo baseado em similaridade entre nós (perfis similares de usuários, por exemplo). Diante desta afirmação, este trabalho tem como objetivo gerar redes complexas utilizando grafos, baseadas no fluxo de navegação dos usuários em plataformas de comércio eletrônico, buscando entender e contribuir com o aumento das efetivações de compra.

1.1 OBJETIVOS

Este trabalho tem como objetivo gerar redes complexas utilizando grafos, baseadas no fluxo de navegação dos usuários em plataformas de comércios eletrônicos, buscando entender e contribuir com o aumento das efetivações de compra.

Os objetivos específicos são:

a) utilizar redes complexas para gerar agrupamentos de produtos ou de comportamentos dos usuários para fazer recomendações;
b) analisar as estruturas das redes complexas, calculando métricas e identificando correlações entre a navegação e as compras;
c) identificar os maiores pontos de saídas do e-commerce a partir das redes geradas;
d) integrar a ferramenta a API Suaview.

2 TRABALHOS CORRELATOS

Neste capítulo são apresentados três trabalhos que apresentam semelhanças com o trabalho proposto. A seção 2.1 discorre sobre um sistema de recomendação baseada em agrupamentos (CARVALHO, 2016). A seção 2.2 apresenta um algoritmo para recomendação de relacionamentos em redes sociais (SILVA, 2010). Por fim, a seção 2.3 aborda a construção de um *webcrawler* e a identificação de grupos de produtos comprados em conjunto (SANTOS, 2019).

2.1 RECOMENDAÇÃO BASEADA EM MODULARIDADE

Carvalho (2016) desenvolveu um sistema para gerar recomendações de produtos com base em agrupamentos, utilizando algoritmos de redes complexas, que precisava ser performático e ter baixa taxa de erro. Para isso, a autora utilizou o LensKit, *framework* que dá suporte a sistemas de recomendação.

Segundo Carvalho (2016) foram utilizadas as bases Movie Lens 100K, MovieTweetings, Movie Lens 1M, BookCrossing e Jester. As três primeiras bases associam usuários a filmes. Já a quarta base associa os usuários a livros e a quinta base faz a associação dos usuários com piadas, todas com as qualificações dadas pelos usuários aos itens. A partir delas, foram executados os algoritmos Louvain com as modularidades de Newman e Girvan, Agrupamento com Movimento de Arestas (AMA) e Agrupamento com Movimento de Vértices (AMV) com a modularidade Suzuki-Wakita.

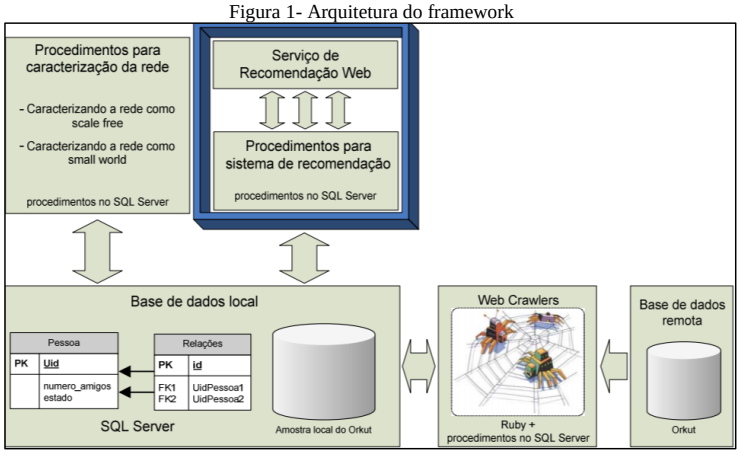
Com base nos experimentos, Carvalho (2016) optou pela a utilização do método Agrupamento com Movimento de Arestas (AMA), pois apresentou os melhores tempos de treinamento e predição. Na análise, a autora gerou 6 grupos de itens e 6 grupos de usuários a partir da base Movie Lens 100K. Utilizando o algoritmo AMV, Carvalho(2016) encontrou alguns padrões, como, um grupo com uma quantidade grande de itens, porém poucas arestas, significando que são filmes pouco avaliados; mulheres assistiram mais os filmes dos gêneros infantil e de animação, além de assistirem poucos filmes de ação. Carvalho (2016) também notou que o grupo que contém em sua maioria usuários entre os 20 e 40 anos assistem mais filmes de suspense e romance. Já o grupo que contém em sua maioria usuários acima dos 60 anos assistem mais filmes de ação e comédia.

Carvalho (2016) conclui que redes complexas e sistemas de recomendações podem gerar bons resultados. Já a hipótese de que quanto maior a métrica de modularidade, menor a taxa de erro, se provou verdadeiro em sua maioria, a não ser na utilização do algoritmo AMV com a modularidade Suzuki-Wakita. Como trabalhos futuros, Carvalho (2016) aponta a necessidade da criação de um algoritmo que encontre a quantidade ideal de grupos, para que não haja a necessidade da passagem de parâmetros, assim como realize a reorganização da rede nas inserções de vértices e faça recomendações quando não for encontrada ligações.

2.2 RECOMENDAÇÃO DE RELACIONAMENTOS EM REDES SOCIAIS BASEADA EM GRAFOS

Silva (2010) desenvolveu um sistema de recomendação de relacionamentos em redes sociais, utilizando grafos. O autor propôs um método híbrido utilizando algoritmo genético e dados topológicos, utilizando a rede social Orkut como base. Para o desenvolvimento do método foi construído um *framework* com 4 módulos. *webcrawlers*, responsáveis pela captura das informações; Base de dados local, responsável pelo armazenamento das informações; Procedimentos para caracterização da rede, responsável pela geração dos grafos; Serviço de Recomendação Web, responsável por buscar dos grafos o grupo de itens para recomendação. Esses módulos são representados pela Figura 1~~Figura 1~~.

Figura 1- Arquitetura do framework



Fonte: Silva (2010).

A partir da Figura 1, pode-se perceber que existe um módulo referente a base de dados, utilizando SQL Server, contendo duas tabelas, uma tabela Pessoa representando os nós e a tabela Relações representado os vértices do grafo. Já o módulo w*ebcrawlers* é responsável pela busca dos dados da rede social Orkut, salvando os dados

encontrados na base de dados local. O *crawler* utiliza um usuário como nó raiz, realizando uma busca em largura para acessar os amigos e fazer os mapeamentos. Foram utilizadas duas bases, uma com usuários brasileiros e outra de indianos. O módulo de caracterização da rede foi desenvolvido em T-SQL a partir de cálculos e constatações das redes *Scale Free* e *Small World*. Por fim, o módulo de serviço de recomendação, também escrito em T-SQL, executa os cálculos de índices, a ponderação e a partir disso, gera as recomendações finais. Esse processo passa por duas etapas, a filtragem e ordenação. A filtragem é responsável por limitar a quantidade de vértices para a ordenação. Neste sentido foi utilizado o conceito de *clustering coefficient*, alcançando os vértices com dois saltos. Já a ordenação utiliza um mecanismo de autoajuste para regular os pesos. Nele, são utilizados 3 índices, o primeiro é composto pela quantidade de amigos em comum. O segundo se refere a densidade do conjunto formado no primeiro índice. No terceiro é a densidade do conjunto formado pelos adjacentes dos itens. A partir desses índices, gera-se a ponderação e a recomendação.

Silva (2010) realizou dois experimentos. No primeiro foi executado a recomendação para dois conjuntos de dados, executando várias vezes ajustando os pesos manualmente com variações pequenas, até que se atingisse o menor valor para a função de otimização. Os melhores resultados obtidos foram um posicionamento médio de 513 e 272 para a sub-rede brasileira e indiana respectivamente, dentro do conjunto de candidatos. Para avaliar a eficiência, o autor utilizou seu próprio usuário, chegando em uma taxa de 75% de acertos nas 20 primeiras recomendações e 50% nas 100 primeiras recomendações.

No segundo experimento Silva (2010) realizou a adição do algoritmo genético e uma validação cruzada com a estratégia *Friend of friend* (FOF). Os dados utilizados vieram da rede social Oro-Aro do C.E.S.A.R, dos quais foram enviadas 10 recomendações para um grupo de 70 usuários, onde 31 realizaram os testes. As recomendações tiveram uma porcentagem de 77.69% de aceitação, 5% superior a estratégia FOF.

Silva (2010) conclui que no primeiro experimento os resultados obtidos foram satisfatórios, porém o método de validação é falho por não possibilitar a reprodução em um ambiente externo. Ele também escreve que a caracterização da rede foi essencial para a elaboração do mecanismo de recomendação. Como melhorias Silva (2010) sugere a adição de outros índices na ordenação para melhorar a ponderação e a implementação de um módulo de algoritmo genético para aprimorar a otimização assim como a realização de testes com outras localizações ou a utilização de alguma API para a extração de dados no lugar do *webcrawler*, pois isso limita apenas ao uso de redes sociais, sendo que qualquer alteração no site pode ocasionar falha de extração dos dados.

2.3 ANÁLISE DA REDE DE PRODUTOS COMPRADOS EM CONJUNTO NO COMÉRCIO ELETRÔNICO

Santos (2019)~~,~~ tinha como objetivos~~,~~ a construção de um *webcrawler* para obter uma base de dados, assim como a construção de uma rede de produtos comprados em conjunto para analisar e identificar correlações entre produtos e validar hipóteses. Algumas dessas hipóteses são, produtos melhores avaliados serem *hubs* de rede e validar se somente as métricas de rede são o suficiente para o treinamento do modelo.

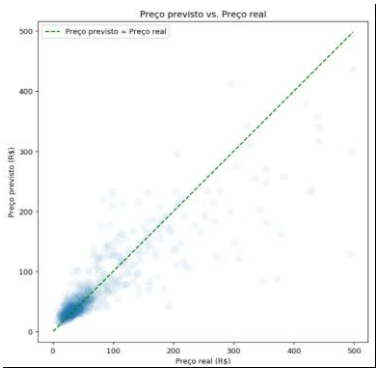
Para montar a base de dados foi criado um *webcrawler*, utilizando o *selenium*, para capturar os produtos comprados em conjunto do e-commerce da Amazon. A busca é iniciada com um produto selecionado manualmente, e posteriormente é executado de forma automática como uma busca em largura pelos outros produtos comprados em conjunto. Ainda segundo Santos (2019) o *crawler* foi desenvolvido em PHP com a execução em `Bash` para fazer execuções em paralelo. Após a extração, os dados são enviados separados em dois arquivos `csv` para uma aplicação em python que utiliza a biblioteca `NetworkX` para processar os dados e gerar as redes, que são salvas em um *dataframe* do pandas, para gerar as previsões de estimativa de preço e as previsões de ligação entre nós. Para criar as estimativas de preços foram utilizados o *Regressor Random Forest* e o coeficiente Gini na montagem das árvores com maiores relevâncias. O resultado obtido pode ser visualizado na Figura 2, onde pode-se ver uma linha tracejada verde que representa o preço previsto do produto sendo igual ao preço real do produto analisado. Já os pontos azuis representam as predições, os que se localizam acima possuem preço previsto superior ao real, e os que se localizam abaixo tem preço previsto inferior ao real.

Para a previsão de ligação entre nós foram utilizados todos os atributos no treinamento e a coluna `existe_ligacao_com_o_1` como alvo e a biblioteca `Scikit_Learn` para o treinamento. Por se tratar de classificações binárias bastante desbalanceadas foi utilizada a matriz de confusão para fazer a análise dos resultados, analisando os falsos positivos e positivos negativos. Dos 768 itens processados, Santos (2019) obteve 634 identificados corretamente, 1 positivo negativo e 133 falsos positivos. Para uma segunda execução foi considerado a utilização da *Receiver Operating Characteristic* Curve (ROC) para identificar o ponto com maior precisão, um ponto onde não seja sensível demais onde terá muitos falsos positivos e nem específico demais, onde obteria muito positivos falsos, algo como regular a sensibilidade de um detector de incêndio, não sendo sensível demais a ponto de disparar a todo momento, tendo muitos casos de falso positivo e nem específico demais, onde pode ter positivos falsos. Com a curva ROC ele obteve um limiar de probabilidade de 0,66, onde valores superiores

Comentado [MH1]: Inseriu uma nova explicação, mas a redação ainda está confusa. Muito "onde"...
A ideia consegue ser captada, mas depende de um esforço de leitura.

eram considerados positivos e inferiores negativos. Após a aplicação do resultado da curva ROC Santos (2019) obteve resultados consideravelmente superiores, reduzindo o número de falsos positivos de 133 para 93.

Figura 2 - Gráfico demonstrando o resultado da estimativa de preço



Fonte: Santos (2019)

Santos (2019) conclui que *hubs* de rede nem sempre serão os mais bem avaliados, porém confirmou a correlação deles serem os mais bem posicionados nas listas de mais vendidos. Santos (2019) também confirmou a hipótese de que a utilização exclusiva dos atributos de métrica foram suficientes para obter melhores resultados em comparação com a *baseline* do modelo de previsão de ligações entre nós, ao contrário dos resultados obtidos na precificação dos produtos.

Santos (2019) sugere a análise da predição de relação entre todos os nós da rede, além de revisar os atributos utilizados no treinamento para poder otimizar o processo. Outro ponto é a integração com APIs de comércios eletrônicos, para que não seja necessária a extração das informações com o *webcrawler*.

3 PROPOSTA DA FERRAMENTA

Neste capítulo são definidas as justificativas para a elaboração deste trabalho, assim como os requisitos funcionais, não funcionais e a metodologia que será aplicada no desenvolvimento.

3.1 JUSTIFICATIVA

No Quadro 1 é apresentado um comparativo entre os trabalhos correlatos. As linhas representam as características e as colunas os trabalhos.

Quadro 1 - Comparativos entre os trabalhos correlatos

Correlatos / Características	Carvalho (2016)	Silva (2010)	Santos (2019)
Obtenção dos dados	API	Webrawler e API	Webcrawler
Método de processamento	LensKit	T-SQL	Scikit Learn
Tipo de item analisado	Avaliações de filmes	Amizades em redes sociais	Produtos comprados em conjunto
Usa grafo bipartido	Sim	Não	Não
Principal algoritmo / técnica	Agrupamento com Movimento de Vértices (AMV)	Algoritmo próprio	Random Forest

Fonte: elaborado pelo autor.

A partir do Quadro 1, observa-se duas estratégias de obtenção de dados, enquanto Carvalho (2016) utilizou uma API que forneciam as informações necessárias para a montagem das redes, Santos (2019) em seu trabalho criou um *webcrawler* que navega pela loja da Amazon e captura os dados para sua base. Já Silva (2010) utiliza ambas as técnicas, criando um *webcrawler* que captura informações da rede social Orkut para o primeiro experimento e utiliza uma API da C.E.S.A.R para o seu segundo experimento.

Carvalho (2016) utilizou grafos bipartidos, onde definiu nós do tipo usuário e nós do tipo filme para a sua análise, enquanto Silva (2010) e Santos (2019) não fizeram uso de tal particularidade dos grafos, já que no projeto

de Silva (2019) somente foram mapeados os usuários e no projeto de Santos (2019) foram mapeados os itens. Quanto aos métodos de processamentos, pode-se observar que dois trabalhos utilizam as bibliotecas da linguagem Python para as predições e recomendações, Carvalho (2016) com o LensKit e Santos (2019) com o Scikit Learn, enquanto Silva (2010) utilizou a linguagem T-SQL. Olhando para os principais algoritmos e técnicas, percebe-se que Carvalho (2016) utiliza o algoritmo AMV para realizar o agrupamento, Santos (2019) aplicou o Random Forest para a criação das estimativas de preços e Silva (2010) desenvolveu seu próprio algoritmo utilizando uma série de técnicas e conceitos para as etapas de classificação, filtragem e ponderação.

Dessa forma, conclui-se que dentro dos trabalhos pesquisados não foi encontrado algum que tenha como objetivo gerar recomendações e análises com base nos acessos dos usuários nos sites ou plataformas de comércio eletrônico. Percebe-se também que nenhum trabalho teve em sua proposta a ideia de realizar testes e validações práticas com o usuário final. Sendo assim, o trabalho proposto se difere pois pretende realizar uma análise alternativa, através do uso dos conceitos de redes complexas, pensada no usuário e sua navegação, tendo como objetivo manter seu interesse por um período maior de tempo, e que o conduza com mais naturalidade ao item desejado. Além disso, a ferramenta a ser desenvolvida será integrada e testada através de uma solução de e-commerce já implantada no mercado. Contudo, a partir deste trabalho, espera-se que os usuários encontrem os itens de interesse mais rapidamente, impulsionando indiretamente a efetivação das vendas.

Comentado [MH2]: Qual? Aqui deveria citar a Suaview, pois senão não há relação entre seu trabalho e essa API em nenhum ponto. Veja que o objetivo específico cita Suaview, mas não se sabe o que ela é. E aqui diz que vai integrar com uma solução de mercado e não se sabe qual. Faltou essa relação direta.

Comentado [MH3]: 'Contudo' é adversativo (porém, todavia, entretanto), ou seja, leva a uma ideia oposta.

3.2 REQUISITOS PRINCIPAIS DO PROBLEMA A SER TRABALHADO

A ferramenta a ser desenvolvida deverá:

a) identificar as páginas que são pontos de saídas (nós folhas) (Requisito Funcional – RF);
b) prever as prováveis próximas páginas a serem acessadas pelo usuário (RF);
c) estabelecer grupos e perfis de usuários a partir das páginas acessadas (RF);
d) gerar recomendações com bases nos grupos de produtos e usuários gerados (RF);
e) utilizar medidas de centralidade, complexidade, distâncias, estrutura de comunidades e distribuição de grau para identificar e estabelecer correlações entre usuários, acessos, saídas e conversões de compras (RF);
f) disponibilizar uma interface gráfica que possibilite ao usuário/gestor visualizar estatísticas de acessos e correlações identificadas (RF);
g) ser implementado na linguagem de programação Python (Requisito Não Funcional – RNF);
h) ser modelada seguindo os princípios de redes complexas (RNF);
i) ser integrado com a API Suaview (RNF).

3.3 METODOLOGIA

O trabalho será desenvolvido observando as seguintes etapas:

a) levantamento bibliográfico: pesquisar e estudar sobre redes complexas, sistemas de recomendação e trabalhos correlatos;
b) levantamento dos requisitos: baseando-se nas informações da etapa anterior, reavaliar os requisitos propostos para a aplicação;
c) integração com API Suaview: implementar a estrutura que irá se comunicar com a API Suaview para a obtenção dos dados;
d) definição das ferramentas para modelagem e armazenamento das redes: pesquisar e escolher as ferramentas mais apropriadas para a modelagem e armazenamento das redes;
e) definição das técnicas de análise de redes: pesquisar e (re)definir as técnicas/algoritmos que serão utilizadas para analisar redes complexas;
f) modelagem da rede: a partir do item (d) modelar a estrutura da rede de forma a facilitar a análise das correlações de acordo com o problema proposto;
g) implementação da rede: implementar as funções de análise da rede considerando os itens (d), (e) e (f), tendo a linguagem Python como base;
h) testes: avaliar a performance assim como validar a aderência e eficiência da rede em relação aos resultados alcançados e os que realmente acontecerão no e-commerce em termos de saída ou efetivação de compras e recomendações de produtos.

As etapas serão realizadas nos períodos relacionados no Quadro 2.

Quadro 2-Cronograma de atividades a serem realizadas

	2021									
	fev.		mar.		abr.		maio		jun.	
etapas / quinzenas	1	2	1	2	1	2	1	2	1	2
levantamento bibliográfico										
levantamento dos requisitos										
integração com API Suaview										
definição das ferramentas para modelagem e armazenamento das redes										
definição das técnicas de análise de redes										
modelagem da rede										
implementação da rede										
testes										

Fonte: elaborado pelo autor.

4 REVISÃO BIBLIOGRÁFICA

Este capítulo tem como objetivo explorar os principais assuntos que fundamentarão o estudo a ser realizado. A seção 4.1 aborda sistemas de recomendação. Já a seção 4.2 discorre sobre redes complexas.

4.1 SISTEMAS DE RECOMENDAÇÃO

De acordo com Andrade (2017), os sistemas de recomendação são um conjunto de algoritmos que utilizam técnicas de aprendizado de máquina e recuperação de informação. Motta *et al.* (2012) destacam que os sistemas de recomendação têm como princípio a ideia, de que, o que é relevante para mim, também será relevante para quem possuir um perfil similar ao meu. Ainda segundo os autores, parte das técnicas de recomendação fazem uso das avaliações feitas pelos usuários, para realizar o entendimento e classificação do perfil.

Motta *et al.* (2012) citam a navegação pelas páginas web e avaliações dos conteúdos como um bom exemplo de dados de entrada, pois possibilitam entender os gostos dos usuários. Motta *et al.* (2012) também apontam que as saídas dos sistemas de recomendação podem ser apresentadas em forma de lista estatística, sugestões, avaliações ou resenhas, dependendo do objetivo de cada recomendação.

Cazella *et al.* (2010) citam algumas técnicas vinculadas aos sistemas de recomendação, que surgiram com o propósito de identificar padrões de comportamento dos usuários. Segundo os autores, tais técnicas definem o funcionamento dos sistemas de recomendação, sendo elas:

a) filtragem de informação: a ideia por trás dessa técnica é diminuir a quantidade de informação transitada. Ela utiliza diversos processos para filtrar as informações e entregá-las somente para quem realmente precisa;
b) filtragem baseada em conteúdo: é uma abordagem que tem foco nas propriedades dos itens, onde a similaridade será determinada a partir das semelhanças entre as propriedades dos itens;
c) filtragem colaborativa: tem foco na relação entre os usuários e os itens. Com a similaridade sendo determinada pelos grupos de usuários que atuam (visualizam, compram ou avaliam) sobre os mesmos itens;
d) filtragem híbrida: a filtragem híbrida tem como objetivo unir os pontos fortes das filtragens baseadas em conteúdo e colaborativa, para gerar um sistema que atenda melhor as necessidades dos usuários;
e) filtragem baseada em outros contextos: essa técnica tende a se aproximar mais das recomendações feitas pelos humanos. ~~Utilizando~~ Ela usa não somente informações normalmente utilizadas, como, preferências do usuário, localidade, idade, entre outras, mas também informações mais complexas, como, traços da personalidade, emoção e habilidades sociais;
f) descoberta de conhecimento em base de dados: essa técnica é aplicada com ferramentas de mineração de dados, que irão gerar relações de associação, classificação ou agrupamento dos itens, usuários, ou itens e usuários;
g) mineração de textos: técnica inspirada na mineração de dados, tem como objetivo extrair informações de texto em linguagem natural. É amplamente utilizada com a filtragem baseada em conteúdo, sendo muito importante na análise do conteúdo que descreve o item a ser recomendado. Essa técnica requer um pré-processamento que limpe o texto a ser analisado, realizando uma redução dos termos para um radical comum, fazendo a remoção de palavras pouco significativas (*stopwords*), como conjunções e focando na manutenção dos substantivos que normalmente são as palavras mais representativas do texto.

Cazella *et al.* (2010) ressaltam que cada técnica tem seu propósito e uso, mas duas ou mais técnicas podem ser utilizadas em conjunto. Os autores citam um sistema de recomendações de cursos que faz uso da técnica de

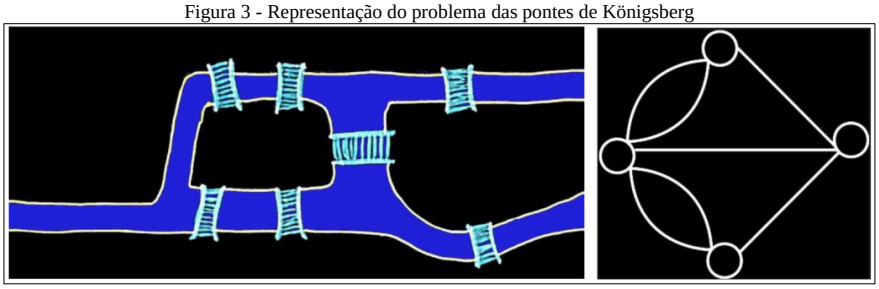
filtragem baseada em outros contextos, no qual é utilizada uma lista de características objetivas, buscadas a partir de uma base de dados, e emocionais respondidas pelo usuário de forma opcional. Neste caso, eles constataram que para os usuários respondentes dos questionários de características emocionais, as recomendações foram mais precisas. Além disso, Cazzela *et al*. (2010) descrevem o uso de técnicas em conjunto, fazendo referência a um sistema de recomendações de páginas web, no qual é utilizado a técnica de filtragem baseada em conteúdo para gerar os perfis dos usuários e, posteriormente compará-los para identificar usuários similares e através da filtragem colaborativa, gerar recomendações. Outros exemplos de sistemas de recomendação são os comércios eletrônicos da Amazon e Ebay. De acordo com Cazella *et al.* (2010), o comércio eletrônico da Amazon faz uso de diversas técnicas, como a mineração de textos dos comentários dos usuários, uso de filtragem colaborativa para recomendar produtos, entre outras técnicas. No Ebay, os autores citam o uso da filtragem de informação, no qual os clientes podem indicar itens que tenham interesse, para que o *site* possa enviar as recomendações em uma periodicidade definida.

4.2 REDES COMPLEXAS

Segundo Carvalho (2012), redes são conjuntos de itens que se conectam entre si, com os elementos da rede sendo vértices e as conexões entre os elementos sendo arestas. Ainda de acordo com o autor, pode-se observar uso das redes em diversas situações, desde o nível subatômico até as estruturas sociais mais complexas.

O estudo das redes teve seu início no século XVIII com Leonhard Euler, onde ele modelou o problema das pontes de Königsberg como um grafo, conforme pode ser visto na Figura 3~~Figura 3~~. Euler modelou os vértices como pedaços de terra e as pontes que os conectavam como arestas. Com essa modelagem, Euler conseguiu provar que não existia um caminho que possibilitasse transitar por todas as sete pontes sem repetir nenhuma (NAHAL, 2014).

Figura 3 - Representação do problema das pontes de Königsberg



Fonte: adaptado de Silva (2020).

Pinto (2018) descreve que um grafo pode ser categorizado como direcional, no qual as arestas têm um sentido, com um vértice de origem e um vértice de destino, ou não direcional, onde o sentido da aresta não importa. Pinto (2018) também explica que arestas podem possuir intensidade, o que significa que cada aresta terá um peso. Esses pesos podem ser utilizados para calcular o melhor caminho entre dois vértices. Carvalho (2012) expõe que a característica principal de um vértice~~,~~ é o seu grau, que representa o número de ligações que o vértice possui. C~~,~~ ~~e~~om o grau dos vértices mensurado, é possível determinar o grau médio da rede, que representa a média de conexões que cada elemento possui. Com os graus dos vértices e grau médio da rede determinados é possível identificar os nós *Hub*. Segundo Nahal (2014), um nó é definido como *Hub* quando ele possui um grau muito superior ao grau médio, sendo determinado como um vértice muito importante pela capacidade de distribuição de informações para um número maior de vértices.

A partir das definições de grafos surgiram as redes complexas, tendo sua definição como um grafo que apresenta uma estrutura topol~~ó~~~~o~~gi~~c~~a não trivial (METZ *et al*., 2007). Bessa *et al.* (2009) descreve que é possível realizar a classificação das redes complexas a partir de suas propriedades estatísticas, sendo elas:

a) redes regulares: são as redes mais simples, no qual todos os vértices possuem o mesmo grau;
b) redes aleatórias: são redes que têm suas conexões adicionadas com a mesma probabilidade para cada elemento, no qual todos os elementos terão aproximadamente a mesma quantidade de conexões;
c) redes de pequeno mundo: são redes que apresentam um caminho médio entre os vértices menor que o obtido em uma rede aleatória com o mesmo número de vértices e arestas. Uma rede de mundo pequeno pode ser gerada a partir de redes regulares ou pela geração de reconexões;
d) redes hierárquicas e modulares: tem como característica a relação de lei de potência entre seu coeficiente de aglomeração do vértice e seu grau. S~~,~~ ~~s~~ão redes onde pode-se separá-la~~ar ela~~ em módulos (*clusters*);

e) redes livres de escala: contêm uma distribuição de conectividade que segue a lei de potência, indicando uma ausência de escala típica. Redes livres de escala permitem a adição de vértices após a geração da rede, recebendo novos vértices em cada passo, normalmente gerando poucos vértices com grau alto (RIBEIRO, 2017).

De acordo com Metz *et al.* (2007), as redes complexas estão sendo aplicadas em diversas áreas, na resolução de problemas dos mais variados tipos. Os autores exemplificam a partir da avaliação de qualidade de texto, no qual foi utilizado um grafo direcional para a modelagem, que tinham as palavras como vértices e suas ligações representando a quantidade de vezes que a sequência de palavras era encontrada. Metz *et al.* (2017) também citam a avaliação e construção de sistemas de sumarização, controle do congestionamento de pacotes e detecção de comunidades em redes sociais.

REFERÊNCIAS

ANDRADE, Michelle H. S. **Entenda como funcionam os Sistemas de Recomendação**. [2017]. Disponível em: https://www.igti.com.br/blog/como-funcionam-os-sistemas-de-recomendacao/. Acesso em: 7 out. 2020.

BARABÁSI, Albert-Laszlo. **Linked: How Everything Is Connected to Everything Else and What It Means for Business, Science, and Everyday Life**, Plume Books, 2003.

BESSA, Aline D. *et al*. I**ntrodução às redes complexas**. 2009. 21f. Curso de Física Estatística e Sistemas Complexos, Universidade Federal da Bahia.

BRITO, Carina. **E-commerce cresce em 2020 impulsionado pela pandemia do coronavírus**. [2020]. Disponível em: https://revistapegn.globo.com/Banco-de-ideias/Varejo/noticia/2020/08/e-commerce-cresce-em-2020-impulsionado-pela-pandemia-do-coronavirus.html. Acesso em: 26 set. 2020.

CARVALHO, Maria A. A. S. **Recomendação baseada em modularidade**. 2016. 126f. Tese (Doutorado em Ciência da Computação) - Curso de Pós-Graduação em Ciência da Computação, Universidade Federal de Pernambuco, Recife.

CARVALHO, Alexsandro M. **Dinâmica de doenças infecciosas em redes complexas**. 2012. 77f. Tese (Doutorado em Ciência) - Curso de Pós-Graduação em Física, Universidade Federal do Rio Grande do Sul, Porto Alegre.

CAZELLA, Silvio C.; NUNES, Maria Augusta S. N.; REATEGUI, Eliseo B. **A Ciência da Opinião:** Estado da arte em Sistemas de Recomendação. 2010

LUNARDI, Guilherme. **12 dados que comprovam o crescimento do e-commerce no Brasil**. [2018]. Disponível em: https://www.ecommercebrasil.com.br/artigos/12-dados-que-comprovam-o-crescimento-do-e-commerce-no-brasil/. Acesso em: 25 set. 2020.

METZ, Jean *et al*. **Redes Complexas**: conceitos e aplicações. São Carlos, 2007.

MOTTA, Claudia L. R. *et al*. **Sistemas Colaborativos**. 1. Rio de Janeiro: Elsevier, 2012

NAHAL, Jessica. **Introdução a redes complexas**. [2014]. Disponível em: https://blog.dp6.com.br/introdução-a-redes-complexas-df73b623d67f. Acesso em: 2 out. 2020.

PINTO, Eduardo R. **Estudo da dinâmica de epidemias em redes complexas**. 2018. 89f. Dissertação (Mestrado em Biometria) - Curso de Pós-Graduação em Biometria, Universidade Estadual Paulista, Botucatu.

RIBEIRO, Larissa F. **Redes sem Escala Típica:** Visão Geral, Modelos Alternativos e Técnicas Computacionais. 2017. 93f. Dissertação (Mestrado em Física) - Curso de Pós-Graduação em Física, Universidade Federal do Rio Grande do Norte, Natal.

SANTOS, Rafael J. P. **Análise da rede de produtos comprados em conjunto no comércio eletrônico**. 2019. 130f. Dissertação (Mestrado em Ciência da Computação e Matemática Computacional) - Curso de Pós-Graduação em Ciência da Computação, Universidade de São Paulo, São Paulo.

SILVA, Marcos H. P. D. **Pontes de Königsberg: destruí-las é a solução**. [2020]. Disponível em: https://www.blogs.unicamp.br/zero/2020/03/20/pontes-de-konigsberg-destrui-las-e-a-solucao/. Acesso em: 19 nov. 2020.

SILVA, Natai B. **Recomendação de relacionamentos em redes sociais baseada em grafos**. 2010. 94f. Dissertação (Mestrado em Ciência da Computação) - Curso de Pós-Graduação em Ciência da Computação, Universidade Federal de Pernambuco, Recife.

TERRA, Tiago. **Comércio eletrônico cresce com pandemia do COVID-19, diz ABComm**. [2020]. Disponível em: https://www.mundodomarketing.com.br/ultimas-noticias/38602/comercio-eletronico-cresce-com-pandemia-do-covid-19-diz-abcomm.html. Acesso em: 21 set. 2020.

ZHONG *et al*. Study on Directed Trust Graph Based Recommendation for E-commerce System**. International Journal of Computers, Communications and Control**. v. 9, n. 4, p. 510-523, jun. 2014

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): Diogo Warmeling

Avaliador(a):
Marcel Hugo

	ASPECTOS AVALIADOS[1]	atende	atende parcialmente	não atende
ASPECTOS TÉCNICOS	1. INTRODUÇÃO O tema de pesquisa está devidamente contextualizado/delimitado?	X		
	O problema está claramente formulado?	X		
	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		
	As referências contemplam adequadamente os assuntos abordados (são indicadas obras atualizadas e as mais importantes da área)?	X		
ASPECTOS METODOLÓGICOS	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: (X–) APROVADO ( ) REPROVADO

Assinatura: _____ Data:
Blumenau, 01/12/2020

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

PROJETO TCC - BCC	ANO/SEMESTRE:	2020.2

USO DE REDES COMPLEXAS PARA ANÁLISE DE CONSUMO EM ECOMMERCE

Diogo Warmeling

Aurélio Faustino Hoppe – orientador

1 INTRODUÇÃO

Com o crescimento dos acessos à internet e a disseminação de dispositivos móveis, os comércios eletrônicos (e-commerce) vêm crescendo de forma exponencial nos últimos anos e, com a pandemia do COVID-19, o crescimento foi ainda maior, tornando-se o método mais comum de compra para diversos setores da economia brasileira (TERRA, 2020). Segundo os dados da 42ª edição do Webshoppers, um estudo sobre e-commerces do Brasil elaborado semestralmente pela Ebit e Elo, o crescimento nas vendas do primeiro semestre de 2020 foi de 47%, caracterizando-se como dobro dos registrados nos últimos anos. Isso, segundo Lunardi (2018), aconteceu devido a entrada de novos adeptos a esse meio de compra, resultando em um crescimento de 40%, chegando a 41 milhões de pessoas.

Brito (2020) destaca que esse crescimento não ficou restrito somente aos *marketplaces*, mas também atingiu muitos comércios eletrônicos de pequeno e médio porte, dos quais, a maior parte não aproveita o grande potencial que a análise dos dados de acessos de seus clientes fornece. Tais análises possibilitam entender o comportamento dos usuários para mantê-los mais tempo em seus *sites*, e conseguir recomendar o produto desejado antes da desistência da compra, assim como auxiliar os clientes a encontrarem as melhores opções disponíveis, aumentando a taxa de efetivação da compra por parte dos usuários. No entanto, a grande maioria dos *sites* apenas disponibilizam listas de produtos semelhantes e associados ao acesso do usuário, desconsiderando o seu perfil, que pode estar seguindo um padrão de acessos.

Segundo Santos (2019) existem diversas pesquisas sobre recomendações em plataformas de comércio eletrônico, que em sua maioria fazem cálculo de similaridade entre entidades (compradores e produtos) para recomendar novos produtos a um potencial comprador. Dentre elas, encontram-se uma grande quantidade de técnicas e algoritmos disponíveis pensadas para análise de dados tais como sistemas de recomendações, algoritmos de aprendizado de máquina e redes complexas, que permitem encontrar relações e agrupar itens, analisar o fluxo de navegação ou o comportamento dos usuários. Tais estruturas são baseadas em grafos, sendo composto por um conjunto de nós, ligados por arestas e que possui uma estrutura topológica não trivial. Elas podem representar diversos aspectos do mundo, como modelar pessoas e suas conexões nas redes sociais, páginas da *web* e seu fluxo de acesso, entre outros inúmeras possibilidades (BARABÁSI, 2003).

Zhong *et al.* (2014) afirmam que um sistema de recomendação baseado em grafos é mais eficiente do que um modelo baseado em similaridade entre nós (perfis similares de usuários, por exemplo). Diante desta afirmação, este trabalho tem como objetivo gerar redes complexas utilizando grafos, baseadas no fluxo de navegação dos usuários em plataformas de comércio eletrônico, buscando entender e contribuir com o aumento das efetivações de compra.

1.1 OBJETIVOS

Este trabalho tem como objetivo gerar redes complexas utilizando grafos, baseadas no fluxo de navegação dos usuários em plataformas de comércios eletrônicos, buscando entender e contribuir com o aumento das efetivações de compra.

Os objetivos específicos são:

a) utilizar redes complexas para gerar agrupamentos de produtos ou de comportamentos dos usuários para fazer recomendações;
b) analisar as estruturas das redes complexas, calculando métricas e identificando correlações entre a navegação e as compras;
c) identificar os maiores pontos de saídas do e-commerce a partir das redes geradas;
d) integrar a ferramenta a API Suaview.

2 TRABALHOS CORRELATOS

Neste capítulo são apresentados três trabalhos que apresentam semelhanças com o trabalho proposto. A seção 2.1 discorre sobre um sistema de recomendação baseada em agrupamentos (CARVALHO, 2016). A seção 2.2 apresenta um algoritmo para recomendação de relacionamentos em redes sociais (SILVA, 2010). Por fim, a seção 2.3 aborda a construção de um *webcrawler* e a identificação de grupos de produtos comprados em conjunto (SANTOS, 2019).

2.1 RECOMENDAÇÃO BASEADA EM MODULARIDADE

Carvalho (2016) desenvolveu um sistema para gerar recomendações de produtos com base em agrupamentos, utilizando algoritmos de redes complexas, que precisava ser performático e ter baixa taxa de erro. Para isso, a autora utilizou o LensKit, *framework* que dá suporte a sistemas de recomendação.

Segundo Carvalho (2016) foram utilizadas as bases Movie Lens 100K, MovieTweetings, Movie Lens 1M, BookCrossing e Jester. As três primeiras bases associam usuários a filmes. Já a quarta base associa os usuários a livros e a quinta base faz a associação dos usuários com piadas, todas com as qualificações dadas pelos usuários aos itens. A partir delas, foram executados os algoritmos Louvain com as modularidades de Newman e Girvan, Agrupamento com Movimento de Arestas (AMA) e Agrupamento com Movimento de Vértices (AMV) com a modularidade Suzuki-Wakita.

Com base nos experimentos, Carvalho (2016) optou pela a utilização do método Agrupamento com Movimento de Arestas (AMA), pois apresentou os melhores tempos de treinamento e predição. Na análise, a autora gerou 6 grupos de itens e 6 grupos de usuários a partir da base Movie Lens 100K. Utilizando o algoritmo AMV, Carvalho(2016) encontrou alguns padrões, como, um grupo com uma quantidade grande de itens, porém poucas arestas, significando que são filmes pouco avaliados; mulheres assistiram mais os filmes dos gêneros infantil e de animação, além de assistirem poucos filmes de ação. Carvalho (2016) também notou que o grupo que contém em sua maioria usuários entre os 20 e 40 anos assistem mais filmes de suspense e romance. Já o grupo que contém em sua maioria usuários acima dos 60 anos assistem mais filmes de ação e comédia.

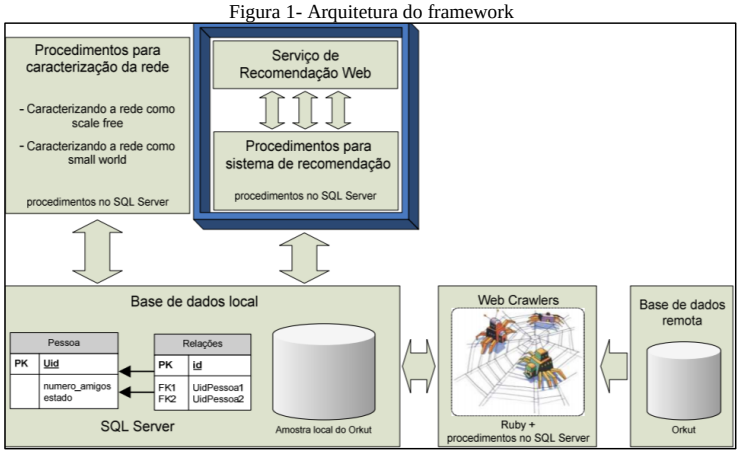
Comentado [AS1]: Rever pontuação. Está confuso.

Carvalho (2016) conclui que redes complexas e sistemas de recomendações podem gerar bons resultados. Já a hipótese de que quanto maior a métrica de modularidade, menor a taxa de erro, se provou verdadeiro em sua maioria, a não ser na utilização do algoritmo AMV com a modularidade Suzuki-Wakita. Como trabalhos futuros, Carvalho (2016) aponta a necessidade da criação de um algoritmo que encontre a quantidade ideal de grupos, para que não haja a necessidade da passagem de parâmetros, assim como realize a reorganização da rede nas inserções de vértices e faça recomendações quando não for encontrada ligações.

2.2 RECOMENDAÇÃO DE RELACIONAMENTOS EM REDES SOCIAIS BASEADA EM GRAFOS

Silva (2010) desenvolveu um sistema de recomendação de relacionamentos em redes sociais, utilizando grafos. O autor propôs um método híbrido utilizando algoritmo genético e dados topológicos, utilizando a rede social Orkut como base. Para o desenvolvimento do método foi construído um *framework* com 4 módulos. *webcrawlers*, responsáveis pela captura das informações; Base de dados local, responsável pelo armazenamento das informações; Procedimentos para caracterização da rede, responsável pela geração dos grafos; Serviço de Recomendação Web, responsável por buscar dos grafos o grupo de itens para recomendação. Esses módulos são representados pela Figura 1.

Figura 1- Arquitetura do framework



Fonte: Silva (2010).

A partir da Figura 1, pode-se perceber que existe um módulo referente a base de dados, utilizando SQL Server, contendo duas tabelas, uma tabela Pessoa representando os nós e a tabela Relações representado os vértices do grafo. Já o módulo w*ebcrawlers* é responsável pela busca dos dados da rede social Orkut, salvando os dados

Comentado [AS2]: Coloque o recurso de referência cruzada para figura/quadro/tabela. Faça isso em todo o texto.

encontrados na base de dados local. O *crawler* utiliza um usuário como nó raiz, realizando uma busca em largura para acessar os amigos e fazer os mapeamentos. Foram utilizadas duas bases, uma com usuários brasileiros e outra de indianos. O módulo de caracterização da rede foi desenvolvido em T-SQL a partir de cálculos e constatações das redes *Scale Free* e *Small World*. Por fim, o módulo de serviço de recomendação, também escrito em T-SQL, executa os cálculos de índices, a ponderação e a partir disso, gera as recomendações finais. Esse processo passa por duas etapas, a filtragem e ordenação. A filtragem é responsável por limitar a quantidade de vértices para a ordenação. Neste sentido foi utilizado o conceito de *clustering coefficient*, alcançando os vértices com dois saltos. Já a ordenação utiliza um mecanismo de autoajuste para regular os pesos. Nele, são utilizados 3 índices, o primeiro é composto pela quantidade de amigos em comum. O segundo se refere a densidade do conjunto formado no primeiro índice. No terceiro é a densidade do conjunto formado pelos adjacentes dos itens. A partir desses índices, gera-se a ponderação e a recomendação.

Silva (2010) realizou dois experimentos. No primeiro foi executado a recomendação para dois conjuntos de dados, executando várias vezes ajustando os pesos manualmente com variações pequenas, até que se atingisse o menor valor para a função de otimização. Os melhores resultados obtidos foram um posicionamento médio de 513 e 272 para a sub-rede brasileira e indiana respectivamente, dentro do conjunto de candidatos. Para avaliar a eficiência, o autor utilizou seu próprio usuário, chegando em uma taxa de 75% de acertos nas 20 primeiras recomendações e 50% nas 100 primeiras recomendações.

No segundo experimento Silva (2010) realizou a adição do algoritmo genético e uma validação cruzada com a estratégia *Friend of friend* (FOF). Os dados utilizados vieram da rede social Oro-Aro do C.E.S.A.R, dos quais foram enviadas 10 recomendações para um grupo de 70 usuários, onde 31 realizaram os testes. As recomendações tiveram uma porcentagem de 77.69% de aceitação, 5% superior a estratégia FOF.

Silva (2010) conclui que no primeiro experimento os resultados obtidos foram satisfatórios, porém o método de validação é falho por não possibilitar a reprodução em um ambiente externo. Ele também escreve que a caracterização da rede foi essencial para a elaboração do mecanismo de recomendação. Como melhorias Silva (2010) sugere a adição de outros índices na ordenação para melhorar a ponderação e a implementação de um módulo de algoritmo genético para aprimorar a otimização, assim como a realização de testes com outras localizações ou a utilização de alguma API para a extração de dados no lugar do *webcrawler*, pois isso limita apenas ao uso de redes sociais, sendo que qualquer alteração no site pode ocasionar falha de extração dos dados.

Formatado: Realce

2.3 ANÁLISE DA REDE DE PRODUTOS COMPRADOS EM CONJUNTO NO COMÉRCIO ELETRÔNICO

Santos (2019), tinha como objetivos, a construção de um *webcrawler* para obter uma base de dados, assim como a construção de uma rede de produtos comprados em conjunto para analisar e identificar correlações entre produtos e validar hipóteses. Algumas dessas hipóteses são, produtos melhores avaliados serem *hubs* de rede e validar se somente as métricas de rede são o suficiente para o treinamento do modelo.

Comentado [AS3]: Não tem vírgula depois da referência pois não existe vírgula entre sujeito e verbo.

Para montar a base de dados foi criado um *webcrawler*, utilizando o *selenium*, para capturar os produtos comprados em conjunto do e-commerce da Amazon. A busca é iniciada com um produto selecionado manualmente, e posteriormente é executado de forma automática como uma busca em largura pelos outros produtos comprados em conjunto. Ainda segundo Santos (2019), o *crawler* foi desenvolvido em PHP com a execução em `Bash` para fazer execuções em paralelo. Após a extração, os dados são enviados separados em dois arquivos `csv` para uma aplicação em python que utiliza a biblioteca `NetworkX` para processar os dados e gerar as redes, que são salvas em um *dataframe* do pandas, para gerar as previsões de estimativa de preço e as previsões de ligação entre nós. Para criar as estimativas de preços foram utilizados o Regressor Random Forest e o coeficiente Gini na montagem das árvores com maiores relevâncias. O resultado obtido pode ser visualizado na Figura 2, onde pode-se ver uma linha tracejada verde que representa o preço previsto do produto sendo igual ao preço real do produto analisado. Já os pontos azuis representam as predições, os que se localizam acima possuem preço previsto superior ao real, e os que se localizam abaixo tem preço previsto inferior ao real.

Comentado [AS4]: Tem vírgula depois da referência em função do “Segundo”.

Comentado [AS5]: Nomes próprios não são escritos em itálico.

Formatado: Fonte: Não Itálico

Para a previsão de ligação entre nós foram utilizados todos os atributos no treinamento e a coluna `existe_ligacao_com_o_1` como alvo e a biblioteca `Scikit_Learn` para o treinamento. Por se tratar de classificações binárias bastante desbalanceadas foi utilizada a matriz de confusão para fazer a análise dos resultados, analisando os falsos positivos e positivos negativos. Dos 768 itens processados, Santos (2019) obteve 634 identificados corretamente, 1 positivo negativo e 133 falsos positivos. Para uma segunda execução foi considerado a utilização da Receiver Operating Characteristic Curve (ROC) para identificar o ponto com maior precisão, um ponto onde não seja sensível demais onde terá muitos falsos positivos e nem específico demais, onde obteria muito positivos falsos, algo como regular a sensibilidade de um detector de incêndio, não sendo sensível demais a ponto de disparar a todo momento, tendo muitos casos de falso positivo e nem específico demais, onde pode ter positivos falsos. Com a curva ROC ele obteve um limiar de probabilidade de 0,66, onde valores superiores

Comentado [AS6]: Nomes de pacotes, classes, entidades, atributos, métodos ou diálogos de interface devem ser escritos em fonte courier. Rever isso em todo o texto.

Comentado [AS7]: Nomes próprios não são escritos em itálico.

Formatado: Fonte: Não Itálico

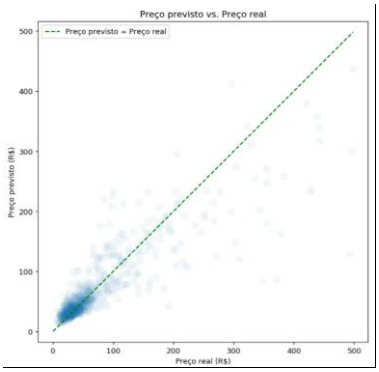
Formatado: Fonte: Não Itálico

Formatado: Fonte: Não Itálico

Comentado [AS8]: Confuso.

eram considerados positivos e inferiores negativos. Após a aplicação do resultado da curva ROC Santos (2019) obteve resultados consideravelmente superiores, reduzindo o número de falsos positivos de 133 para 93.

Figura 2 - Gráfico demonstrando o resultado da estimativa de preço



Fonte: Santos (2019)

Santos (2019) conclui que *hubs* de rede nem sempre serão os mais bem avaliados, porém confirmou a correlação deles serem os mais bem posicionados nas listas de mais vendidos. Santos (2019) também confirmou a hipótese de que a utilização exclusiva dos atributos de métrica foram suficientes para obter melhores resultados em comparação com a *baseline* do modelo de previsão de ligações entre nós, ao contrário dos resultados obtidos na precificação dos produtos.

Santos (2019) sugere a análise da predição de relação entre todos os nós da rede, além de revisar os atributos utilizados no treinamento para poder otimizar o processo. Outro ponto é a integração com APIs de comércios eletrônicos, para que não seja necessária a extração das informações com o *webcrawler*.

3 PROPOSTA DA FERRAMENTA

Neste capítulo são definidas as justificativas para a elaboração deste trabalho, assim como os requisitos funcionais, não funcionais e a metodologia que será aplicada no desenvolvimento.

3.1 JUSTIFICATIVA

No Quadro 1 é apresentado um comparativo entre os trabalhos correlatos. As linhas representam as características e as colunas os trabalhos.

Quadro 1 - Comparativos entre os trabalhos correlatos

Correlatos / Características	Carvalho (2016)	Silva (2010)	Santos (2019)
Obtenção dos dados	API	Webrawler e API	Webcrawler
Método de processamento	LensKit	T-SQL	Scikit Learn
Tipo de item analisado	Avaliações de filmes	Amizades em redes sociais	Produtos comprados em conjunto
Usa grafo bipartido	Sim	Não	Não
Principal algoritmo / técnica	Agrupamento com Movimento de Vértices (AMV)	Algoritmo próprio	Random Forest

Fonte: elaborado pelo autor.

A partir do Quadro 1, observa-se duas estratégias de obtenção de dados, enquanto Carvalho (2016) utilizou uma API que forneciam as informações necessárias para a montagem das redes, Santos (2019) em seu trabalho criou um *webcrawler* que navega pela loja da Amazon e captura os dados para sua base. Já Silva (2010) utiliza ambas as técnicas, criando um *webcrawler* que captura informações da rede social Orkut para o primeiro experimento e utiliza uma API da C.E.S.A.R para o seu segundo experimento.

Carvalho (2016) utilizou grafos bipartidos, onde definiu nós do tipo usuário e nós do tipo filme para a sua análise, enquanto Silva (2010) e Santos (2019) não fizeram uso de tal particularidade dos grafos, já que no projeto

de Silva (2019) somente foram mapeados os usuários e no projeto de Santos (2019) foram mapeados os itens. Quanto aos métodos de processamentos, pode-se observar que dois trabalhos utilizam as bibliotecas da linguagem Python para as predições e recomendações, Carvalho (2016) com o LensKit e Santos (2019) com o Scikit Learn, enquanto Silva (2010) utilizou a linguagem T-SQL. ~~Olhando para~~Ao analisar os principais algoritmos e técnicas, percebe-se que Carvalho (2016) utiliza o algoritmo AMV para realizar o agrupamento, Santos (2019) aplicou o Random Forest para a criação das estimativas de preços e Silva (2010) desenvolveu seu próprio algoritmo utilizando uma série de técnicas e conceitos para as etapas de classificação, filtragem e ponderação.

Dessa forma, conclui-se que dentro dos trabalhos pesquisados não foi encontrado algum que tenha como objetivo gerar recomendações e análises com base nos acessos dos usuários nos sites ou plataformas de comércio eletrônico. Percebe-se também que nenhum trabalho teve em sua proposta a ideia de realizar testes e validações práticas com o usuário final. Sendo assim, o trabalho proposto se difere pois pretende realizar uma análise alternativa, através do uso dos conceitos de redes complexas, pensada no usuário e sua navegação, tendo como objetivo manter seu interesse por um período maior de tempo, e que o conduza com mais naturalidade ao item desejado. Além disso, a ferramenta a ser desenvolvida será integrada e testada através de uma solução de e-commerce já implantada no mercado. Contudo, a partir deste trabalho, espera-se que os usuários encontrem os itens de interesse mais rapidamente, impulsionando indiretamente a efetivação das vendas.

Comentado [AS9]: Acredito que aqui você queria dizer “Com isso”

3.2 REQUISITOS PRINCIPAIS DO PROBLEMA A SER TRABALHADO

A ferramenta a ser desenvolvida deverá:

a) identificar as páginas que são pontos de saídas (nós folhas) (Requisito Funcional – RF);
b) prever as prováveis próximas páginas a serem acessadas pelo usuário (RF);
c) estabelecer grupos e perfis de usuários a partir das páginas acessadas (RF);
d) gerar recomendações com bases nos grupos de produtos e usuários gerados (RF);
e) utilizar medidas de centralidade, complexidade, distâncias, estrutura de comunidades e distribuição de grau para identificar e estabelecer correlações entre usuários, acessos, saídas e conversões de compras (RF);
f) disponibilizar uma interface gráfica que possibilite ao usuário/gestor visualizar estatísticas de acessos e correlações identificadas (RF);
g) ser implementado na linguagem de programação Python (Requisito Não Funcional – RNF);
h) ser modelada seguindo os princípios de redes complexas (RNF);
i) ser integrado com a API Suaview (RNF).

3.3 METODOLOGIA

O trabalho será desenvolvido observando as seguintes etapas:

a) levantamento bibliográfico: pesquisar e estudar sobre redes complexas, sistemas de recomendação e trabalhos correlatos;
b) levantamento dos requisitos: baseando-se nas informações da etapa anterior, reavaliar os requisitos propostos para a aplicação;
c) integração com API Suaview: implementar a estrutura que irá se comunicar com a API Suaview para a obtenção dos dados;
d) definição das ferramentas para modelagem e armazenamento das redes: pesquisar e escolher as ferramentas mais apropriadas para a modelagem e armazenamento das redes;
e) definição das técnicas de análise de redes: pesquisar e (re)definir as técnicas/algoritmos que serão utilizadas para analisar redes complexas;
f) modelagem da rede: a partir do item (d) modelar a estrutura da rede de forma a facilitar a análise das correlações de acordo com o problema proposto;
g) implementação da rede: implementar as funções de análise da rede considerando os itens (d), (e) e (f), tendo a linguagem Python como base;
h) testes: avaliar a performance assim como validar a aderência e eficiência da rede em relação aos resultados alcançados e os que realmente acontecerão no e-commerce em termos de saída ou efetivação de compras e recomendações de produtos.

As etapas serão realizadas nos períodos relacionados no Quadro 2.

Quadro 2-Cronograma de atividades a serem realizadas

	2021									
	fev.		mar.		abr.		maio		jun.	
etapas / quinzenas	1	2	1	2	1	2	1	2	1	2
levantamento bibliográfico										
levantamento dos requisitos										
integração com API Suaview										
definição das ferramentas para modelagem e armazenamento das redes										
definição das técnicas de análise de redes										
modelagem da rede										
implementação da rede										
testes										

Fonte: elaborado pelo autor.

4 REVISÃO BIBLIOGRÁFICA

Este capítulo tem como objetivo explorar os principais assuntos que fundamentarão o estudo a ser realizado. A seção 4.1 aborda sistemas de recomendação. Já a seção 4.2 discorre sobre redes complexas.

4.1 SISTEMAS DE RECOMENDAÇÃO

De acordo com Andrade (2017), os sistemas de recomendação são um conjunto de algoritmos que utilizam técnicas de aprendizado de máquina e recuperação de informação. Motta *et al.* (2012) destacam que os sistemas de recomendação têm como princípio a ideia, de que, o que é relevante para mim, também será relevante para quem possuir um perfil similar ao meu. Ainda segundo os autores, parte das técnicas de recomendação fazem uso das avaliações feitas pelos usuários, para realizar o entendimento e classificação do perfil.

Motta *et al.* (2012) citam a navegação pelas páginas web e avaliações dos conteúdos como um bom exemplo de dados de entrada, pois possibilitam entender os gostos dos usuários. Motta *et al.* (2012) também apontam que as saídas dos sistemas de recomendação podem ser apresentadas em forma de lista estatística, sugestões, avaliações ou resenhas, dependendo do objetivo de cada recomendação.

Cazella *et al.* (2010) citam algumas técnicas vinculadas aos sistemas de recomendação, que surgiram com o propósito de identificar padrões de comportamento dos usuários. Segundo os autores, tais técnicas definem o funcionamento dos sistemas de recomendação, sendo elas:

a) filtragem de informação: a ideia por trás dessa técnica é diminuir a quantidade de informação transitada. Ela utiliza diversos processos para filtrar as informações e entregá-las somente para quem realmente precisa;
b) filtragem baseada em conteúdo: é uma abordagem que tem foco nas propriedades dos itens, onde a similaridade será determinada a partir das semelhanças entre as propriedades dos itens;
c) filtragem colaborativa: tem foco na relação entre os usuários e os itens. ~~C~~ com a similaridade sendo determinada pelos grupos de usuários que atuam (visualizam, compram ou avaliam) sobre os mesmos itens;
d) filtragem híbrida: a filtragem híbrida tem como objetivo unir os pontos fortes das filtragens baseadas em conteúdo e colaborativa, para gerar um sistema que atenda melhor as necessidades dos usuários;
e) filtragem baseada em outros contextos: essa técnica tenta se aproximar mais das recomendações feitas pelos humanos. Utilizando não somente informações normalmente utilizadas, como, preferências do usuário, localidade, idade, entre outras, mas também informações mais complexas, como, traços da personalidade, emoção e habilidades sociais;
f) descoberta de conhecimento em base de dados: essa técnica é aplicada com ferramentas de mineração de dados, que irão gerar relações de associação, classificação ou agrupamento dos itens, usuários, ou itens e usuários;
g) mineração de textos: técnica inspirada na mineração de dados, tem como objetivo extrair informações de texto em linguagem natural. É amplamente utilizada com a filtragem baseada em conteúdo, sendo muito importante na análise do conteúdo que descreve o item a ser recomendado. Essa técnica requer um pré-processamento que limpe o texto a ser analisado, realizando uma redução dos termos para um radical comum, fazendo a remoção de palavras pouco significativas (*stopwords*), como conjunções e focando na manutenção dos substantivos que normalmente são as palavras mais representativas do texto.

Comentado [AS10]: Normalmente frases não iniciam no gerúndio. Gerúndio complementa alguma ideia.

Cazella *et al.* (2010) ressaltam que cada técnica tem seu propósito e uso, mas duas ou mais técnicas podem ser utilizadas em conjunto. Os autores citam um sistema de recomendações de cursos que faz uso da técnica de

filtragem baseada em outros contextos, no qual é utilizada uma lista de características objetivas buscadas a partir de uma base de dados e emocionais respondidas pelo usuário de forma opcional. Neste caso, eles constataram que para os usuários respondentes dos questionários de características emocionais, as recomendações foram mais precisas. Além disso, Cazzela *et al*. (2010) descrevem o uso de técnicas em conjunto, fazendo referência a um sistema de recomendações de páginas web, no qual é utilizado a técnica de filtragem baseada em conteúdo para gerar os perfis dos usuários e, posteriormente compará-los para identificar usuários similares e através da filtragem colaborativa, gerar recomendações. Outros exemplos de sistemas de recomendação são os comércios eletrônicos da Amazon e Ebay. De acordo com Cazella *et al.* (2010), o comércio eletrônico da Amazon faz uso de diversas técnicas, como a mineração de textos dos comentários dos usuários, uso de filtragem colaborativa para recomendar produtos, entre outras técnicas. No Ebay, os autores citam o uso da filtragem de informação, no qual os clientes podem indicar itens que tenham interesse, para que o *site* possa enviar as recomendações em uma periodicidade definida.

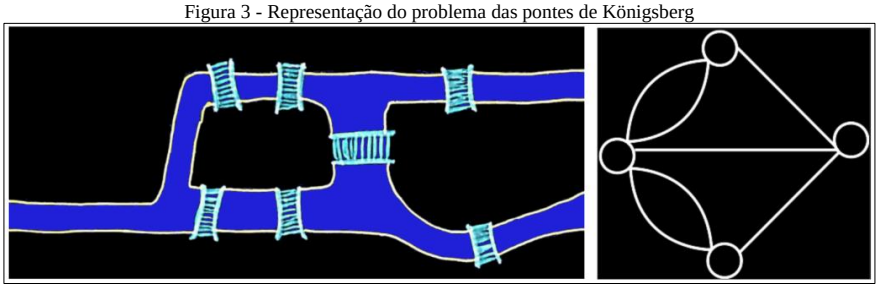
Comentado [AS11]: confuso

4.2 REDES COMPLEXAS

Segundo Carvalho (2012), redes complexas? são conjuntos de itens que se conectam entre si, com os elementos da rede sendo vértices e as conexões entre os elementos sendo arestas. Ainda, de acordo com o autor, pode-se observar uso das redes em diversas situações, desde o nível subatômico até as estruturas sociais mais complexas.

O estudo das redes complexas? teve seu início no século XVIII com Leonhard Euler, onde ele modelou o problema das pontes de Königsberg como um grafo, conforme pode ser visto na Figura 3. Euler modelou os vértices como pedaços de terra e as pontes que os conectavam como arestas. Com essa modelagem, Euler conseguiu provar que não existia um caminho que possibilitasse transitar por todas as sete pontes sem repetir nenhuma (NAHAL, 2014).

Figura 3 - Representação do problema das pontes de Königsberg



Fonte: adaptado de Silva (2020).

Pinto (2018) descreve que um grafo pode ser categorizado como direcional, no qual as arestas têm um sentido, com um vértice de origem e um vértice de destino, ou não direcional, onde o sentido da aresta não importa. Pinto (2018) também explica que arestas podem possuir intensidade, o que significa que cada aresta terá um peso. Esses pesos podem ser utilizados para calcular o melhor caminho entre dois vértices. Carvalho (2012) expõe que a característica principal de um vértice é o seu grau, que representa o número de ligações que o vértice possui, com o grau dos vértices mensurado, é possível determinar o grau médio da rede, que representa a média de conexões que cada elemento possui. Com os graus dos vértices e grau médio da rede determinados, é possível identificar os nós *Hub*. Segundo Nahal (2014), um nó é definido como *Hub* quando ele possui um grau muito superior ao grau médio, sendo determinado como um vértice muito importante pela capacidade de distribuição de informações para um número maior de vértices.

Comentado [AS12]: rever pontuação. Confuso.

A partir das definições de grafos surgiram as redes complexas, tendo sua definição como um grafo que apresenta uma estrutura topologia não trivial (METZ *et al.*, 2007). Bessa *et al.* (2009) descreve que é possível realizar a classificação das redes complexas a partir de suas propriedades estatísticas, sendo elas:

a) redes regulares: são as redes mais simples, no qual todos os vértices possuem o mesmo grau;
b) redes aleatórias: são redes que têm suas conexões adicionadas com a mesma probabilidade para cada elemento, no qual todos os elementos terão aproximadamente a mesma quantidade de conexões;
c) redes de pequeno mundo: são redes que apresentam um caminho médio entre os vértices menor que o obtido em uma rede aleatória com o mesmo número de vértices e arestas. Uma rede de mundo pequeno pode ser gerada a partir de redes regulares ou pela geração de reconexões;
d) redes hierárquicas e modulares: tem como característica a relação de lei de potência entre seu coeficiente de aglomeração do vértice e seu grau, são redes onde pode-se separar ela em módulos

(*clusters*);

e) redes livres de escala: contêm uma distribuição de conectividade que segue a lei de potência, indicando uma ausência de escala típica. Redes livres de escala permitem a adição de vértices após a geração da rede, recebendo novos vértices em cada passo, normalmente gerando poucos vértices com grau alto (RIBEIRO, 2017).

De acordo com Metz *et al.* (2007), as redes complexas estão sendo aplicadas em diversas áreas, na resolução de problemas dos mais variados tipos. Os autores exemplificam a partir da avaliação de qualidade de texto, no qual foi utilizado um grafo direcional para a modelagem, que tinham as palavras como vértices e suas ligações representando a quantidade de vezes que a sequência de palavras era encontrada. Metz *et al.* (2017) também citam a avaliação e construção de sistemas de sumarização, controle do congestionamento de pacotes e detecção de comunidades em redes sociais.

REFERÊNCIAS

ANDRADE, Michelle H. S. **Entenda como funcionam os Sistemas de Recomendação**. [2017]. Disponível em: https://www.igti.com.br/blog/como-funcionam-os-sistemas-de-recomendacao/. Acesso em: 7 out. 2020.

BARABÁSI, Albert-Laszlo. **Linked: How Everything Is Connected to Everything Else and What It Means for Business, Science, and Everyday Life**, Plume Books, 2003.

BESSA, Aline D. *et al*. I**ntrodução às redes complexas**. 2009. 21f. Curso de Física Estatística e Sistemas Complexos, Universidade Federal da Bahia.

BRITO, Carina. **E-commerce cresce em 2020 impulsionado pela pandemia do coronavírus**. [2020]. Disponível em: https://revistapegn.globo.com/Banco-de-ideias/Varejo/noticia/2020/08/e-commerce-cresce-em-2020-impulsionado-pela-pandemia-do-coronavirus.html. Acesso em: 26 set. 2020.

CARVALHO, Maria A. A. S. **Recomendação baseada em modularidade**. 2016. 126f. Tese (Doutorado em Ciência da Computação) - Curso de Pós-Graduação em Ciência da Computação, Universidade Federal de Pernambuco, Recife.

CARVALHO, Alexsandro M. **Dinâmica de doenças infecciosas em redes complexas**. 2012. 77f. Tese (Doutorado em Ciência) - Curso de Pós-Graduação em Física, Universidade Federal do Rio Grande do Sul, Porto Alegre.

CAZELLA, Silvio C.; NUNES, Maria Augusta S. N.; REATEGUI, Eliseo B. **A Ciência da Opinião:** Estado da arte em Sistemas de Recomendação. 2010

LUNARDI, Guilherme. **12 dados que comprovam o crescimento do e-commerce no Brasil**. [2018]. Disponível em: https://www.ecommercebrasil.com.br/artigos/12-dados-que-comprovam-o-crescimento-do-e-commerce-no-brasil/. Acesso em: 25 set. 2020.

METZ, Jean *et al*. **Redes Complexas**: conceitos e aplicações. São Carlos, 2007.

MOTTA, Claudia L. R. *et al*. **Sistemas Colaborativos**. 1. Rio de Janeiro: Elsevier, 2012

NAHAL, Jessica. **Introdução a redes complexas**. [2014]. Disponível em: https://blog.dp6.com.br/introdução-a-redes-complexas-df73b623d67f. Acesso em: 2 out. 2020.

PINTO, Eduardo R. **Estudo da dinâmica de epidemias em redes complexas**. 2018. 89f. Dissertação (Mestrado em Biometria) - Curso de Pós-Graduação em Biometria, Universidade Estadual Paulista, Botucatu.

RIBEIRO, Larissa F. **Redes sem Escala Típica:** Visão Geral, Modelos Alternativos e Técnicas Computacionais. 2017. 93f. Dissertação (Mestrado em Física) - Curso de Pós-Graduação em Física, Universidade Federal do Rio Grande do Norte, Natal.

SANTOS, Rafael J. P. **Análise da rede de produtos comprados em conjunto no comércio eletrônico**. 2019. 130f. Dissertação (Mestrado em Ciência da Computação e Matemática Computacional) - Curso de Pós-Graduação em Ciência da Computação, Universidade de São Paulo, São Paulo.

SILVA, Marcos H. P. D. **Pontes de Königsberg: destruí-las é a solução**. [2020]. Disponível em: https://www.blogs.unicamp.br/zero/2020/03/20/pontes-de-konigsberg-destrui-las-e-a-solucao/. Acesso em: 19 nov. 2020.

SILVA, Natai B. **Recomendação de relacionamentos em redes sociais baseada em grafos**. 2010. 94f. Dissertação (Mestrado em Ciência da Computação) - Curso de Pós-Graduação em Ciência da Computação, Universidade Federal de Pernambuco, Recife.

TERRA, Tiago. **Comércio eletrônico cresce com pandemia do COVID-19, diz ABComm**. [2020]. Disponível em: https://www.mundodomarketing.com.br/ultimas-noticias/38602/comercio-eletronico-cresce-com-pandemia-do-covid-19-diz-abcomm.html. Acesso em: 21 set. 2020.

ZHONG *et al*. Study on Directed Trust Graph Based Recommendation for E-commerce System**. International Journal of Computers, Communications and Control**. v. 9, n. 4, p. 510-523, jun. 2014

Comentado [AS13]: Colocar em ordem alfabética de nome

Comentado [AS14]: Faltou o nome do autor.

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): Diogo Warmeling _____

Avaliador(a): Andreza Sartori_____

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

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

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

PARECER: (x) APROVADO ( ) REPROVADO

Assinatura: _____ Data: 01/12/2020 _____

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