UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
DEPARTAMENTO DE INFORMÁTICA

GABRIELA VIEIRA LEON

AGRUPAMENTO DE GRAFOS E SISTEMA DE RECOMENDAÇÃO: UM ESTUDO DE CASO DAS AVALIAÇÕES DA AMAZON

TRABALHO DE CONCLUSÃO DE CURSO

CURITIBA
2021

GABRIELA VIEIRA LEON

AGRUPAMENTO DE GRAFOS E SISTEMA DE RECOMENDAÇÃO: UM ESTUDO DE CASO DAS AVALIAÇÕES DA AMAZON

Clustering graphs and recommendation systems: a case study of Amazon ratings

Trabalho de Conclusão de Curso apresentado(a) como requisito para obtenção do título(grau) de Especialista em Ciência de Dados e suas Aplicações, do Departamento de Informática, da Universidade Tecnológica Federal do Paraná (UTFPR).

Orientador: Prof. Dr. Marcelo de Oliveira Rosa

CURITIBA

2021

Ministério da Educação
UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
UTFPR - CAMPUS CURITIBA
DIRETORIA-GERAL - CAMPUS CURITIBA
DIRETORIA DE PESQUISA E PÓS-GRADUAÇÃO - CAMPUS CURITIBA
DEPARTAMENTO DE APOIO DAS ESPECIALIZAÇÕES LATO-SENSU DOS CURSOS DE INFORMÁTICA - CAMPUS CURITIBA

UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

TERMO DE APROVAÇÃO

AGRUPAMENTO DE GRAFOS E SISTEMA DE RECOMENDAÇÃO: UM ESTUDO DE CASO DAS AVALIAÇÕES DA AMAZON

por

Gabriela Vieira Leon

Este Trabalho de Conclusão de Curso foi apresentado às 19h00min do dia 21 de julho de 2021 por videoconferência como requisito parcial à obtenção do grau de Especialista em Ciência de Dados e suas Aplicações na Universidade Tecnológica Federal do Paraná - UTFPR - Campus Curitiba. A aluna foi arguida pela Banca de Avaliação abaixo assinados. Após deliberação, a Banca de Avaliação considerou o trabalho aprovado.

_____

Prof. Dr. Marcelo de Oliveira Rosa (Presidente)

DAELT - CT - UTFPR

_____

Prof. Dr. Matheus Garibalde Soares de Lima (Membro)

GRUPO BOTICÁRIO

_____

Profa. Dra. Rita Cristina Galarraga Berardi (Membro)

DAINF - CT - UTFPR

O Termo de Aprovação assinado encontra-se no sistema SEI- Processo nº 23064.027988/2021-48

Dedico este trabalho a minha família e aos meus amigos, pelos momentos de ausência.

AGRADECIMENTOS

Este trabalho não poderia ser terminado sem a ajuda de diversas pessoas às quais presto minha homenagem. Certamente esses parágrafos não irão atender a todas as pessoas que fizeram parte dessa importante fase de minha vida. Portanto, desde já peço desculpas àquelas que não estão presentes entre estas palavras, mas elas podem estar certas que fazem parte do meu pensamento e de minha gratidão.

A minha família, pelo carinho, incentivo e total apoio em todos os momentos da minha vida, especialemente neste último ano que foi o mais desafiador.

À coordenadora do curso, Rita, pela dedicação em fazer do curso o melhor possível, mesmo com todas as adaptações e desafios deste ano pandêmico.

Ao meu orientador, Marcelo, pela confiança depositada e que, não mediu esforços para me auxiliar nas suas aulas mesmo com a diferença de fuso horário e compreendeu o momento que eu estava vivendo.

A todos os professores e colegas do departamento, que ajudaram de forma direta e indireta na conclusão deste trabalho.

Aos meus colegas, que tornaram as noites mais leves e compartilharam suas experiências.

Enfim, a todos os que de alguma forma contribuíram para o meu desenvolvimento durante o curso e para a realização deste trabalho.

RESUMO

LEON, Gabriela Vieira. **Agrupamento de grafos e sistema de recomendação: um estudo de caso das avaliações da Amazon**. 2021. 34 f. Trabalho de Conclusão de Curso (Especialização em Ciência de Dados e suas Aplicações) – Universidade Tecnológica Federal do Paraná. Curitiba, 2021.

Na pandemia, devido ao distanciamento social e as restrições do comércio, as vendas *online* cresceram. Sem os vendedores, é preciso que o próprio *site* faça recomendações aos clientes para que eles comprem outros produtos. Os dados das vendas e avaliações e suas relações podem ser descritos como grafos, mais especificamente, redes bipartidas, e estas podem ser utilizadas para recomendar produtos a clientes. O objetivo deste trabalho é identificar os grupos de clientes com preferências semelhantes utilizando agrupamento de grafos pelo algoritmo BRIM e, a partir da suas preferências, recomendar outros produtos. Foram estudados quatro cenários, considerando o uso de agrupamento e de número de conexões, e diferentes etapas de limpeza da base de dados. O melhor cenário indica que devem ser desconsiderados usuários que avaliaram apenas um produto e aqueles que avaliaram todos os produtos com nota máxima e avaliações duplicadas. O agrupamento se mostrou eficaz com uma modularidade de 0.843. Por requerer poucos dados e ser de fácil implementação, a metodologia pode contribuir para o aumento de vendas e evitar o fechamento de pequenos comércios. As limitações encontradas foram o tempo para gerar o agrupamento e que a recomendação só é realizada para usuários que fizeram mais de uma avaliação. Por isso, muitos produtos nem entram na base de avaliação devido à limpeza e, consequentemente, o sistema acaba não recomendando produtos pouco comprados ou recém adicionados.

Palavras-chave: Agrupamento. Grafos. Recomendação. Redes bipartidas. BRIM.

ABSTRACT

LEON, Gabriela Vieira. **Clustering graphs and recommendation systems: a case study of Amazon ratings**. 2021. 34 p. Trabalho de Conclusão de Curso (Especialização em Ciência de Dados e suas Aplicações) – Universidade Tecnológica Federal do Paraná. Curitiba, 2021.

In the pandemic, due to social distance and trade restrictions, online sales grew. Without the sellers, the site itself needs to make recommendations to customers so that they buy other products. Sales and valuation data and their relationships can be described as graphs, more specifically, bipartite networks and these can be used to recommend products to customers. The objective of this work is to identify groups of customers with similar preferences using the BRIM algorithm of graph clustering and, based on their preferences, recommend other products that may be of interest. Four scenarios were studied, considering the use of clustering and number of connections, and different steps of database cleaning. The best scenario indicates that users who rated only one product and those who rated all products with the maximum grade and duplicate ratings should be disregarded. The clustering proved to be effective with a modularity of 0.843. Because it requires a small quantity of data and is easy to implement, the methodology can contribute to increasing sales and avoiding the closing of small businesses. The limitations found were the time to generate the clusters and that the recommendation is only made for users who have performed more than one assessment. For this reason, many products are not even considered in the evaluation base due to cleaning and, therefore, the system ends up not recommending products that are rarely purchased or recently added.

Keywords: Clustering. Graphs. Recommendation. Bipatites networks. BRIM.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Uma aplicação importante da inteligência artificial é a identificação de semelhanças entre as instâncias de um conjunto de dados. Mesmo que muitos desses dados sejam heterogêneos, é possível agrupar instâncias de acordo com algumas medidas de similaridade. Essa tarefa é conhecida como *clustering* e os grupos criados, *clusters*.

Em alguns casos, essas instâncias estão de alguma forma relacionadas e podem ser organizadas em uma estrutura especial chamada de grafos. Os grafos são compostos de nós (instâncias) e arestas - conexões entre nós que representam sua relação. Os grafos representam redes complexas.

O agrupamento de grafos consiste em reunir os nós mais semelhantes levando em consideração suas relações. Os nós mais relacionados, com mais arestas conectando-os, estarão no mesmo *cluster* e os nós menos relacionados, com poucas arestas entre eles, pertencerão a diferentes clusters (SCHAEFFER, 2009).

O resultado do agrupamento de um grafo depende da medida de similaridade usada e do algoritmo. Como em outros campos da IA, muitos métodos foram desenvolvidos nos últimos anos.

As aplicações mais comuns de análise de grafos são de redes biológicas, redes sociais e redes de transporte. Mas grafos também são empregados para analisar comportamentos de consumidores das áreas de *marketing* e vendas, especialmente de vendas *online*, cuja quantidade de dados é vasta.

Estes dados e suas relações podem ser utilizados para recomendar produtos a clientes. Esse tipo de tarefa é conhecida como sistema de recomendação.

Na pandemia, devido ao distanciamento social e as restrições do comércio, as vendas online cresceram 75% (VILELA, 2021). Sem os vendedores, é preciso que o próprio *site* faça recomendações aos clientes para que eles comprem outros produtos. Por isso, o objetivo deste trabalho é identificar os grupos de clientes com preferências semelhantes e, a partir da suas preferências, recomendar outros produtos que possam ser do seu interesse. Quanto mais refinado o sistema, maior é a conversão da sugestão em venda.

A filtragem colaborativa baseada no usuário é, provavelmente, a estratégia mais bem sucedida para a construção de sistemas de recomendação (BREESE *et al.*, 1998) (RESNICK *et al.*, 1994). No entanto, apesar de seu sucesso, há uma limitação importante quanto a sua escalabilidade

(SARWAR *et al.*, 2000) (SCHAFER *et al.*, 1999). Como alternativa, a recomendação baseada em item nomeado é proposta (DESHPAND, 2004) (SARWAR *et al.*, 2001). Nessas abordagem, a informação histórica é analisada para identificar as relações entre pares, de modo que a compra de um item muitas vezes leva ao compra de outro item semelhante.

Numa abordagem tradicional, Linden et. al (LINDEN *et al.*, 2003) estudou a recomendação de itens da Amazon de acordo com as avaliações atribuídas pelos usuários. Utilizando dados semelhantes, Wang (WANG *et al.*, 2006) utilizou grafos para encontrar as semelhanças e um algoritmo chamado *Item Rating Smoothness Maximization* (IRSM) para fazer as recomendações. Nos dois estudos, os autores consideram o método efetivo tanto em relação ao tempo computacional quanto às recomendações geradas.

Utilizando uma estratégia parecida, o objetivo deste trabalho é recomendar produtos de beleza da Amazon (JIANMO *et al.*, 2019) a partir das avaliações feitas pelos usuários. Para encontrar a semelhança entre os usuários e itens, será utilizada uma técnica de agrupamento de grafo. Também serão abordadas algumas propriedades, medidas de distâncias e métodos de avaliação de grafos, além da metodologia empregada e dos resultados obtidos. A metodologia poderá ser empregada para outros conjuntos de dados, visto que não requer muitos dados de entrada. Além disso, pequenos comerciantes - os mais afetados pela pandemia (RODRIGUES, 2021) - podem empregar a metodologia, que é de fácil implementação, para melhorar suas vendas.

2 REVISÃO DA LITERATURA

2.1 GRAFOS

Um grafo é composto por um par de conjuntos de nós (ou vértices, V) e arestas (E) e pode ser classificado de acordo com as arestas. Em um grafo não-direcionado, o caminho de um nó a outro pode ser percorrido em ambos os sentidos. Se as arestas são direcionadas, existe uma orientação específica para se passar de um vértice a outro. Em um grafo não ponderado, todas as arestas possuem o mesmo peso (geralmente assume-se o valor 1), enquanto em um grafo ponderado as arestas têm diferentes níveis de importância (pesos) e isso muda as relações e características do grafo (SCHAEFFER, 2009).

O comprimento de um caminho é o número de arestas visitados entre dois nós. A distância entre os nós é o caminho mais curto que os conecta. Um grafo é considerado conectado se todos os nós possam ser alcançados de outros e desconectado se houver pelo menos um que não possa.

A importânica desse ramo da matemática é evidenciada haja vista que muitas estruturas complexas podem ser representadas por grafos.

2.2 AGRUPAMENTO DE GRAFOS

Clustering é uma técnica de aprendizado de máquina não supervisionado, o que significa que funciona com instâncias não rotuladas. Neste caso, eles são agrupados com base em alguma medida de similaridade, que ajuda a entender os relacionamentos existentes em um conjunto de dados. Basicamente o agrupamento de grafos visa particionar um conjunto de grafos em grupos diferentes que compartilham alguma forma de similaridade (ERRICA *et al.*, 2020). Esses grupos, também conhecidos como *clusters*, podem ser chamados de comunidades (NEWMAN; GIRVAN, 2003) (GIRVAN; NEWMAN, 2002).

No entanto, nem sempre está claro se um nó deve ser atribuído a um cluster ou se ele poderia ter diferentes "níveis de associação" em vários *clusters*. Esta incerteza é comum em agrupamento de documentos, por exemplo. Para tarefas gerais de agrupamento, algoritmos de agrupamento *fuzzy* foram propostos, mas, diferentemente de outros campos, não há muitos trabalhos a serem encontrados (FORTUNATO, 2009).

2.3 PROPRIEDADES

A propriedade mais desejável é que cada *cluster* seja conectado intuitivamente, o que significa que haja pelo menos um - e de preferência muitos - caminhos conectando cada par de nós de um *cluster*. Se um nó não puder ser alcançado a partir de outro nó, eles não devem ser agrupados no mesmo *cluster*. Além disso, os caminhos mais curtos devem ser internos ao *cluster* (SCHAEFFER, 2009).

É geralmente aceito que um bom *cluster* é denso e tem relativamente poucas conexões dos nós incluídos para nós de outros *clusters* (do resto do grafo). A densidade interna de um bom *cluster* deve ser notavelmente maior do que a densidade do grafo e a densidade do *intercluster* deve ser menor do que a densidade do grafo (NEWMAN, 2004).

Outra medida que ajuda a avaliar a dispersão de conexões do *cluster* para o resto do grafo é o tamanho do corte. O tamanho do corte é o número de arestas que conectam os nós em um *cluster* aos nós no resto do grafo. Quanto menor o tamanho do corte, melhor "isolado" o *cluster* e, consequentemente, menor a densidade *intercluster* (SCHAEFFER, 2009).

2.4 FUNÇÕES PARA AGRUPAMENTO

Existem duas maneiras principais de determinar os clusters: calculando algumas semelhanças de nós ou otimizando algumas funções que representam características.

2.4.1 Medidas de similaridade

A similaridade é estimada calculando uma métrica de distância e pode ser dividida em três categorias: medidas baseadas na contagem de pares, correspondência de cluster e teoria da informação (SCHAEFFER, 2009).

A contagem de pares significa calcular o número de pares de vértices que são classificados nos mesmos *clusters* em dois agrupamentos. Um exemplo famoso é o índice Rand. Uma métrica comumente usada é a distância do cosseno. Para dados categóricos, o índice de Jaccard é apropriado. Para dados tipo texto, uma métrica típica é distância *edit*. Outra opção é usar índices padronizados como *z-score*, já que as magnitudes das pontuações podem dar uma percepção equivocada sobre a similaridade efetiva.

A correspondência de *cluster* visa estabelecer uma correspondência entre pares de

clusters de diferentes partições com base no tamanho de sua sobreposição - propriedades estruturais do grafo utilizando, por exemplo, medidas baseadas em adjacência. Uma medida popular é a fração de vértices detectados corretamente, introduzida por Girvan e Newman. Calcular o coeficiente de correlação de Pearson também é uma boa estratégia.

A similaridade também pode ser estimada computando, dado um *cluster*, a quantidade adicional de informação que se precisa ter para atribuir a instância a outro *cluster*. Se os *clusters* forem semelhantes, pouca informação será necessária para ir de um para o outro. A informação mútua normal tem sido usada regularmente para calcular a similaridade dos *clusters* na literatura. No entanto, a medida é sensível ao número de *clusters* detectados. Uma medida mais promissora é a variação da informação (SCHAEFFER, 2009). Outra maneira intuitiva é calcular o comprimento dos caminhos e estabelecer um limite.

2.4.2 Medidas baseadas em otimização

Nas otimizações, são definidas funções que avaliam a qualidade de um determinado *cluster*. Essas funções também podem avaliar e comparar algoritmos de agrupamento. A ideia é tentar identificar *clusters* que atendam a uma determinada propriedade desejável. Critérios usuais incluem variantes de medidas de densidade, medidas baseadas na fração ou no número de arestas presentes no *cluster*, entre outras (SCHAEFFER, 2009).

Alguns algoritmos procuram *clusters* com densidade acima de um limite, outros irão otimizar a independência (conectividade) do *cluster* com base no tamanho do corte e condutância mínima (KANNAN *et al.*, 2004).

2.5 AVALIAÇÃO DA QUALIDADE

Para comparar dois ou mais algoritmos de agrupamento ou mesmo aceitar um resultado como “bom”, algumas funções de qualidade são necessárias. Nesse sentido, propriedades de um bom agrupamento foram discutidas por Kleinberg (KLEINBERG, 2002). Dado um conjunto S de pontos, uma função de distância positiva definida e simétrica $d()$ é definida. O objetivo é encontrar um agrupamento com base nas distâncias entre os pontos com as seguintes propriedades:

1. Invariância de escala: multiplicar qualquer função de distância por uma dada constante resulta no mesmo agrupamento.

2. Riqueza: qualquer partição possível de um determinado conjunto de pontos pode ser recuperada se alguém escolher uma função de distância adequada. Em outras palavras, é a capacidade de produzir todos os agrupamentos ao escolher um conjunto de arestas apropriado.

3. Consistência: dado um agrupamento, qualquer modificação da função de distância que não diminua a distância entre pontos de diferentes *clusters* e que não aumente a distância entre pontos de um mesmo *cluster* resulta no mesmo agrupamento.

No entanto, Kleinberg provou o importante teorema da impossibilidade: dado um conjunto S de pontos, uma função de distância $d()$ é definida, que é definida positiva e simétrica, é impossível encontrar um agrupamento f com base nas distâncias entre os pontos que satisfaça ao mesmo tempo as três propriedades acima.

Além disso, é útil ter um critério quantitativo da qualidade de um agrupamento de um grafo. Uma função de qualidade é uma função que atribui um número a cada *cluster* de um grafo. O *cluster* com a maior pontuação é, por definição, o melhor.

Um exemplo de função de qualidade é o desempenho P, que conta o número de pares de vértices agrupados corretamente, dois vértices pertencentes à mesma comunidade e conectados por uma aresta, ou dois vértices pertencentes a comunidades diferentes e não conectados por uma aresta.

Outro exemplo é a cobertura, que é a razão do número de arestas intracomunitárias pelo número total de arestas: por definição, uma estrutura de *cluster* ideal, no qual os *clusters* são desconectados uns dos outros, resulta em uma cobertura de 1, já que todas as arestas do grafo caem dentro de aglomerados (FORTUNATO, 2009).

Ademais, a função de qualidade mais popular é a modularidade de Newman e Girvan (GIRVAN; NEWMAN, 2002). Ela é baseada na ideia de que não se espera que um grafo aleatório tenha uma estrutura de *cluster*. Então, comparando a densidade de enlace de uma comunidade com a densidade de enlace obtida para o mesmo grupo de nós para uma rede religada aleatoriamente, pode-se decidir se a comunidade original corresponde a um subgrafo denso, ou se seu padrão de conectividade surgiu por acaso. Para uma maior modularidade, o peso total das arestas do *intracluster* é grande e o peso total das arestas do *intercluster* é pequeno. É o equivalente teórico do grafo para minimizar a soma dos quadrados das distâncias dentro dos *clusters* e maximizá-la entre os *clusters*.

Por fim, a probabilidade de classificação integrada é uma medida que pode ser usada para escolher o número de *clusters* (BIERNACKI *et al.*, 2000). Assume-se que os dados seguem um modelo probabilístico denominado modelo de mistura finita. O objetivo é estimar a probabilidade de que uma determinada observação pertença a um determinado *cluster*.

2.6 MÉTODOS

Os métodos mais tradicionais são de agrupamento hierárquico (por divisão ou aglomeração), agrupamento por partição e agrupamento espectral. Também foram desenvolvidos algoritmos baseados na otimização da modularidade, algoritmos dinâmicos e baseados em inferência estatística (FORTUNATO, 2009) (SCHAEFFER, 2009) (PORTER *et al.*, 2009).

A maioria dos métodos considera a situação ideal e clássica: redes não direcionadas e não ponderadas. Embora algumas extensões tenham sido propostas, mais estudos são necessários para entender as arestas direcionadas e ponderadas, as redes bipartidas e os clusters sobrepostos (FORTUNATO, 2009) (PORTER *et al.*, 2009) (NETWORK. . . ,) (SCHAEFFER, 2009).

Muitos métodos foram desenvolvidos e aprimorados para se tornarem mais precisos e/ou mais rápidos. Porém, a maioria deles exige um conhecimento prévio da estrutura da rede que, normalmente, não está disponível.

2.7 SISTEMAS DE RECOMENDAÇÃO

A tarefa de recomendação é gerar uma lista com os itens com as melhores avaliações (SARWAR, 2001). Nos sistemas de recomendação, a utilidade de um item normalmente é representada por uma nota. Em bases de dados como as da Amazon (*All_Beauty*), os usuários podem classificar os produtos com notas de um (ruim) a cinco (bom).

A partir desta base de dados, é construída uma rede bipartida, em que uma partição são os usuários e na outra, os produtos. Não há ligação direta entre usuários ou entre produtos, apenas entre usuário e produto. Os usuários geralmente avaliam um subconjunto muito menor que o conjunto dos produtos disponíveis, por isso a rede não é completamente conectada.

Segundo Shoham e Balabanovic (BALABANOVIC; SHOHAM, 1997), a classificação de sistemas de recomendação é dividida em três categorias: filtragem por conteúdo, fitragem colaborativa e filtragem híbrida.

Este estudo se classifica como filtragem colaborativa, pois baseia-se na ideia de que

uma pessoa tende a aceitar a sugestão de um grupo de pessoas próximas, ou semelhantes a ela e recomenda produtos baseados na avaliação dessas pessoas similares (ADOMAVICIUS; TUZHILIN, 2005) (SARWAR, 2001). Em outras palavras, a ideia é usar a inteligência coletiva de um grupo de pessoas para fazer recomendações para outras (ALAG; MAXMANUS, 2009) (SEGARAN, 2008).

Entretanto, não é possível avaliar se a recomendação proposta resulta em ganhos de venda, pois ela não será aplicada pela marca.

3 MATERIAL E MÉTODOS

3.1 SOFTWARE

A implementação foi realizada no notebook Jupyter, usando Python 3. Os pacotes incluídos que foram explorados são os seguintes:

- **NetworkX:** Biblioteca Python usada para criar os grafos e calcular as semelhanças dos nós
- **Numpy, Random, and Intertools:** Biblioteca Python usada para operações matemáticas

3.2 CENÁRIOS

Foram testados quatro cenários diferentes para avaliar qual seria a metodologia que melhor desempenharia a recomendação e entender até que ponto é vantajoso fazer a limpeza da base de dados.

No Cenário 1, foram consideradas as duas primeiras etapas de limpeza detalhadas na subseção 3.3.1, o agrupamento dos items pelo algoritmo BRIM (seção 3.5) e a recomendação pelos vizinhos. No Cenário 2, foram consideradas as duas primeiras etapas de limpeza e a recomendação pelos vizinhos. No Cenário 3, foram aplicadas as três etapas de limpeza, o agrupamento e a recomendação dos vizinhos. No Cenário 4, também foram aplicadas todas as etapas de limpeza, porém testou-se apenas a recomendação dos vizinhos.

As características utilizadas na comparação dos resultados dos cenários foram:

- Número de avaliações
- Número de usuários
- Número de items
- Item mais avaliado e número de avaliações recebidas
- Número de *clusters*
- Qualidade do agrupamento (modularidade)

- Item mais recomendado e número de vezes que foi a primeira recomendação: cada par usuário-item do *dataset* foi submetido a função de recomendação e salvou-se o primeiro produto recomendado, depois computou-se qual produto foi recomendado mais vezes.

3.3 BASE DE DADOS

O conjunto de dados utilizado é uma versão atualizada do conjunto de dados de revisão da Amazon lançado em 2014. Ele inclui o código do produto, código do usuário e a nota atribuída. Foi utilizada a base de dados de beleza (*All_Beauty*), contendo 361605 votos de 324038 usuários e 32586 produtos.

Tabela 1 – Cinco primeiros registros do conjunto de dados utilizado.

Item	Usuário	Peso
014789302X	A26PO1B2Q2G1CS	1.0
1620213982	A1Z8A548Z31SUB	2.0
1620213982	A1Z7KJ7SBYTDA8	5.0
1620213982	A1T2B5PFIP9TY1	5.0
1620213982	AP3KNPYPC9WSH	4.0

Fonte: Autoria própria.

3.3.1 Limpeza da base de dados

Foram aplicadas até três etapas de limpeza. A primeira consistiu em remover notas duplicadas do mesmo usuário para um mesmo produto, mantendo a última nota registrada. Depois, retirou-se os usuários - e suas notas para os produtos - que tinham avaliado apenas um produto. Por fim, tirou-se os usuários que deram nota máxima em todas suas avaliações (eterno feliz), pois considerou-se que não seriam tão confiáveis.

3.4 CONSTRUÇÃO DA REDE BIPARTIDA

A rede foi construída com a biblioteca Networkx. Para cada linha do conjunto de dados (que contém informações de produto, usuário e nota), adicionava-se dois nós e uma aresta entre esses nós. Para cada nó adicionado era indicado se este pertencia ao grupo de usuários ou produtos. A nota daquele usuário para aquele produto em questão era atribuída como peso da arestra adicionada.

3.5 ALGORITMO DE AGRUPAMENTO

Para identificar os clusters, foi utilizado um algoritmo específico para redes ponderadas bipartidas, BRIM, proposto por Barber (BARBER, 2007).

O algoritmo é baseado na ideia de que as partições nas duas partes da rede são dependentes, com cada parte sendo mutuamente usada para induzir os vértices da outra parte nos módulos. Este é um método de otimização de modularidade. No entanto, é calculada uma extensão da modularidade de Newman e Girvan, uma matriz de modularidade bipartida que enfatiza a decomposição de valor singular em vez da decomposição espectral.

Essa extensão altera o modelo nulo, de forma com que os graus esperados coincidam com os graus na rede real e com a restrição adicional de que cada aresta liga dois conjuntos separados.

Então, seja p o número de vértices de usuários e q o número de vértices de produtos; isso implica $n = p + q$. Sem perda de generalidade, suponha que os vértices sejam indexados de modo que os vértices de usuários são identificados como $1, 2, \ldots, p$ e os vértices de produtos são rotulados como $p + 1, p + 2, \cdots, p + q$. A matriz de adjacência então tem uma forma de bloco fora da diagonal de:

$$A = \begin{bmatrix} 0_{p\times p} & \tilde{A}_{p\times q} \\ \tilde{A}^{T}_{q\times p} & 0_{q\times q} \end{bmatrix} \tag{1}$$

na qual $0_{i\times j}$ é a matriz totalmente nula com i linhas e j colunas e $\tilde{A}$ é a matriz dos graus dos nós de usuários. É requerida a mesma estrutura de bloco para P que é exibido por A, dando:

$$P = \begin{bmatrix} 0_{p\times p} & \tilde{P}_{p\times q} \\ \tilde{P}^{T}_{q\times p} & 0_{q\times q} \end{bmatrix} \tag{2}$$

em que $\tilde{P}$ a matriz dos graus esperados dos nós dos usuários e P é a matriz adjacência do modelo nulo. A probabilidade de uma aresta estar presente entre dois vértices é proporcional ao produto dos graus dos vértices. Para a rede bipartida, isso se torna $\tilde{P}_{i,j} = Gkidj$ para alguma constante G.

Esta forma para P atribui probabilidade zero às arestas entre vértices do mesmo grupo (usuários ou produtos), impedindo tais arestas no modelo nulo. A matriz de modularidade B, por

sua vez, tem uma forma de bloco fora da diagonal de:

$$B = \begin{bmatrix} 0_{p\times p} & \tilde{B}_{p\times q} \\ \tilde{B}^{T}_{q\times p} & 0_{q\times q} \end{bmatrix} \tag{3}$$

na qual $\tilde{B} = \tilde{A} - \tilde{P}$. Os blocos totalmente nulos na diagonal são as contribuições potenciais de modularidade dos pares de vértices de mesmo grupo presentes em um módulo; todas as contribuições significativas, positivas ou negativas, para a modularidade, portanto, são feitas por pares de vértices de grupos distintos. Assim, o número esperado de arestas no modelo nulo deve ser igual ao número de arestas na rede real.

Com isso, a modularitidade de redes bipartidas proposta por Barber é:

$$Q = \frac{1}{2m}\sum_{i,j}(A_{ij} - P_{ij})\delta(g_i,g_j) \tag{4}$$

Considerando que c é o número de comunidades identificadas em uma rede e a matriz de índice de uma divisão de comunidade, S, tem dimensões de nxc com cada linha como um vetor de índice de (0,1) elemento correspondente a um nó, tal que 1 se o nó i pertence à comunidade j e 0 caso contrário. Em redes bipartidas, S pode ser particionado para que:

$$S = \begin{bmatrix} R_{p\times c} \\ T_{p\times c} \end{bmatrix} \tag{5}$$

sendo que R e T são matrizes de índice para nós de usuários e os nós de produtos, respectivamente. Uma expressão equivalente da modularidade é:

$$Q = \frac{1}{m}\sum_{i=1}^{p}\left[\sum_{k=1}^{c} R_{ik}(\tilde{B}T)_{ik}\right] \tag{6}$$

Uma vez que cada linha de R consiste em um único 1 com todos os outros elementos sendo 0, maximizar Q se torna simples: só é preciso atribuir o nó do usuário i à comunidade k, como $(\tilde{B}T)_{ik}$ é o máximo da i-ésima linha de $\tilde{B}T$. $\tilde{B}T$ é completamente determinada pela divisão de nós de produtos e, portanto, é simples induzir a divisão de nós de usuários a partir da divisão de nós de produtos, com o objetivo de otimizar a modularidade bipartida. Da mesma forma, pode-se deduzir:

$$Q = \frac{1}{m}\sum_{j=1}^{p}\left[\sum_{k=1}^{c} T_{jk}(\tilde{B}^{T}R)_{jk}\right] \tag{7}$$

que indica como induzir a divisão de nós de produtos a partir da divisão de nós de usuário. Juntas, essas duas precedências dão origem ao algoritmo BRIM: partindo de uma divisão de nós de usuário (também pode começar de uma divisão de nós de produto; a divisão inicial é chamada de divisão inicial), é induzida a divisão de nós de produto. A partir da divisão dos nós do produto, é induzida a divisão do nó do usuário e o algoritmo segue iterando. Desta forma, a modularidade bipartida aumenta até que um máximo local seja alcançado e, consequentemente, uma divisão da comunidade seja gerada (LIU; MURATA, 2010).

3.6 SISTEMA DE RECOMENDAÇÃO

Na comunidade de física estatística, a abordagem usual adotada para identificar *clusters* em redes bipartidas é primeiro construir uma projeção unipartida de uma parte da rede e, em seguida, identificar os módulos nessa projeção usando métodos para redes unipartidas (BARBER, 2007).

A função que faz a recomendação recebe a rede bipartida, os nós de itens, o usuário em questão, o item que foi avaliado e os *dataframes* contendo a indicação dos *clusters* dos nós dos usuários e dos itens. As seguintes etapas são realizadas:

1. Checa-se se o usuário e o produto passado estão conectados, ou seja, se o usuário já avaliou aquele determinado produto.

2. Constrói-se a projeção ponderada unipartida dos produtos e encontra-se os vizinhos do determinado produto. O grafo projetado ponderado é a projeção da rede bipartida, nos nós especificados com pesos representando o número de vizinhos compartilhados. Os nós retêm seus atributos e estão conectados no grafo resultante se eles têm uma aresta para um nó comum no grafo original (BORGATTI; HALGIN, 2014).

3. Verifica-se se algum dos vizinhos já foi avaliado pelo usuário e o retira da lista de recomendações.

4. Classifica-se a recomendação de acordo com o número de conexões compartilhadas.

5. Verifica-se o *cluster* do produto e do usuário obtido pelo agrupamento pelo algoritmo BRIM.

6. Verifica-se o *cluster* dos vizinhos é o mesmo do produto ou do usuário. Atribui-se 1 caso positivo e 0 caso negativo.

7. Classifica-se a recomendação de acordo com o pertencimento do mesmo *cluster* e pelo número de conexões compartilhadas e retorna uma lista de outros produtos que não foram comprados pelo usuário, mas foram conectados ao produto por meio de outras pessoas. Essa lista é classificada com base no número de pessoas compartilhadas (do maior para o menor) e se os produtos estão no mesmo cluster (similar). A classificação ajudará na recomendação.

A função retorna duas listas, a primeira que considera a classificação de acordo com o agrupamento e o número de conexões e a segunda que considera apenas o número de conexões.

3.7 ESQUEMA

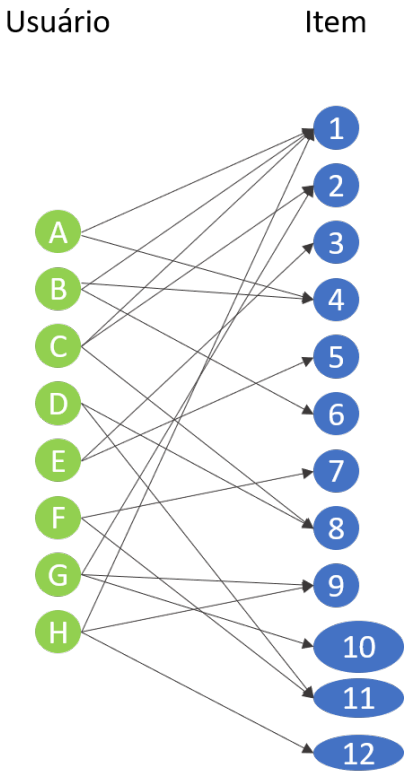
Por exemplo, caso fosse sugerido apenas o item mais avaliado, de acordo com a rede da Figura 1, não importa qual fosse o usuário e se ele já avaliou ou não aquele produto, o produto sugerido seria o produto 1, que recebeu quatro avaliações.

Entretando, a metodologia proposta a ser validada é o agrupamento e a recomendação segundo a semelhança. A primeira etapa é a clusterização segundo o algoritmo de BRIM. Na Figura 2, é possível identificar o agrupamento segundo a cores das bordas dos círculos.

Se o objetivo fosse sugerir um item ao usuário “A”. Os itens que não são do mesmo cluster, seriam ignorados (em cinza na Figura 3), bem como aqueles já avaliados por ele. Na sequência, os vizinhos mais próximos seriam procurados, ou seja, usuários que avaliaram os mesmos itens que “A” de forma similar e quais são os outros itens que esses usuários avaliariam.

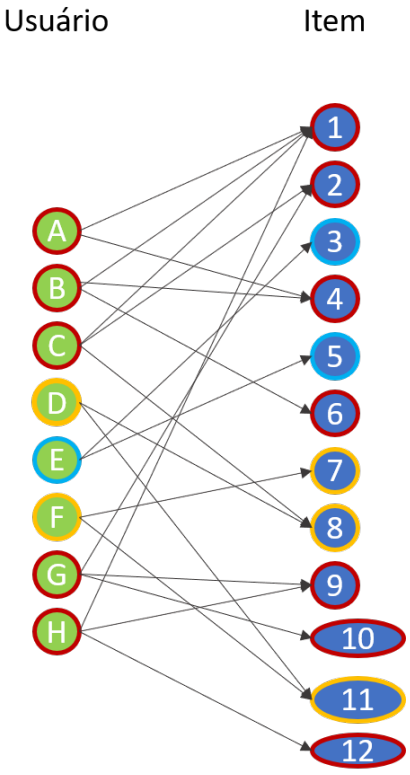
Desta forma, o item 6 seria o mais recomendado para o usuário “A”, pois foi avaliado pelo usuário “B” que também avaliou os dois produtos avaliados por “A”. Em seguida, viriam os itens 2, 9 e 12 que foram avaliados pelos usuários “C” e “H”, que assim como “A” avaliaram o item 1.

Figura 1 – Representação de uma rede bipartida.



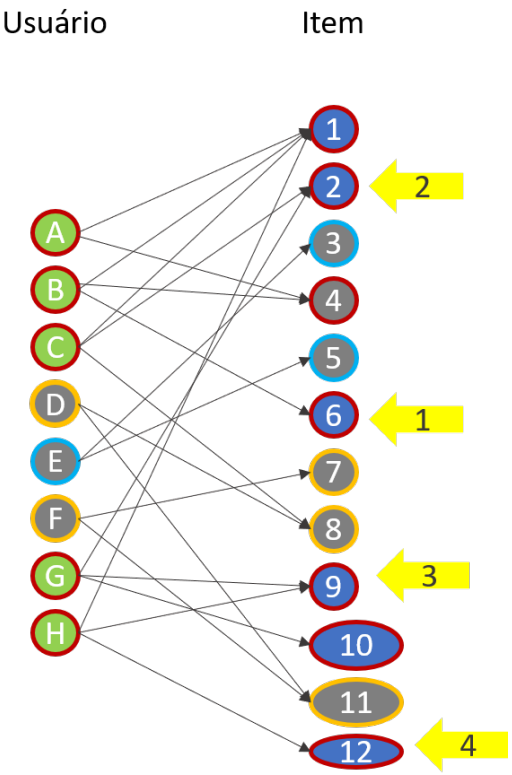
Fonte: Autoria própria

Figura 2 – Representação do agrupamendo da rede bipartida.



Fonte: Autoria própria.

Figura 3 – Representação da ordem de recomendação dos itens para o usuário "A".



Fonte: Autoria própria.

4 RESULTADOS E DISCUSSÃO

Na Tabela 2, estão resumidos os resultados de cada cenário testado.

Tabela 2 – Resultados obtidos para os quatro cenários propostos.

	Cenário 1	Cenário 2	Cenário 3	Cenário 4
Avaliações	67205		31176	
Usuários	11055		7728	
Itens	29638		13621	
Item mais avaliado, número de avaliações	B000FOI48G, 8302		B000FOI48G, 2878	
Limpeza	Remoção de duplicatas e dos usuários que avaliaram apenas uma vez		Remoção de duplicatas, dos usuários que avaliaram apenas uma vez e dos usuários que deram nota máxima em todas as avaliações	
Agrupamento	Sim	Não	Sim	Não
Clusters	825	1	672	1
Modularidade	0.817	-	0.843	-
Item mais recomendado, número de avaliações	B00W259T7G, 18311	B0091V8B5A, 16411	B00W259T7G, 6900	B0091V8B5A, 5672

Fonte: Autoria própria.

A primeira coisa que se nota é que o sistema de recomendação, seja ele com ou sem agrupamento, tem efeito sobre os itens que serão recomendados ao usuário, visto que para todos os cenários o item mais recomendado é diferente do item mais vezes avaliado. Além disso, se fosse sugerir os produtos mais votados, a recomendação seria sempre a mesma, ou seja, não seria personalizada ao usuário e não, necessariamente, seria algo de seu interesse.

Em seguida, percebe-se que tanto o produto mais vezes avaliado quanto os produtos mais recomendados são os mesmos independentemente do tamanho do *dataset* (ou número de avaliações). Logo, é preferível trabalhar com uma base menor, pois se reduz a complexidade do problema, tempo e custo computacional. Além disso, é possível notar que a qualidade do agrupamento melhora com a redução da base, visto que há aumento da modularidade.

Na Tabela 3 é possível identificar a posição no *ranking* dos items mais recomendados segundo os cenários 3 e 4 de acordo com a limpeza da base de dados. Em geral, não foi aplicada nenhuma limpeza, no Cenário 1 foram removidas duplicatas e usuários que avaliaram apenas um produto e, no Cenário 3, além dessas duas etapas de limpeza, também foram removidos os usuários que deram nota máxima para todos os itens que avaliaram.

Tabela 3 – Posição dos itens mais recomendados em relação aos items mais votados de cada cenário.

	Mais avaliado	posição do B00W259T7G nos mais votados	posição do B0091V8B5A nos mais votados
Geral	B000FOI48G, 8656	7	13032
Cenário 1	B000FOI48G, 8302	10	10032
Cenário 3	B000FOI48G, 2878	3	5120

Fonte: Autoria própria.

Percebe-se que o produto mais recomendado pelo cenário 4, o “B0091V8B5A”, é um item que foi pouco avaliado. Já o produto mais recomendado pelo cenário 3, o “B00W259T7G”, foi bastante avaliado e tem mais conexões entre os usuários. Portanto, a metodologia aplicada no cenário 3 tende a ser mais confiável e dar melhores resultados (recomendação que vira compra efetiva).

Entretanto, não é possível afirmar qual metodologia - com ou sem agrupamento - fornece melhores resultados. Para isso, seria preciso testar ambas e analisar qual delas resultou numa maior taxa de conversão de compra/recomendação.

Como a proposta do trabalho é utilizar o agrupamento de redes bipartidas para a recomendação de produtos e, segundo os resultados da Tabela 3 é a metodologia mais confiável, os resultados e análises detalhadas abaixo são provenientes do Cenário 3.

4.1 LIMPEZA DE DADOS

Após a limpeza, o conjunto de dados passou a conter 31176 votos, de 13621 usuários e 7728 produtos. Tal diminuição de 91,4% em seu tamanho facilita o processamento, ao custo de um decréscimo de 76,3% de produtos e de 95,8% de usuários englobados.

Isso significa que usuários novos ou que não avaliam (ou avaliaram apenas uma vez) os produtos que compraram não recebem recomendações de outros produtos porque não tem nenhuma conexão. Além disso, produtos recém adicionados no sistema ou de pouca saída não serão recomendados aos clientes.

4.2 AGRUPAMENTO

A rede foi construída em 6.14 segundos e a estrutura comunidade inicial em 0.08 segundos. A modularidade do agrupamento inicial foi de 0.839, um valor alto que indica boa

separação entre os *clusters*.

As matrizes foram computadas em 1.55 segundos e a otimização demorou 75.65 segundos. A modularidade optimizada foi de 0.843, com um ganho de apenas 0.5%.

Como resultado, foram encontrados 672 clusters.

Apesar de computados, os resultados de tempo não podem ser tratados como uma avaliação de performance dos algoritmos, pois foi executado em um ambiente em nuvem, tendo a influência da própria rede, de flutuações em *threads* e processos.

4.3 SISTEMA DE RECOMENDAÇÃO

Como exemplo de recomendação, escolheu-se três pares de usuário e item. O primeiro foi o usuário "UA2GJX2KCUSR0EI" e o item "IB001F51RAG". Nesse caso, de todos os produtos, o produto sugerido pelo sistema seria o produto que recebesse a melhor avaliação dos usuários que, como nosso cliente escolhido, adquiriram o produto com o referido ID e que estivesse no mesmo cluster que o usuário ou o produto avaliado.

Foram encontrados 71 vizinhos e os primeiros cinco produtos recomendados podem ser vistos na Tabela 4. Entre parênteses, são indicadas as posições do item quanto à limpeza geral, limpeza do cenário 1 e limpeza do cenário 3 respectivamente.

Tabela 4 – Cinco primeiras recomendações para o usuário *UA2GJX2KCUSR0EI* baseado no produto *IB001F51RAG*.

Item	Peso	Mesmo cluster
IB00W259T7G (7,10,3)	8	1
IB005IHT94S (38, 65, 74)	2	1
IB002GP80EU (445, 52, 34)	2	1
IB016V8YWBC (132,94,79)	1	1
IB000NKJIXM (1468, 265, 138)	1	1

Fonte: Autoria própria.

Para comparação, a Tabela 5 indica quais seriam os produtos sugeridos considerando o sistema de recomendação proposto ou os produtos mais avaliados de acordo com cada etapa de limpeza (Geral, Cenário 1 e Cenário 2).

O segundo usuário escolhido foi o "UAENH50GW3OKDA" e o produto "IB00OR1D37A". Foram encontrados 11 vizinhos e os primeiros cinco produtos recomendados podem ser vistos na Tabela 6.

Tabela 5 – Cinco primeiras sugestões para o usuário *UA2GJX2KCUSR0EI* baseado no produto *IB001F51RAG* segundo a recomendação proposta e os produtos mais votados.

Recomendado	Mais avaliado		
	Geral	Cenário 1	Cenário 3
IB00W259T7G	B000FOI48G	B000FOI48G	B000FOI48G
IB005IHT94S	B000GLRREU	B000GLRREU	B000GLRREU
IB002GP80EU	1620213982	B0012Y0ZG2	B00W259T7G
IB016V8YWBC	B001QY8QXM	B000URXP6E	B00021DJ32
IB000NKJIXM	B01DKQAXC0	B006WYJM8Y	B019809F9Y

Fonte: Autoria própria.

Tabela 6 – Cinco primeiras recomendações para o usuário *UAENH50GW3OKDA* baseado no produto *IB00OR1D37A*.

Item	Peso	Mesmo cluster
IB00OR1OFA4 (2781,1839,2043)	2	1
IB00EVO39EA (16412,8210,5266)	1	1
IB01D5LPRC6 (5717, 3350,4098)	1	1
IB00SHV7ABQ (4259,1942,2375)	1	1
IB00YSZSM80 (3812,9752,3609)	1	1

Fonte: Autoria própria.

Para comparação, a Tabela 7 indica quais seriam os produtos sugeridos considerando o sistema de recomendação proposto ou os produtos mais avaliados de acordo com cada etapa de limpeza (Geral, Cenário 1 e Cenário 2).

Tabela 7 – Cinco primeiras sugestões para o usuário *UAENH50GW3OKDA* baseado no produto *IB00OR1D37A* segundo a recomendação proposta e os produtos mais votados.

	Mais avaliado		
Recomendado	Geral	Cenário 1	Cenário 3
IB00OR1OFA4	B000FOI48G	B000FOI48G	B000FOI48G
IB00EVO39EA	B000GLRREU	B000GLRREU	B000GLRREU
IB01D5LPRC6	1620213982	B0012Y0ZG2	B00W259T7G
IB00SHV7ABQ	B001QY8QXM	B000URXP6E	B00021DJ32
IB00YSZSM80	B01DKQAXC0	B006WYJM8Y	B019809F9Y

Fonte: Autoria própria.

Por fim, o terceiro usuário escolhido foi o “UA1KSC91G9AIY2Z” e o produto “IB00NR90T0W”. Foram encontrados 24 vizinhos e os primeiros cinco produtos recomendados podem ser vistos na Tabela 8.

Tabela 8 – Cinco primeiras recomendações para o usuário *UA1KSC91G9AIY2Z* baseado no produto *IB00NR90T0W*.

Item	Peso	Mesmo cluster
IB0010ZBORW (51, 27, 9)	5	1
IB00W259T7G (7,10,3)	4	1
IB00CZH3K1C (1871,352,271)	1	1
IB00EF1QRMU (651,244,143)	1	1
IB019FWRG3C (181,21,10)	1	1

Fonte: Autoria própria.

Para comparação, a Tabela 9 indica quais seriam os produtos sugeridos considerando o sistema de recomendação proposto ou os protudos mais avaliados de acordo com cada etapa de limpeza (Geral, Cenário 1 e Cenário 2).

Tabela 9 – Cinco primeiras sugestões para o usuário *UA1KSC91G9AIY2Z* baseado no produto *IB00NR90T0W* segundo a recomendação proposta e os produtos mais votados.

	Mais avaliado		
Recomendado	Geral	Cenário 1	Cenário 3
IB0010ZBORW	B000FOI48G	B000FOI48G	B000FOI48G
IB00W259T7G	B000GLRREU	B000GLRREU	B000GLRREU
IB00CZH3K1C	1620213982	B0012Y0ZG2	B00W259T7G
IB00EF1QRMU	B001QY8QXM	B000URXP6E	B00021DJ32
IB019FWRG3C	B01DKQAXC0	B006WYJM8Y	B019809F9Y

Fonte: Autoria própria.

As Tabelas 5, 7 e 9 comprovam que, ao sugerir apenas produtos mais avalidos (sem considerar peso ou similaridade entre usuários), todos os usuários recebem a mesma sugestão. Ou seja, não é personalizado.

Por outro lado, ao comparar as recomendações feitas pelo sistema para cada usuário, percebe-se que eles são diferentes e individualizados. Isso porque são baseadas nas similaridades entre avaliações dos usuários, aumentando a chance de que o usuário se interesse pela sugestão e efetive a compra.

5 CONCLUSÕES E PERSPECTIVAS

Para a área de vendas, a tarefa de agrupamento de grafos pode ser uma boa aliada para fazer recomendações de novos produtos.

Aplicação do algoritmo BRIM apresentou bom resultado de agrupamento (modularidade acima de 0.8). Porém, o tempo requerido inviabiliza seu uso para mega lojas, visto que o volume de vendas e avaliações é muito grande e o agrupamento teria que ser refeito com uma frequência elevada.

Porém, para pequenos negócios, o tempo é insignificante. Além disso, por requerer poucos dados (apenas a avaliação do usuário ao produto) e por ser de fácil implementação, pode contribuir para o aumento de vendas e evitar o fechamento do comércio.

Outra limitação é que a recomendação só é realizada pra usuários que fizeram mais de uma avaliação. Por isso, muitos produtos nem entram na base de avaliação devido à limpeza e, por isso, o sistema acaba não recomendando produtos pouco comprados ou recém adicionados.

Uma alternativa é fazer uma boa campanha de marketing para novos produtos, colocá-los em destaque no site e dar incentivos aos clientes para que avaliem os produtos (especialmente os novos), como descontos ou brindes.

Ainda assim, visto que a maioria dos micro e pequenos negócios não tem nenhum sistema de recomendação e que representam 99% das empresas brasileiras (GOVERNO..., 2020), esta melhoria ainda é válida.

REFERÊNCIAS

ADOMAVICIUS, G; TUZHILIN, A. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. **IEEE Transactions on Education**, v. 17, n. 6, p. 734–749, 2005.

ALAG, S.; MAXMANUS, R. Collective intelligence in action. **Manning New York**, 2009.

BALABANOVIC, M.; SHOHAM, Y. Fab: Content-based, collaborative recommendation. **Communications of the ACM**, v. 40, n. 4, p. 66–72, 1997.

BARBER, M.J. Modularity and community detection in bipartide network. **Phys. Rev.**, v. 76, p. 1–9, 2007.

BIERNACKI, C.; CELEUX, G.; GOVAERT, G. Assessing a mixture model for clustering with the integrated completed likehood. **IEEE Trans. Pattern Anal. Mach. Intell.**, p. 719–725, 2000.

BORGATTI, S. P.; HALGIN, D. S. The sage handbook of social network analysis. *In*: ____. *[S.l.]*: Sage Publications, 2014. cap. 28.

BREESE, J. S.; HECKERMAN, D.; KADIE, C. Empirical analysis of predictive algorithms for collaborative filtering. **UAI**, 1998.

DESHPAND, G. Karypis M. Item-based top-n recommendation algorithms. **ACM Trans. On Information Systems**, 2004.

ERRICA, Davide Bacciua Federico; MICHELI, Alessio; PODDA, Marco. A gentle introduction do deep learning for graphs. v. 46, n. 1, p. 5, 2020.

FORTUNATO, Santo. Community detection in graphs. **Physics Reports**, p. 90–132, 2009.

GIRVAN, M.; NEWMAN, M.E.J. Community structure in social and biological networks. **Proccedings of the National Academy of Sciences**, p. 8271–8276, 2002.

GOVERNO destaca papel da Micro e Pequena Empresa para a economia do país. 2020. Ministério da Economia, Governo Federal. Disponível em: https://www.gov.br/economia/pt-br/assuntos/noticias/2020/outubro/governo-destaca-papel-da-micro-e-pequena-empresa-para-a-economia-do-pais#:~:text=Governo\%20destaca\%20papel\%20da\%20Micro\%20e\%20Pequena\%20Empresa\

%20para\%20a\%20economia\%20do\%20pa\%C3\%ADs,-Empreendimentos\
%20representam\%2099&text=Juntas\%2C\%20elas\%20representam\%2099\%25\
%20dos,dos\%20empregos\%20gerados\%20no\%20Brasil.

JIANMO, Ni; JIACHENG, Li; JULIAN, McAuley. **Justifying recommendations using distantly-labeled reviews and fined-grained aspects**. 2019. Empirical Methods in Natural Language Processing (EMNLP).

KANNAN, R.; VEMPALA, S.; VETTA, A. On clusterings - good, bad and spectral. **Journal of AXM**, p. 497–515, 2004.

KLEINBERG, J. An impossibility theorem for clustering. **MIT Press**, 2002.

LINDEN, G.; SMITH, B.; YORK, J. Amazon.com recommendations: Item-to-titem collaborative filtering. **IEEE Internet Computing**, 2003.

LIU, X.; MURATA, T. **Community Detection in Large-scale Bipartide Networks**. 2010.

NETWORK Science by Albert-László Barabási. Disponível em: http://networksciencebook.com/.

NEWMAN, M.E.J. Detecting community structure in networks. **The European Physical Journal**, p. 321–330, 2004.

NEWMAN, M.E.J.; GIRVAN, M. Mixing patterns and community structure in networks. **Notes in Physics**, 2003.

PORTER, Mason A.; ONNELA, Jukka-Pekka; MUCHA, Peter J. Communities in networks. **Notices of the AMS**, p. 1086–1095, 2009.

RESNICK, P.; IACOVOU, N.; SUSHAK, P. Bergstrom M.; RIEDL, J. Grouplens: An open architcture for collaborative filtering of netnews. **SIGCSCW**, 1994.

RODRIGUES, Léo. **CNC aponta fechamento de 75 mil lojas em 2020**. 2021. Agência Brasil. Disponível em: https://agenciabrasil.ebc.com.br/economia/noticia/2021-03/cnc-aponta-fechamento-de-75-mil-lojas-em-2020.

SARWAR, B. Item-based collaborative filtering recommendation algorithms. **Proceedings of the 10th international conference on World Wide Web**, p. 285–295, 2001.

SARWAR, B. M.; KARYPIS, G.; KONSTAN, J. A.; RIEDL, J. Application of dimensionality reduction in recommender systems - a case study. **ACM WebKDD**, 2000.

SARWAR, B. M.; KARYPIS, G.; KONSTAN, J. A.; RIEDL, J. Item-based collaborative filtering recommendation algorithms. **WWW**, 2001.

SCHAEFFER, Satu Elisa. Graph clustering. **Computer Science review**, p. 27–57, 2009.

SCHAFER, J.; KONSTAN, J.; RIEDL, J. Recommender systems in e-commerce. **Proceedings of ACM E-Commerce**, 1999.

SEGARAN, T. **Programming collective intelligence: building smart web 2.0 applications**. 2008.

VILELA, Luiza. **E-commerce: o setor que cresceu 75% em meio à pandemia**. 2021. No Varejo Consumidor Moderno. Disponível em: https://www.consumidormoderno.com.br/2021/02/19/e-commerce-setor-cresceu-75-crise-coronavirus/.

WANG, F.; MA, S.; YANG, L.; LI, T. Recommendation on item graphs. **ICDM**, 2006.