Submission date: 26/02/2019
Resubmission date: 07/06/2019
Resubmission date: 28/07/2019
Camera ready submission: 19/08/2019

Section: regular article

1st round notification: 09/05/2019
2nd round notification: 01/07/2019
3rd round notification: 07/08/2019
Available online: 21/08/2019
Publication date: 23/12/2019

Distância de Barbieri: uma métrica para identificar similaridade entre perfis de consumidores

Title: Barbieri Distance: a metric to identify similarity between consumer profiles

Luiz Eugênio Barbieri[1], Cristiano Roberto Cervi[1]

[1]Instituto de Ciências Exatas e Geociências – Universidade de Passo Fundo (UPF)
Passo Fundo, Rio Grande do Sul – Brasil

133780@upf.br, cervi@upf.br

Abstract. *This work is part of the study of recommender systems with content-based filtering, having as motivation the observation of user behavior in an Enterprise Resource Planning (ERP). The main contribution of the work is the development of Barbieri Distance, a metric whose purpose is to measure the similarity between buyers based on their purchase history. The metric is for situations where there is no buyer valuation data for the product purchased. Since it does not require ratings for items, because similarity happens when buyers buy too much or too little of the same product, it is possible to identify the similarity of the consumer profile based on their purchase history. In order to perform the metric validation experiments, a comparison method between buyer profiles is used, which presented satisfactory results in the calculation of similarity.*

Keywords. *Content-based filtering; Levenshtein distance; Recommender systems; Similarity.*

Resumo. *O presente trabalho insere-se nos estudos de sistemas de recomendação com filtragem baseada em conteúdo, tendo como motivação para seu desenvolvimento a observação do comportamento de usuários em um sistema de gestão empresarial. A principal contribuição do trabalho é o desenvolvimento da Distância de Barbieri, uma métrica cujo propósito é medir a similaridade entre compradores com base em seu histórico de compras. A métrica foi desenvolvida para situações nas quais não existem dados de avaliação do comprador para o produto adquirido. Como não necessita de ratings para os itens, pois a similaridade acontece quando compradores adquirem muito ou pouco de um mesmo produto, é possível identificar a similaridade do perfil de consumidores com base em seu histórico de compras. Para realizar os experimentos de validação da métrica, utilizou-se um método de comparação entre perfis de compradores, que apresentou resultados satisfatórios no cálculo da similaridade.*

Palavras-Chave. *Distância Levenshtein; Filtragem baseada em conteúdo; Similaridade; Sistema de recomendação.*

Cite as: *Barbieri, L. E. & Cervi, C. R. (2020). Barbieri Distance: a metric to identify similarity between consumer profiles (Distância de Barbieri: uma métrica para identificar similaridade entre perfis de consumidores). iSys: Revista Brasileira de Sistemas de Informação (Brazilian Journal of Information Systems), 13(1), 144-158.*

1. Introdução

A área de sistemas de recomendação tem avançado no contexto de recomendação de produtos e serviços. Tais sistemas são amplamente aplicados em muitos domínios, como bancos, celulares, músicas, livros e assim por diante [Yao et. al. 2019].

Um sistema de recomendação combina várias técnicas computacionais para selecionar itens personalizados com base nos interesses dos usuários e conforme o contexto no qual estão inseridos [Ricci et. al. 2011]. O tipo de sistema de recomendação baseado em conteúdo utiliza palavras-chave para descrever itens e construir um perfil de usuário com base nos itens que representam as preferências e os interesses do usuário. Nesta perspectiva, os algoritmos desenvolvidos para tais sistemas visam recomendar itens similares àqueles que determinado usuário gostou no passado ou está examinando no presente. Por meio da comparação entre itens previamente classificados, os itens com maior correlação são recomendados [Aggarwal 2016].

Neste trabalho é apresentada a Distância de Barbieri, um algoritmo que tem por objetivo identificar a similaridade entre compradores em cenários onde não há *ratings*. Aplicando a Distância de Barbieri, buscou-se identificar a similaridade entre perfis de compradores com base nas quantidades de produtos comprados pelos compradores experimentados. No decorrer deste artigo, será explicado o funcionamento do algoritmo, cujo surgimento se deu pela percepção de uma lacuna nestes cenários em que não se dispõe dos *ratings*. Com base em análises de dados adquiridos por meio do uso de sistema de gestão empresarial, identificou-se que neste contexto os compradores tendem a comprar de forma recorrente. Nas recorrências, identificou-se também que os compradores, frequentemente, repetem a compra de um determinado produto. Com a recompra de determinado produto nota-se a formação de um perfil de compra implícito.

Nas abordagens convencionais, como nos sistemas de recomendação da Amazon (www.amazon.com) e da Netflix (www.netflix.com), o modelo de perfil gerado é sobre um usuário que está utilizando o sistema. Um perfil de usuário consiste, principalmente, de conhecimento sobre as preferências individuais que determinam o comportamento do usuário [Barth 2010]. Diferentemente das abordagens convencionais, o presente trabalho visa identificar a similaridade de perfil entre compradores dentre um conjunto de compradores, tendo como fonte de dados o histórico de vendas proveniente do sistema de gestão empresarial do estabelecimento.

Na seção 2, serão discutidos conceitos e abordagens da literatura, os quais fundamentam este trabalho. Na seção 3, dar-se-á ênfase no detalhamento da abordagem proposta. Na seção 4 constam os experimentos e resultados obtidos, isto é, a fase de testes. Por fim, na seção 5 constam as considerações finais do trabalho.

2. Fundamentação Teórica

Os sistemas de recomendação, conforme definidos nos trabalhos de Herlocker (2000) e Vieira (2013), têm um foco específico: prever quais itens ou informações um usuário achará interessante ou útil, com a função de conduzir o usuário na escolha de um produto ou serviço ao disponibilizar sugestões personalizadas e individuais de acordo com os interesses do mesmo. Resnick e Varian (1997) e Santana (2018) caracterizam como típico o sistema de recomendação no qual as pessoas fornecem recomendações como entradas, que o sistema agrega e dirige aos destinatários apropriados.

Os sistemas de recomendação são utilizados em diversos contextos, como ressalta Maria (2017), podendo ser aplicados na venda de produtos, sugestões de serviços e/ou de pessoas, educação, entre outros. O assunto manifestou-se claramente como área de pesquisa independente em meio à década de 1990, momento em que os pesquisadores passaram a focar em problemas de recomendação que nitidamente se utilizavam de estruturas de avaliação (*ratings*). Este impulso na discussão se deu por meio do sistema de recomendação Tapestry [Goldberg et al. 1992] e do surgimento dos primeiros artigos sobre filtragem colaborativa, que é uma das técnicas utilizadas para realizar a recomendação.

Outra técnica é a filtragem baseada em conteúdo. O propósito deste método é recomendar itens que pertençam a um grupo de itens que o usuário esteve recentemente interessado. De acordo com Meurer (2014), essa filtragem analisa diferentes informações armazenadas sobre os itens com o objetivo de encontrar àqueles de particular interesse ao usuário. Podemos afirmar que o presente trabalho se insere nesta técnica, afinal, são usadas as últimas transações do histórico do perfil do usuário com a finalidade de abranger as mudanças comportamentais do usuário. Dentro de cada transação o grupo de itens que tiver mais acessos tem um peso maior que os restantes. É uma técnica de aprendizagem supervisionada, onde o perfil serve como dados de treino e os itens podem ser avaliados como relevantes ou não [Croft 2010] segundo uma medida de similaridade.

A modelagem de perfil é uma técnica utilizada para identificar preferências e interesses de um usuário com base em características significativas ou estereótipos que o insiram em um grupo que demonstra um comportamento similar em uma coleção de usuários [Plumbaum 2015]. Estas preferências são todas aquelas informações diretamente necessárias para a adaptação do comportamento do sistema aos interesses do usuário, possibilitando prever quais produtos o usuário poderia gostar.

O estudo sobre modelagem de perfil de usuário se deu no final da década de 1970 com o trabalho de Rich (1979). O trabalho descreve os problemas envolvidos na construção e exploração de modelos de usuários individuais, a fim de orientar o desempenho de um sistema interativo. Nele, um sistema chamado Grundy, que recomenda romances para pessoas, é descrito e analisado como um fórum no qual essas questões são exploradas.

Outro estudo que também trabalha a modelagem de perfil são os de Cervi, Galante e Oliveira (2013a; 2013b), no qual é abordada a questão de identificar o perfil de pesquisadores e medir sua reputação. O estudo resultou na definição de um modelo chamado Rep-Model e de uma métrica chamada Rep-Index. A métrica utiliza o modelo de perfil para identificar individualmente a reputação de pesquisadores.

Para gerar uma modelagem de perfil é preciso capturar características do usuário. A modelagem de perfil pode ser dividida em três tipos de detecção: Modelagem Explícita, na qual o usuário possui interação direta, como a captura de informações de questionários respondidos ou avaliações do usuário sobre produtos; Modelagem Implícita, que captura os dados e a forma de navegação do usuário para modelar seu perfil, que pode se dar com a observação do comportamento de um usuário, seja analisando o tempo que um usuário observa um produto ou a quantidade de vezes que ele interage com o mesmo [Barth 2010]; e Modelagem Híbrida, que utiliza as duas modelagens anteriores. Para que a recomendação seja eficaz e precisa, é importante ter uma boa medida de similaridade [Henriques 2016]. A similaridade, especialmente entre perfis de usuários, é essencial para

a geração de uma boa recomendação, pois as características de preferências e interesses dos usuários são levadas em consideração.

O coeficiente da correlação de Tanimoto (também conhecida como Similaridade de Jaccard) e a medida de similaridade Log-Likelihood podem ser utilizadas para identificar a similaridade de dois conjuntos de dados, sendo que a última se diferencia da primeira por calcular também o quão provável é a sobreposição dos elementos dos conjuntos [Vivian 2017]. Apesar de semelhantes à métrica proposta no presente trabalho, não se aplicam ao mesmo. A correlação de Tanimoto, bem como a similaridade Log-Likelihood apresentam seu resultado de similaridade buscando identificar a interseção dos elementos dos conjuntos comparados. Já no presente trabalho, busca-se identificar a similaridade do perfil do usuário comparando os elementos de mesma posição do vetor. Ou seja, dado dois conjuntos de valores com 6 elementos cada, serão comparados o elemento da primeira posição do primeiro vetor com o primeiro elemento do segundo vetor. Em seguida é comparado o segundo elemento do primeiro vetor com o segundo elemento do segundo vetor e assim consecutivamente até chegar ao último elemento. Desta forma é possível identificar o quão similar dois compradores foram ao realizar a compra dos mesmos produtos. Assim, enquanto as medidas de Tanimoto e Log-Likelihood apresentam maior similaridade quando encontram mais elementos iguais independente de sua posição, a Distância de Barbieri irá apresentar maior similaridade quando os elementos de mesma posição dos diferentes vetores forem mais próximos dentre o conjunto de elementos comparados.

Considerando modelagens que geram um perfil com base em palavras-chave, a identificação da similaridade entre esses perfis se dá pela comparação de quão próximas são as palavras-chave. A Distância de Levenshtein (1966) é uma das formas utilizadas para se verificar a similaridade entre palavras por meio do método de comparação por proximidade, ou seja, ela se dá pelo número mínimo de passos necessários para transformar uma palavra na outra. Para ser utilizado como índice de similaridade, é preciso dividir a distância de Levenshtein obtida pelo comprimento da maior palavra e depois subtraindo o resultado de um. O índice varia de zero a um, sendo zero quando for necessário substituir todas as letras da palavra e um para palavras iguais [Bastos 2009].

Em cenários nos quais se dispõe do histórico de compras dos compradores, nenhuma das abordagens sozinha é capaz de suprir a demanda deste cenário, pois observa-se que a filtragem colaborativa é insuficiente quando não se possuem dados de avaliação dos usuários, e a filtragem baseada em conteúdo necessita de palavras-chave para identificar similaridade entre perfis. Propõe-se uma abordagem que, baseada no conteúdo que é o histórico de compras, modele o perfil dos compradores identificando a similaridade entre os mesmos.

3. Distância de Barbieri

Contrapondo situações nas quais o perfil é formado por palavras-chave, neste trabalho o perfil foi modelado com base no histórico de compras do comprador. Cada comprador possui seu histórico de compras, dispondo da informação de qual produto e quantas unidades foram adquiridas pelo mesmo. As similaridades entre os perfis obtêm-se identificando os compradores que adquiriram quantidades similares dos mesmos produtos.

Observando que a Distância de Levenshtein identifica a similaridade entre palavras-chave com base no número mínimo de passos para igualar duas sentenças, aplicou-se a mesma lógica visando identificar o número mínimo de passos para igualar quantidades. Observa-se também que não é necessário que compradores tenham adquirido exatamente a mesma quantidade para torna-los similares, mas sim uma quantidade aproximada. Para identificar o quão aproximado são dois valores dentre um conjunto de valores, é gerado um vetor de cem posições, onde cada posição representa um conjunto de valores. Todas as posições possuem o mesmo alcance e compreende-se que dois valores são iguais dentre o conjunto (cem por cento de similaridade) quando estão na mesma posição do vetor. Assim, quanto mais próximas as posições de vetor dos valores comparados, mais similares eles são.

Visto que a Distância de Levenshtein somente é aplicável a palavras, manifestou-se a necessidade de uma nova abordagem que informa a distância entre dois valores. Para se obter a similaridade entre os valores origem $V(o)$ e valor alvo $V(a)$ dentre uma coleção de valores, foi proposta a

Equação 1, denominada Distância de Barbieri.

$$S = 1 - \frac{D}{\max(V_o, V_a)}$$

Equação 1. A similaridade entre dois valores é igual a normalização da distância entre os valores origem e alvo, dividida pelo valor máximo entre ambos e tendo seu resultado subtraído de 1 (um).

Primeiramente é preciso calcular a distância $D = |P(V_o) - P(V_a)|$ entre os valores, sendo esta o resultado absoluto da subtração entre a posição do valor origem $P(V_o)$ e a posição do valor alvo $P(V_a)$. Para encontrar a posição, é preciso antes gerar um arranjo de números reais $\mathbb{R}$, chamado de Torre. O tamanho t da torre $\mathbb{R}$ é o resto absoluto $\mathbb{Z}$ da subtração entre o valor máximo da coleção $V_{max} = max\,\mathbb{R}$ e o valor mínimo da coleção $V_{min} = min\,\mathbb{R}$, assim sendo representado por $t_{\mathbb{R}} = \mathbb{Z}(V_{max} - V_{min})$.

A torre tem seus elementos definidos pelo quociente resultante da divisão entre o resto do valor máximo subtraído pelo mínimo, dividido pelo tamanho da torre e somado pelo valor do índice anterior. Um elemento de posição $i > 0$ na torre tem seu valor representado por $\mathbb{R}_i = \mathbb{R}_{i-1} + \left(\frac{V_{max} - V_{min}}{t}\right)$, com a observação de que o elemento de posição 0 possui o valor mínimo $\mathbb{R}_0 = V_{min}$.

A posição de um valor na torre é o índice no qual o valor é maior que o valor do elemento que se está buscando saber a posição $V > \mathbb{R}_i$ e menor ou igual ao valor do elemento posterior ao elemento que se busca saber a posição $V \leq \mathbb{R}_{i+1}$, representado pela fórmula $P_v = \mathbb{R}_i < V \leq \mathbb{R}_{i+1}$. Assim, a distância é obtida através do resto absoluto entre a posição do valor origem na torre, subtraído pela posição do valor alvo na torre. O exemplo a seguir ilustra o funcionamento da fórmula. Dado um conjunto de elementos igual a

$$\mathbb{R} = \begin{bmatrix} 5{,}2 & 0{,}6 & 6{,}9 & 1{,}1 & 6{,}6 & 11{,}4 \\ 9{,}1 & 7 & 1{,}4 & 3{,}7 & 7 & 12 \end{bmatrix}$$

Identifica-se o $min\,\mathbb{R}$ como o menor valor

$min\ \mathbb{R} = 0{,}6$

e $max\ \mathbb{R}$ como o maior valor

$max\ \mathbb{R} = 12$

visto que o tamanho t é a diferença inteira da subtração do maior pelo menor

$t = \mathbb{Z}(max\ \mathbb{R} - min\ \mathbb{R})$

$t = \mathbb{Z}(12 - 0{,}6)$

$t = \mathbb{Z}(11{,}4)$

$t = 11$

Assim, um arranjo de 11 elementos é montado, com o elemento de índice 0 o valor igual a $min\ \mathbb{R}$ e seus elementos posteriores com valor representado pelo resultado da fórmula $\mathbb{R}_i = \mathbb{R}_{i-1} + \left(\frac{V_{max} - V_{min}}{t}\right)$.

Tendo o elemento de índice 0 o valor de $min\ \mathbb{R}$ que é 0,6, o elemento de índice $i = 1$ pode ser calculado como

$\mathbb{R}_i = \mathbb{R}_{i-1} + \left(\frac{V_{max} - V_{min}}{t}\right)$

$\mathbb{R}_1 = \mathbb{R}_{1-1} + \left(\frac{12 - 0{,}6}{11}\right)$

$\mathbb{R}_1 = \mathbb{R}_0 + \left(\frac{11{,}4}{11}\right)$

$\mathbb{R}_1 = 0{,}6 + (1{,}03)$

$\mathbb{R}_1 = 1{,}63$

Para calcular ao elemento de índice $i = 2$ aplica-se a mesma fórmula

$\mathbb{R}_2 = \mathbb{R}_{2-1} + \left(\frac{12 - 0{,}6}{11}\right)$

$\mathbb{R}_2 = \mathbb{R}_1 + \left(\frac{11{,}4}{11}\right)$

$\mathbb{R}_2 = 1{,}63 + 1{,}03$

$\mathbb{R}_2 = 2{,}66$

E assim sucessivamente até completar todos os elementos. Dessa forma, obtêm-se uma torre como a da Tabela 1.

Tabela 1. Tabela que representa o arranjo obtido aplicando a fórmula que gera a torre das distâncias entre os elementos.

Índice	0	1	2	3	4	5	6	7	8	9	10
Valor	0,6	1,63	2,66	3,69	4,72	5,75	6,78	7,81	8,84	9,87	10,9

A posição de qualquer valor do arranjo pode ser encontrada seguindo a fórmula $P(V) = f_i < V \leq f_{i+1}$. Assim, sabe-se que a posição do valor 5,2 no arranjo é 4, pois 5,2 está entre os valores 4,72 e 5,75, que pertencem respectivamente às posições 4 e 5. Considerando que se procura saber a posição de 5,2 atribuímos à V o valor de 5,2. Logo,

$P(5{,}2) = f_0 < 5{,}2 \leq f_1$

Substituindo f_0 pelo valor que representa o índice de 0 e f_1 pelo valor que representa o índice de 1, obtêm-se a seguinte afirmação

$P(5{,}2) = 0{,}6 < 5{,}2 \leq 1{,}63$

Vê-se que a posição do valor na torre não é 0, pois, 5,2 não é menor ou igual à 1,63, tornando a afirmação falsa. O teste pode ser repetido sucessivamente até chagarmos em $i = 4$, onde

$P(5{,}2) = 4{,}72 < 5{,}2 \leq 5{,}75$

Assim, para saber qual a similaridade entre 5,2 e 9,1 dentre o conjunto de elementos proposto aplicando a fórmula $S = 1 - \frac{D}{\max(V_o, V_a)}$ e considerando que $V_o = 5{,}2$ e $V_a = 9{,}1$, encontramos a distância com $D = |P(5{,}2) - P(9{,}1)|$

$D = |P(V_o) - P(V_a)|$

Sabemos que $P(5{,}2) = 4$, e que $P(9{,}1) = 8$, pois

$P(9{,}1) = f_8 < 9{,}1 \leq f_9$

Substitui-se $f_8 = 8{,}84$ e $f_9 = 9{,}87$, assim

$P(9{,}1) = 8{,}84 < 9{,}1 \leq 9{,}87$

Logo

$D = |4 - 8|$

$D = |-4|$

$D = 4$

Também se sabe que $\max(V_o, V_a) = 9{,}1$, pois, 9,1 é maior que 5,2, sendo assim é o valor máximo entre ambos. Seguindo a métrica

$S = 1 - \frac{4}{9{,}1}$

$S = 1 - 0{,}44$

$S = 0{,}56$

Ou seja, considerando o conjunto de elementos $\mathbb{R}$, a similaridade entre 5,2 e 9,1 é de 0,56 ou 56%.

4. Experimentos e Resultados

Esta seção apresenta os experimentos realizados e tem como objetivo avaliar a Distância de Barbieri para verificar se ela atende aos propósitos planejados. Para isso, apresenta-se uma introdução acerca do *baseline* utilizado, o detalhamento dos experimentos, assim como a análise dos resultados obtidos.

4.1. Base de Dados

Para realizar os experimentos, foram utilizados dados provenientes de empresas usuárias do ERP360, que é um sistema de gestão empresarial desenvolvido pela Ren9ve Softwares. O desenvolvimento da metodologia se deu em parceria com a Ren9ve

Softwares, que intermediou a liberação do uso dos dados para fins acadêmicos, os quais foram disponibilizados pelas empresas usuárias do sistema.

Os dados são provenientes de uma das empresas usuárias do ERP360. Foram escolhidos cinco compradores de forma aleatória na base de dados de determinada empresa, com a restrição de que o comprador houvesse comprado no máximo sete produtos distintos. A restrição de compras de produtos se deu para que os resultados pudessem ser transcritos em sua totalidade no presente trabalho. Coletou-se então o histórico de compras dos compradores selecionados, sendo que o histórico é composto pelo produto, quantidade de itens e data em que ocorreu a compra. O único tratamento realizado nos dados foi a substituição dos nomes dos compradores e dos produtos por nomes fictícios.

4.2. Abordagem de Uso da Distância de Barbieri

Para realizar os experimentos, desenvolveu-se um algoritmo na linguagem C#. Tal algoritmo recebe por parâmetro os dados, sendo estes a lista de compradores com seu histórico de compras, o comprador base que será utilizado como parâmetro para identificação da similaridade e um arranjo de dados que é obtido conforme descrito na seção 3, representando a torre utilizada na Distância de Barbieri.

Considerando que o histórico de compras possui uma quantidade volátil de elementos, que a Distância de Barbieri é aplicada para cada produto comprado por ambos compradores e assim é obtido como resultado uma lista de N similaridades, é preciso realizar uma média aritmética com a finalidade de obter um valor percentual único que represente a similaridade entre os compradores comparados, pois sem o valor resultante não é possível observar a similaridade geral entre os compradores.

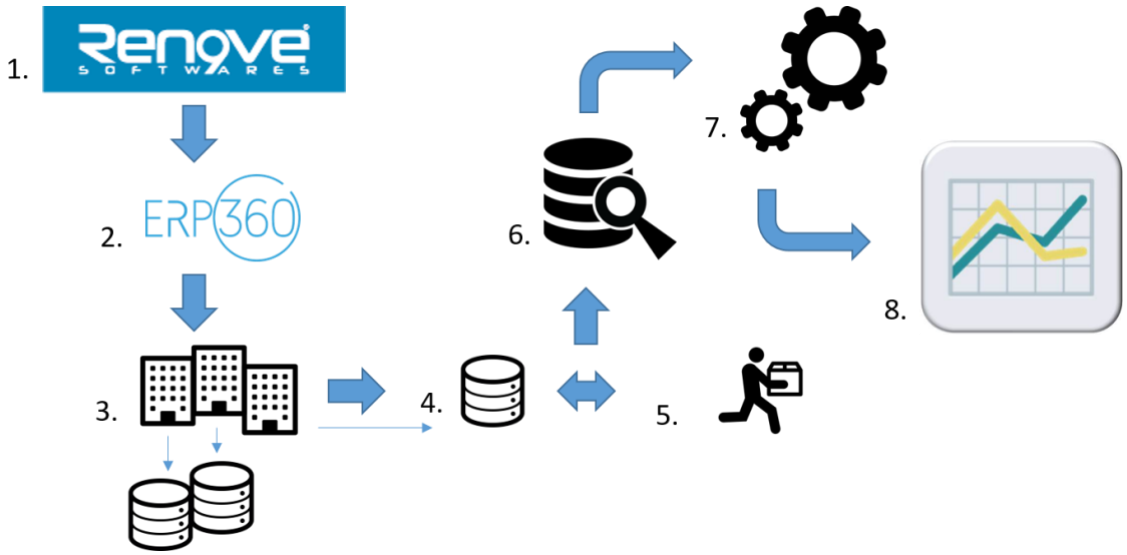

Figura 1. Fluxograma seguido nas experimentações.

Conforme é possível observar na Figura 1, a Ren9ve Softwares (1) é detentora do sistema de gestão empresarial ERP360 (2). Atualmente o sistema é utilizado por mais de setecentas empresas, sendo que cada empresa possui a sua base de dados (3). A base de dados (4) possui informações de todas as vendas realizadas pela empresa usuária (5). Nos

registros, encontram-se informações como os dados do comprador, produtos vendidos, datas das vendas, quantidades adquiridas, entre outras. Com uma base de dados em mãos (6), é realizada uma seleção aleatória de cinco compradores e coletado o histórico de compras dos mesmos. Na etapa (6) um tratamento nos dados abstrai informações sensíveis como o nome dos compradores e produtos, visando manter a privacidade das informações. Na sequência, os dados são enviados ao algoritmo (7) que faz o processamento destes dados, tendo como objetivo identificar a similaridade dos compradores com base em seu histórico de compras. Por fim, o algoritmo apresenta como retorno os resultados (8), sendo estes a representação da similaridade percentual entre os compradores comparados.

Para a realização dos experimentos, utilizou-se os dados disponibilizados e, mediante um conjunto de compradores, cada um com sua lista de compras e com respectivas quantidades, aplicou-se a métrica da Distância de Barbieri. O objetivo foi identificar o comprador com maior similaridade a dois usuários: Ana e Noemi. Para isso, duas situações distintas foram experimentadas. Na primeira situação, Experimentos 1 e 2, foi considerada a singularidade entre os compradores, ou seja, para identificar o V_{max} e o V_{min} e, consequentemente, o tamanho t do arranjo foram consideradas apenas as compras dos compradores comparados.. Na segunda situação, Experimentos 3 e 4, a singularidade é desconsiderada, ou seja, para identificar os valores V_{max} e V_{min}, toda a lista de compras é considerada e não mais apenas a lista de compras dos compradores comparados.

A seguir, na Tabela 2, apresenta-se os dados dos compradores e o histórico de compras de cada um deles.

Tabela 2. Lista de compradores e seus respectivos históricos de compras.

	Ana	**Noemi**	**Eugênio**	**Mauro**	**Larissa**
Produto A	3	369	3	13	318
Produto B	4	171	4	0	167
Produto C	7	437	8	16	354
Produto D	1	107	3	28	173
Produto E	6	159	5	2	0
Produto F	31	194	20	24	326
Produto G	12	31	11	26	0

A seguir, os experimentos são apresentados em detalhes.

4.3. Experimento 1

No primeiro experimento, considerando a singularidade dos compradores, ao comparar Ana com Noemi foi identificado $V_{max} = 437$ e $V_{min} = 1$. Assim, gera-se um arranjo de tamanho 436. Ao comparar os valores de Ana e Noemi para o Produto A, observa-se que os mesmos possuem uma similaridade de 0,81%, enquanto a mesma comparação para o Produto G gera uma similaridade de 38,7%. A similaridade dos demais produtos podem ser observados na Tabela 3.

Tabela 3. Similaridade entre os itens comprados por Ana e Noemi considerando a singularidade do histórico de compras entre ambas.

	A	B	C	D	E	F	G
Ana	3	4	7	1	6	31	12
Noemi	369	171	437	107	159	194	31
Similaridade	0,81%	2,33%	1,83%	0,93%	3,77%	16,00%	38,7%

Usando a Distância de Barbieri, é possível obter o percentual de similaridade de cada um dos itens comprados por ambos os compradores. Para se obter a similaridade entre os perfis dos compradores utilizados na comparação, é preciso considerar a similaridade de todos os produtos comprados.

É necessário considerar que a quantidade de itens comprados pelos perfis comparados é variável. Assim, realiza-se a soma da similaridade de cada um dos itens comprados e a divide pela quantidade de itens comparados. Isto tem a finalidade de identificar a média da similaridade do histórico de compras dos perfis comparados. Portanto, constata-se que a similaridade entre Ana e Noemi é de 9,19% com base na média aritmética das similaridades dos itens, isto é, são perfis pouco similares, o que significa que as compradoras possuem interesses distintos.

Ao realizar a comparação entre os perfis de Ana e Eugênio, considerando a singularidade dos perfis, foi identificado V_{max} = 31 e V_{min} = 1. Neste caso, o arranjo é de tamanho 30. Com isso, realizando a comparação entre os valores de Ana e Eugênio para o produto A, observa-se uma similaridade de 100%, visto que ambos compraram a mesma quantidade do produto em questão. A menor similaridade neste caso é para o produto D, no qual é possível observar uma similaridade de 33,33%. Usando média aritmética, observa-se neste caso uma similaridade de 80,51% para o perfil de Ana com Eugênio, conforme valores apresentados na Tabela 4.

Tabela 4. Similaridade entre os itens comprados por Ana e Eugênio considerando a singularidade do histórico de compras entre ambos.

	A	B	C	D	E	F	G
Ana	3	4	7	1	6	31	12
Eugênio	3	4	8	3	5	20	11
Similaridade	100%	100%	87,5%	33,33%	83,33%	67,74%	91,66%

Após realizar a comparação de Ana para com todos os demais compradores, obtêm-se um resultado conforme o apresentado na Tabela 5.

Tabela 5. Similaridade entre Ana e os demais compradores, considerando a singularidade entre o histórico de compras do comprador base e o comprador comparado.

Comprador	Noemi	Eugênio	Mauro	Larissa
Similaridade	9,19%	80,51%	32,93%	2,23%

Com isso, é possível afirmar que o comprador com perfil mais similar a Ana é Eugênio, com uma similaridade de 80,51%. Assim, pode-se afirmar que ao oferecer um

produto que Ana tenha comprado e Eugênio não, existe 80,51% de chance que Luiz compre o produto oferecido. Já o comprador com perfil menos similar a Ana é Larissa, com uma similaridade de 2,23%. Neste caso, ao oferecer um produto que Ana comprou e Larissa não, a chance de que Larissa compre o produto oferecido é de 2,23%.

4.4. Experimento 2

No segundo experimento, o objetivo é identificar qual dos compradores tem o perfil mais similar ao de Noemi. Para isso, aplicou-se a métrica de similaridade comparando Noemi com cada um dos demais compradores.

Novamente, considerando a singularidade dos compradores, ao comparar Noemi com Mauro foi identificado V_{max} = 437 e V_{min} = 0. Assim, gera-se um arranjo de tamanho 437. Ao comparar os valores de Noemi e Mauro para o Produto A, observa-se que os mesmos possuem uma similaridade de 3,52%, enquanto a mesma comparação para o produto G gera uma similaridade de 83,87%. A similaridade dos demais produtos podem ser observados exemplificados na Tabela 6.

Tabela 6. Similaridade entre os itens comprados por Noemi e Mauro considerando a singularidade do histórico de compras entre ambos.

	A	B	C	D	E	F	G
Noemi	369	171	437	107	159	194	31
Mauro	13	0	16	28	2	24	26
Similaridade	3,52%	0%	3,89%	26,16%	1,25%	12,37%	83,87%

Com a média aritmética aplicada, constata-se uma similaridade de 18,72% entre Noemi e Mauro.

Ao realizar a comparação entre os perfis de Noemi e Larissa, considerando a singularidade dos perfis, foi identificado V_{max} = 437 e V_{min} = 0. Novamente o arranjo é de tamanho 437. Comparando os valores de Noemi e Larissa para o produto A, observa-se uma similaridade de 86,17%. Já a menor similaridade neste caso é para os Produtos E e G, onde é possível observar uma similaridade de 0%, visto que Larissa não comprou os produtos em questão. A similaridade para os perfis de Noemi e Larissa é de 55,2%, conforme valores apresentados na Tabela 7.

Tabela 7. Similaridade entre os itens comprados por Noemi e Larissa considerando a singularidade do histórico de compras entre ambas.

	A	B	C	D	E	F	G
Noemi	369	171	437	107	159	194	31
Larissa	318	167	354	173	0	326	0
Similaridade	86,17%	97,66%	81,23%	61,84%	0%	59,50%	0%

Após realizar a comparação de Noemi para com todos os demais compradores, obtêm-se um resultado conforme o apresentado na Tabela 8.

Tabela 8. Similaridade entre Noemi e os demais compradores, considerando a singularidade entre o histórico de compras do comprador base e comprador comparado.

Comprador	Ana	Eugênio	Mauro	Larissa
Similaridade	9,19%	8,13%	18,72%	55,2%

Desta forma, é possível afirmar que o comprador com perfil mais similar a Noemi é Larissa, com uma similaridade de 55,2%. Afirma-se que ao oferecer um produto que Noemi tenha comprado e Larissa não, há 55,2% de chance que Larissa compre o produto oferecido. Já o comprador com perfil menos similar a Noemi é Eugênio, com uma similaridade de 8,13%. Neste caso, ao oferecer um produto que Noemi comprou e Eugênio não, a chance de que Eugênio compre o produto oferecido é de 8,13%.

4.5. Experimento 3

No terceiro experimento, a singularidade de Ana com os demais compradores e Noemi com os demais compradores não é considerada. Isso significa que o valor máximo, valor mínimo e consequentemente o tamanho do arranjo serão os mesmos para calcular a similaridade entre todos os perfis. Todos os demais passos do experimento foram seguidos da mesma forma que o anterior, exceto o de encontrar o valor máximo, mínimo e tamanho do arranjo. Ou seja, considera-se V_{max} = 437 e V_{min} = 0 em todas as comparações. Também é considerado 437 o tamanho do arranjo em todas as comparações.

Tabela 9. Similaridade entre Ana e os demais compradores, desconsiderando a singularidade entre o histórico de compras do comprador base e o comprador comparado.

Comprador	Noemi	Eugênio	Mauro	Larissa
Similaridade	9,19%	80,05%	32,47%	2,19%

Após comparar Ana com os demais compradores, é possível observar a similaridade apresentada na Tabela 9.

Tabela 10. Comparativo entre os resultados obtidos quando aplicada a métrica da Distância de Barbieri considerando e desconsiderando a singularidade entre o histórico de compras de Ana com o histórico de compras dos compradores comparados.

Comprador	**Similaridade Singular**	**Similaridade Geral**	**Diferença**
Noemi	9,19%	9,19%	0%
Eugênio	80,51%	80,05%	-0,57%
Mauro	32,93%	32,47%	-1,4%
Larissa	2,23%	2,19%	-1,79%

É possível observar que não há mudança significativa entre a similaridade calculada considerando a singularidade ou não, visto que a diferença de resultados é ínfima, conforme pode-se observar no comparativo feito na Tabela 10.

4.6. Experimento 4

No quarto experimento a singularidade de Noemi com os demais compradores não é considerada. Observam-se as similaridades apresentadas na Tabela 11.

Tabela 11. Similaridade entre Noemi e os demais compradores, desconsiderando a singularidade entre o histórico de compras do comprador base e o comprador comparado.

Comprador	Ana	Eugênio	Mauro	Larissa
Similaridade	9,19%	8,13%	18,72%	55,2%

Neste caso foi possível observar a total nulidade de diferença entre a comparação considerando a singularidade ou não, visto que não houve diferença nos resultados, conforme pode-se observar na Tabela 12.

Tabela 12. Comparativo entre os resultados obtidos quando aplicada a métrica da Distância de Barbieri considerando e desconsiderando a singularidade entre o histórico de compras de Noemi com o histórico de compras dos compradores comparados.

Comprador	**Similaridade Singular**	**Similaridade Geral**	**Diferença**
Ana	9,19%	9,19%	0%
Eugênio	8,13%	8,13%	0%
Mauro	18,72%	18,72%	0%
Larissa	55,2%	55,2%	0%

Com isso constata-se que é possível obter a similaridade de perfil entre os compradores considerando ou não a singularidade entre os perfis comparados.

5. Considerações Finais

Considerando que a metodologia proposta neste trabalho visa identificar a similaridade entre compradores e recomendar qual comprador possui maior probabilidade de adquirir um produto, pode-se afirmar que os resultados dos experimentos se mostram positivos.

O desenvolvimento desta metodologia foi impulsionado pela percepção de uma lacuna neste cenário, vindo a contribuir para identificar a similaridade de perfis de consumidor com base no histórico de suas compras. A Distância de Barbieri vem a contribuir também com quaisquer necessidades de identificar similaridade entre valores em um conjunto.

A representação do funcionamento do algoritmo da Distância de Barbieri utilizada nos experimentos demonstra a funcionalidade e a relevância da abordagem proposta. Acredita-se que ela possa ser implementada num cenário de gestão empresarial real, onde permitirá ao gestor conhecer de forma objetiva quais são os interesses de seus clientes, fornecendo ofertas personalizadas e remodelando as relações entre empresa e cliente.

A aplicação deste método neste cenário viria a reforçar a eficácia da métrica como forma de identificar similaridade entre perfis de consumidores em que não se dispõe das avaliações ou ratings para os produtos comprados. Outra possibilidade a ser explorada é

o desenvolvimento de um sistema que identifica um produto com baixo giro de estoque, identifica o último comprador deste produto, identifica compradores com o perfil similar ao do último comprador, e recomenda o produto com baixo giro aos potenciais compradores, com base na similaridade dos perfis.

Referências Bibliográficas

Aggarwal, C. C. (2016) "Recommender Systems: The Textbook" 1st edition Springer Publishing Company, Inc. doi: 10.1007/978-3-319-29659-3.

Barth, F. J. (2010) "Modelando o perfil do usuário para a construção de sistemas de recomendação: um estudo teórico e estado da arte", Revista de Sistemas de Informação da FSMA, n.6, p. 59-71.

Bastos, W. M. (2009) "Metodologia para recomendação de consultores ad-hoc baseada na extração de perfis do currículo Lattes", Universidade de Brasília, mês 7.

Cervi, C. R., Galante, R. and Oliveira, J. P. M. (2013) "Application of Scientific Metrics to Evaluate Academic Reputation in Different Research Areas", In International Journal of Social, Behavioral, Educational, Economic, Business and Industrial Engineering, World Academy of Science, Engineering and Technology, v.7, n.10, p. 2778-2788, http://waset.org/publications/17174.

Cervi, C. R., Galante, R. and Oliveira, J. P. M. (2013) "Comparing the Reputation of Researchers Using a Profile Model and Scientific Metrics", In: 2013 IEEE 16th International Conference on Computational Science and Engineering, p. 353-359, doi: 10.1109/CSE.2013.61.

Croft, W. B., Metzler, D. and Strohman, T. (2010) "Search engines: Information retrieval in practice", Addison-Wesley Reading, v.283.

Goldberg, D. *et al.* (1992) "Using Collaborative Filtering to Weave an Information Tapestry", Commun. ACM, New York, NY, USA, v.35, n.12, p. 61-70. doi: 10.1145/138859.138867.

Henriques, P. M. M. (2016) "Smart search in a distributed environment", Faculdade De Engenharia Da Universidade Do Porto, https://repositorio-aberto.up.pt/handle/10216/85259.

Herlocker, J. L. (2000) "Understanding and Improving Automated Collaborative Filtering Systems", University of Minnesota, Minneapolis, MN, USA.

Levenshtein, V. I. (1966) "Binary codes capable of correcting deletions, insertions, and reversals", In: Soviet physics doklady, v.10, n.8, p. 707-710, https://nymity.ch/sybilhunting/pdf/Levenshtein1966a.pdf.

Maria, S. A. A. (2017) "RecETC : uma funcionalidade baseada na recomendação de conteúdo para auxiliar no processo de escrita coletiva digital", Universidade Federal do Rio Grande do Sul, Porto Alegre, mês 5, http://hdl.handle.net/10183/170310.

Meurer, H. (2014) “Ferramenta de gerenciamento e recomendação como recurso na aprendizagem baseada em projeto em design”, Universidade Federal do Rio Grande do Sul, Porto Alegre, mês 12, http://hdl.handle.net/10183/115721.

Plumbaum, T. (2015) “User modeling in the social semantic web”, Technische Universität Berlin, Berlim, mês 12, http://dl.acm.org/citation.cfm?id=2887675.2887684.

Resnick, P. and Varian, H. R. (1997) “Recommender Systems”, Commun. ACM, New York, NY, USA v.40, n.3, p. 56-58, doi: 10.1145/245108.245121.

Ricci, F., Rokach, L. and Shapira, B. (2011) "Introduction to Recommender Systems Handbook”, In: Recommender Systems Handbook, Springer Publishing Company, Inc., Boston, MA, p. 1-35, doi: 10.1007/978-0-387-85820-3_1.

Rich, E. (1979) “Building and Exploiting User Models”, In: Proceedings of the 6th International Joint Conference on Artificial Intelligence – v.2, Tokyo, Japan, series: IJCAI'79, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, p. 720-722, http://dl.acm.org/citation.cfm?id=1623050.1623079.

Santana, L. L. B. S. (2018) “Explorando relações entre usuários em um sistema de recomendação híbrido baseado em filmes”, Universidade Federal da Bahia, Salvador, mês 10.

Vieira, P. K. M. (2013) “Recomendação semântica de conteúdo em ambientes de convergência digital”, Universidade Federal da Paraíba, João Pessoa, mês 3, https://repositorio.ufpb.br/jspui/handle/tede/7826.

Vivian, G. R. (2017) “Recomendação de carreira de pesquisadores: uma abordagem baseada em personalização, similaridade de perfil e reputação”, Universidade de Passo Fundo, Passo Fundo, http://tede.upf.br:8080/jspui/handle/tede/1425.

Yao, L., Xu, Z., Zhou, X. and Lev, B. (2019) “Synergies Between Association Rules and Collaborative Filtering in Recommender System: An Application to Auto Industry.” In: García Márquez F., Lev B. (eds) Data Science and Digital Business. Springer, Cham.