Sistemas de Recomendação: Um estudo de caso

Matheus Pellizzaro[1], **Nilson Ribeiro Modro**[1], **Luiz Cláudio Dalmolin**[1], **Alex Luiz de Sousa**[1]

[1]Universidade do Estado de Santa Catarina (UDESC)
Centro de Educação do Planalto Norte (CEPLAN)

matpellizzaro@gmail.com, nilson.modro@udesc.br, lcdalmolin@gmail.com, alex.sousa@udesc.br.

Resumo. *Com o crescimento contínuo da internet e o aumento do volume de informações geradas a cada dia, se torna premente a utilização de sistemas de recomendação como uma alternativa para viabilizar a "navegação" nesse novo contexto. Esta monografia se propõe a apresentar uma análise sobre como funcionam os sistemas de recomendação e para tanto, detalha alguns dos métodos mais usados na hora de criar sistemas de recomendação, tais como: baseado em conteúdo, filtragem colaborativa e sistemas híbridos. Foram utilizadas diversas ferramentas e bibliotecas do Python, e uma base de dados publica (Movielens) para realização de testes, estudos e obtenção dos resultados. Este estudo contribui para que você entenda a base sobre sistemas de recomendação e lhe apresenta um guia nos estudos nesta extensa área de pesquisa.*

Abstract. *With the continued growth of the Internet and the increasing volume of information generated every day, if pressing makes use of recommender systems as an alternative to enable the "navigation" in this new context. This paper aims to present an analysis of how the recommendation systems and to do so, details some of the methods most used when creating recommender systems, such as content-based, collaborative filtering and hybrid systems. various tools and Python libraries were used, and a publishing database (Movielens) for performing tests, studies and the results obtained. This study helps you to understand the basis on recommendation systems and presents you with a guide on studies in this large area of research*

1. Introdução

O avanço da internet em tempo de uso diário pelas pessoas vem abrindo novas oportunidades de negócios lucrativos. A poucos anos atrás, por exemplo, era incipiente o comércio eletrônico. Atualmente, os chamados e-commerces são os que mais crescem em vendas e lucros pelo mundo.

Segundo [Diego Ivo 2015], a estimativa é que o comércio eletrônico tenha gerado mais de 136 milhões de pedidos no último ano, com um valor médio por compra de R$ 316. Um dos fatores de sucesso foi o bom desempenho apresentado nas datas comemorativas. Somente na Black Friday, por exemplo, o e-commerce ultrapassou a marca de R$1 bilhão em um único dia, um grande recorde do setor.

A explicação para este crescimento gigantesco é simples, com cada dia mais estímulos e novidades, pessoas estão ficando sem paciência e tempo, desta forma tendem a querer economizar em tarefas que antes eram feitas com mais tranquilidade,

por exemplo, consultas a saldo de conta corrente, pagamento de faturas (água, luz, telefone, aluguel, etc.), compras variadas (eletrodomésticos, eletrônicos, comida, etc.), entre diversos outros.

Além da motivação de tempo, outro fator tem atraído cada vez mais usuários para as compras on-line: os preços melhores e facilidade para encontra-los com as mais diversas ferramentas de comparação.

Pensando neste sentido, lojas online trabalham não só para ter uma loja com bastantes ofertas tentadoras, mas para atrair cada vez mais o público para suas ofertas e tentar fazer com que comprem não somente o necessário, mas também o desnecessário, gerando vendas por impulso.

É neste momento que entram em cena as mais variadas técnicas de marketing digital, como: SEO (*Search Engine Optimization* – Otimização para Buscadores), E-mail Marketing, Marketing de Conteúdo, Sistemas de Recomendação, etc.

Estas diversas técnicas têm como objetivo principal permitir que o usuário encontre mais facilmente produtos e ofertas que tenham maiores chances de despertar seu interesse e consequente evolua para uma ação/compra.

Sistemas de recomendação são considerados uma inovação tecnológica recente, pois modificam a forma como usuários encontram informações, é possível afirmar que estamos deixando a era da informação para entrar na era da recomendação. O Google pode ser considerado um bom exemplo de sistema de recomendação, que não está limitado apenas a mostrar páginas catalogadas, mas também fornecer ao usuário mais informações sobre o que ele pesquisa em sua página de resultados como: biografias, resultado de equações matemáticas, conversão de unidades, etc.

Este artigo tem como objetivo apresentar de um modo simples e objetivo os sistemas de recomendação e seu funcionamento, e um exemplo de sistema que faz uso dos conceitos de sistemas de recomendação para recomendar filmes.

2. Revisão Bibliográfica

Nesta seção será possível entender um pouco sobre os principais conceitos, métodos e bibliotecas usadas em sistemas de recomendação. Esses conceitos e explicações ajudarão no entendimento do estudo de caso apresentado na seção 3.

2.1. Sistemas de Recomendação

Com a quantidade de informações e com a disponibilidade facilitada das mesmas pelo uso da Internet, as pessoas se deparam com uma diversidade muito grande de opções. Muitas vezes um indivíduo possui muito pouca ou quase nenhuma experiência pessoal para realizar escolhas entre as várias alternativas que lhe são apresentadas. A questão relevante neste momento refere-se a como proceder nestes casos? Para minimizar as dúvidas e necessidades que temos frente à escolha entre alternativas, geralmente confiamos nas recomendações que são passadas por outras pessoas, as quais podem chegar de forma direta [Maes & Shardanand 1995].

Os sistemas de recomendação auxiliam no aumento da capacidade e eficácia deste processo de indicação já bastante conhecida na relação social entre seres humanos [Resnick & Varian 1997], ou seja, a indicação de um produto ou serviço, se bem-feita, tem grandes chances de gerar uma venda.

O que um sistema de recomendação faz basicamente é predizer o quanto se pode gostar de um certo produto ou serviço. Ele usará uma lista de vários itens ordenada de acordo com o interesse, além disso, ele ainda "explicará", algumas vezes, porque o item está sendo recomendado. A figura 1 a seguir mostra uma recomendação em um sistema de vendas *online*.

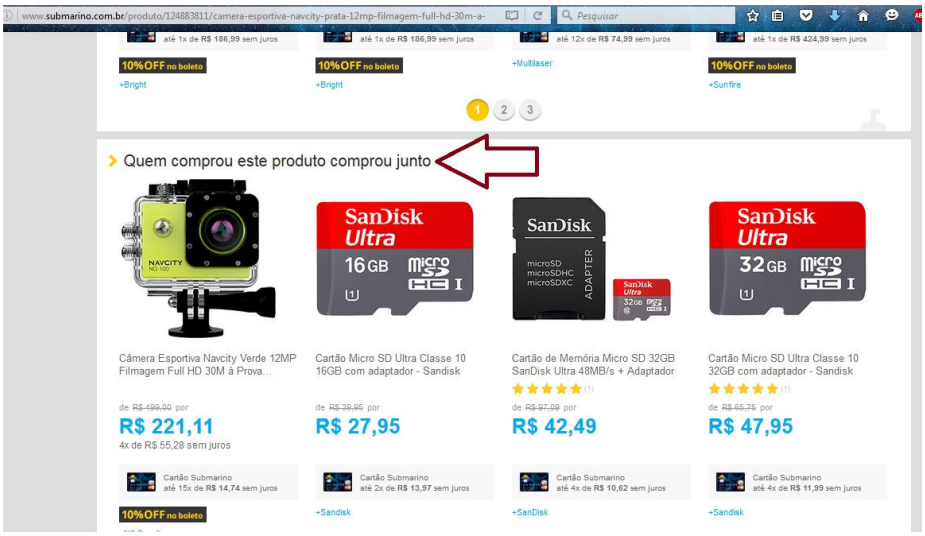

Figura 1 – Indicação de produto Submarino.com.br.

Na figura 1 é possível ver o sistema de recomendação do Submarino, que recomenda produtos ao usuário, baseado no que outros usuários compraram ao adquirir o produto que ele está vendo no momento. Esta tática tenta fazer o cliente comprar não só o produto que lhe despertou interesse primeiro, mas também comprar "complementos" à esse produto e fazê-lo adquirir produtos adicionais.

A maioria das empresas produz seu sistema de recomendação baseado não apenas no que o usuário comprou ou está interessado em comprar, mas os sistemas de recomendação vão além, oferecendo produtos que o usuário "nem sabe que quer", baseado em outros usuários que compraram produtos semelhantes e tem gostos/perfis parecidos com o do cliente em questão.

Há diversos exemplos no mercado de empresas investindo em sistemas de recomendação melhores para aumentar suas vendas e engajamento do público com seus E-commerces. Um dos exemplos mais claros disso é a Netflix, que antes alugava DVDs pela internet e hoje é um poderoso site de *streaming* (armazenamento de vídeos na nuvem para visualização dos usuários) de filmes.

Netflix é uma empresa de *streaming* de filmes, eles fazem recomendações baseadas nos filmes que os clientes assistiram. No final de 2006 eles anunciaram um prêmio de US$ 1 milhão para a primeira pessoa que melhora-se a precisão do seu sistema de recomendação em 10%. Milhares de equipes de todo o mundo entraram na disputa e, a partir de abril de 2007, o líder da equipe conseguiu uma melhoria de 7%. Usando dados sobre os filmes que cada cliente gostou, desta forma o Netflix ganhou a capacidade de recomendar filmes para outros clientes, filmes que eles sequer tinham ouvido falar, fazendo com que eles voltassem para assistir mais. Qualquer maneira de melhorar o seu sistema de recomendação vale a pena para a Netflix. [Segaran 2007].

Existem três tipos de sistemas de recomendação, são eles:

- Sistemas baseados em conteúdo – Avaliam o conteúdo que o usuário gostou, e sugere conteúdos semelhantes. Exemplo: Usuário gostou do filme Toy Story (Desenho), o sistema poderia sugerir A Era do Gelo (Desenho).
- Sistemas baseados em filtragem colaborativa – Avalia o que usuários semelhantes gostaram, e sugere conteúdos que esses usuários gostaram e que o usuário ainda não viu. Exemplo: Matheus é semelhante à Andrea. Andrea gosta do filme Os Vingadores, sistema recomenda Os Vingadores para Matheus.

- Sistemas de recomendação híbridos – Mescla diversas técnicas dos dois estilos anteriores, e através de ontologias (dados simbólicos), consegue realizar recomendações efetivas.

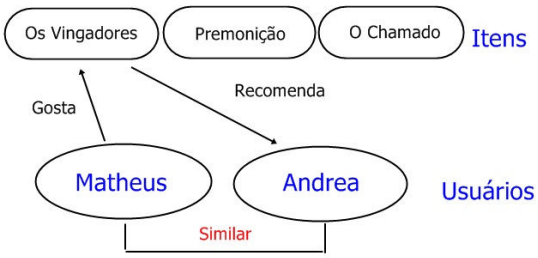

Figura 2. Exemplo de sistema baseado em filtragem colaborativa.

Na figura 2 encontra-se um exemplo bem simples de como funcionam sistemas baseados em filtragem colaborativa. Este é considerado por muitos, como um dos melhores métodos para recomendação de conteúdo.

Ainda há uma última forma utilizada para recomendar conteúdos, a recomendação baseada em localização. Este tipo de recomendação leva em conta o local que o usuário está para realizar uma recomendação, um aplicativo que faz uso disso é o Foursquare.

O Foursquare é um aplicativo bastante conhecido que recomenda locais e serviços próximos ao usuário, como restaurantes, lojas, etc.

Nos próximos tópicos será possível entender como os sistemas de recomendação podem ser "treinados" para recomendar e como eles conseguem enfim fazer suas recomendações.

2.2 Tipos comuns de sistemas de recomendação

A aprendizagem por árvore de decisão conforme [Takahashi 2015], é um algoritmo de aprendizagem supervisionada que utiliza uma árvore de decisão como modelo preditivo que mapeia as observações de um item para concluir sobre o valor desejado acerca do item.

As árvores de decisão recebem como entrada um conjunto de atributos e retornam uma decisão que é o valor predito para a entrada. Sua estrutura é formada por nós e ramos, sendo que cada nó contém um teste em um atributo e cada ramo representa um possível valor do atributo [Maimon & Rokach 2008].

A principal desvantagem deste método é que é necessário criar ou treinar todas as regras e suas respectivas respostas, o que torna o processo progressivamente lento, porém quando bem treinado, pode ser bastante eficiente.

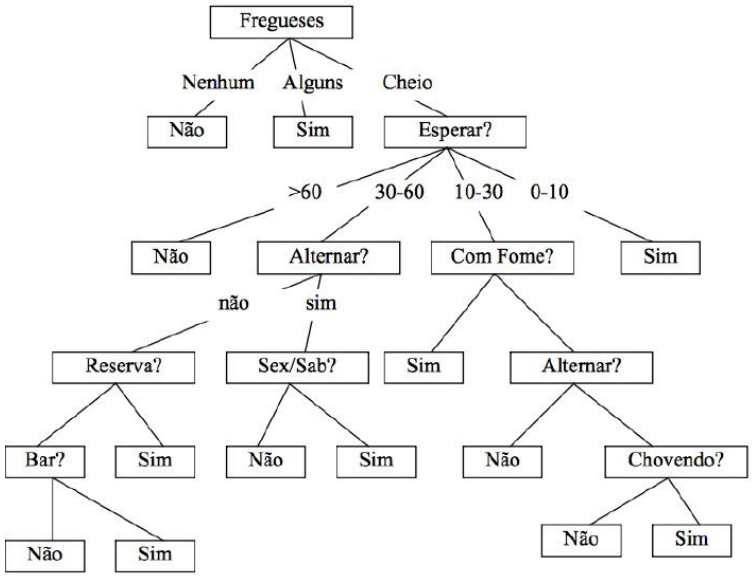

Figura 3. Exemplo de árvore de decisão.

A figura 3 exemplifica como seria uma árvore de decisão ao procurar um restaurante para jantar ou almoçar.

Outro tipo amplamente usado são as redes bayesianas, que são modelos baseados em grafos para tomada de decisão onde há incerteza, os nós representam as variáveis a serem trabalhadas (discreta ou contínua), e os arcos representam a conexão direta entre os nós.

Um exemplo de rede Bayesiana pode ser a relação de probabilidade entre doenças e sintomas. Dados os sintomas, a rede pode ser usada para calcular as probabilidades de presença das doenças [Takahashi 2015].

Já a análise de agrupamentos (ou Clustering) é a tarefa de agrupar um conjunto de dados de forma que os objetos que pertencem ao mesmo grupo (chamado de cluster) sejam mais semelhantes (de acordo com alguns critérios) entre eles do que aos que estão em outros grupos (clusters). [Bailey 1994].

Segundo [Takahashi 2015], existem diversos tipos de algoritmos para agrupamento dos objetos, seja por distância entre os pontos ou por aproximação dos centroides, e cada um pode retornar um resultado diferente.

Um exemplo prático seria a separação de pessoas em uma rede social por grupos, por exemplo, família, melhores amigos, bloqueados, etc.

2.5 MovieLens

MovieLens é um conjunto de dados, nomeadamente avaliações sobre filmes que foram coletados pelo Projeto de Pesquisa GroupLens na Universidade de Minnesota, através do site movielens.umn.edu [Grouplens 2016].

Os dados foram coletados durante o período de sete meses a partir de 19 de setembro de 1997 à 22 de abril de 1998. Estes dados foram processados - usuários que tinha menos de 20 classificações ou não tinham dados demográficos completos foram removidos [Grouplens 2016].

3. Estudo de Caso

Esta seção mostrará uma análise breve sobre um algoritmo desenvolvido na Universidade da Islândia (*Háskóli Íslands*) por Einar Héðinsson, Ari Tómasson, Arnar

Magnússon e Sigurður H. Ólafsson. Este sistema usa o conjunto de dados MovieLens para fazer recomendações de filmes e recomendar usuários semelhantes.

As análises ajudarão a entender como são feitas as recomendações de filmes aos usuários cadastrados, quais critérios e cálculos são usados para que a recomendação seja possível.

3.1. Coeficiente de Correlação de Pearson

O coeficiente de correlação de Pearson é usado para medir o grau de correlação entre duas variáveis, e qual a "direção" (positiva ou negativa) entre duas variáveis de escala métrica. O método usado pelo sistema é baseado no coeficiente de correlação de Pearson, que gera um número, variando de 1 à -1.

Quanto mais próximo de 1 o coeficiente esta, mais semelhantes são os usuários analisados, quanto mais próximo à -1, maior a diferença entre eles.

	A	B	C	D
1		**usuario 1**	**usuario 2**	**usuario 3**
2	**filme 1**	5	5	3
3	**filme 2**	3	2	5
4	**filme 3**	1	2	5
5	**filme 4**	5	5	2
6	**filme 5**	4	4	1
7				
8				
9				
10		**1 x 2**	**1 x 3**	**2 x 3**
11	**Correlação**	0,906327	-0,71826	-0,7925

Figura 4. Exemplos de coeficiente de correlação de Pearson.

Conforme é possível observar na figura 4, o usuário 1 é mais semelhante ao usuário 2, enquanto o usuário 1 é mais semelhante ao usuário 3.

A figura 5 contém a fórmula para calcular o coeficiente de correlação de Pearson. Onde "n" é a quantidade de "filmes" à analisar, multiplicado pela somatória das somas das notas do "usuário 1" por "usuário 2", menos somatória de todas as notas do "usuário 1", multiplicado pela somatória das notas do "usuário 2", tudo isso dividido por raiz quadrada de "n", vezes somatória das notas do "usuário 1" elevado à dois, menos, somatória de todas as notas do "usuário 1", e o resultado dessa somatória elevado à dois, e dentro de outra raiz quadrada, idêntica a primeira, só que com os valores do "usuário 2" e multiplicado pelo valor da primeira.

$$r = \frac{n.\sum(x.y) - (\sum x).(\sum y)}{\sqrt{n.\sum x^2 - (\sum x)^2}.\sqrt{n.\sum y^2 - (\sum y)^2}}$$

Figura 5. Fórmula para cálculo do Coeficiente de Correlação de Pearson

3.2. Leitura do MovieLens com Pandas

O Pandas não é usado no sistema de recomendação da Islândia, porém é uma opção na hora de manipular e acessar dados do conjunto de dados MovieLens. É uma biblioteca do Python que trabalha com o conceito de "DataFrame", que é uma estrutura de dados criada pela biblioteca para o acesso aos dados presentes no conjunto.

A figura 6 apresentada a seguir mostra um código bem simples, que mostra os 25 filmes com mais classificações, presentes no MovieLens de duas formas diferentes.

```
import pandas as pd
import numpy as np
# Leitura dos dados da base.
u_cols = ['user_id', 'age', 'sex', 'occupation', 'zip_code']
users = pd.read_csv('ml-100k/u.user', sep='|', names=u_cols,
                    encoding='latin-1')

r_cols = ['user_id', 'movie_id', 'rating', 'unix_timestamp']
ratings = pd.read_csv('ml-100k/u.data', sep='\t', names=r_cols,
                      encoding='latin-1')

# Apenas considera as primeiras 5 colunas, as outras como gênero, etc, não são carregadas neste momento.
m_cols = ['movie_id', 'title', 'release_date', 'video_release_date', 'imdb_url']
movies = pd.read_csv('ml-100k/u.item', sep='|', names=m_cols, usecols=range(5),
                     encoding='latin-1')

# Cria DataFrame unindo todas as leituras de arquivos.
movie_ratings = pd.merge(movies, ratings)
lens = pd.merge(movie_ratings, users)

#Os 25 melhores
most_rated = lens.groupby('title').size().sort_values(ascending=False)[:25]
print (most_rated)

#Os 25 melhores de outra forma
print (lens.title.value_counts()[:25])

# Chamando por titulo e mostrando os cabecas
#movie_stats = lens.groupby('title').agg({'rating': [np.size, np.mean]})
#print (movie_stats.head())
```

Figura 6. Código fonte para leitura do MovieLens com Pandas.

3.2. Recomendação de usuário semelhante

Como explicado no tópico anterior, o coeficiente de correlação de Pearson que gera a "decisão" sobre qual usuário é mais semelhante ao outro. Descobrindo os mais semelhantes é possível então fazer a recomendação.

No sistema da Islândia há opção de encontrar usuários similares, além da função principal, que é conseguir recomendações de filmes para o usuário.

A recomendação de filmes é feita, neste sistema, da seguinte forma: primeiro o sistema identifica quais usuários tem maior coeficiente positivo com o usuário que está necessitando recomendações, na maioria das vezes mesmo sendo bem semelhantes certos usuários, sempre haverá filmes que um ou outro não viu, e que seus semelhantes viram e avaliaram. O sistema então pega esses filmes, considerando sempre o usuário mais semelhante e organiza pelos de melhor nota, e vai sugerindo ao usuário. Após chegar ao final dos filmes diferentes visto pelo usuário mais semelhante, ele passará ao próximo usuário mais semelhante, até que as recomendações se esgotem.

Na figura 7 é possível observar o algoritmo em Python responsável pela classificação dos usuários, baseado em suas avaliações.

```python
def correlation_of_users(self, x, y, userID):
    self.x = x
    self.y = y
    self.userID = userID

    if len(x) != 0: # and len(x) != 1:

      n = len(x)
      X_sum = float(sum(x))
      Y_sum = float(sum(y))
      X_sum_sq = sum(map(lambda x: pow(x, 2), x))
      Y_sum_sq = sum(map(lambda x: pow(x, 2), y))
      psum = sum(imap(lambda x, y: x * y, x, y))
      num = psum - (X_sum * Y_sum/n)
      den = pow((X_sum_sq - pow(X_sum, 2) / n ) * (Y_sum_sq - pow(Y_sum, 2) / n), 0.5)
      if den == 0: return 0
      return num/den
    else:
      return 0
```

Figura 7. Coeficiente de Correlação de Pearson em Python.

4. Conclusões

Sistemas de recomendação estão por todas as partes hoje em dia, principalmente em lojas online e aplicativos móveis. Estão ficando cada vez mais aprimorados, afinal é sempre necessário aumentar o engajamento dos usuários e consequentemente o lucro, usando a mesma estrutura ou até diminuindo os gastos necessários em recursos.

Além disso, sistemas de recomendação ajudam a instigar os clientes, os fazendo encontrar produtos úteis, que nem mesmo eles sabiam que seriam uteis a eles.

Há diversas ferramentas e frameworks disponíveis no mercado para o desenvolvimento de sistemas de recomendação, basta apenas focar em um deles, aproveitando suas funções pré-programadas, com a combinação delas é possível criar um sistema de recomendação próprio.

A linguagem de programação mais usada para escrever sistemas de recomendação é o Python, devido a agilidade no processamento de uma grande quantidade de dados. Além disso, Python é uma linguagem própria para trabalhar com cálculos matemáticos avançados, os quais são usados em algoritmos de sistemas de recomendação.

A recomendação é ainda uma área relativamente nova no mercado e merece mais estudos feitos. Em muitos casos, é possível vislumbrar nestes sistemas, conceitos e aplicações de Inteligência Artificial. Além de sugerir o que é esperado como interesse, fazer sugestões inusitadas de produtos, textos, etc. que não estão na mesma categoria ou área do conteúdo visualizado no momento, mas que possui similaridades sazonais, por exemplo, pessoas que gostam de matemática e compram calculadoras, podem vir a gostar de palavras cruzadas, que não tem relação direta com matemática, porém ambos envolvem uso de lógica.

O sistema operacional mais adequado para trabalhar com sistemas de recomendação é o Linux, pois a grande maioria dos algoritmos de sistemas de recomendação são feitos em Python, que é uma linguagem que possui alto desempenho e poder de processamento em grande quantidade de dados. O Python e diversas de suas bibliotecas já vêm instalados como padrão no Linux.

6. Referências

Bailey, Kenneth D. (1994). Numerical Taxonomy and Cluster Analysis. SAGE. Thousand Oaks, p. 35-66.

Grouplens. (2016). README: SUMMARY & USAGE LICENSE. Disponível em <http://files.grouplens.org/datasets/movielens/ml-100k-README.txt>. Acesso em 14 de mai. 2016.

Ivo, Diego. (2015). Brasil.com.br, que País é esse?. Disponível em: < http://www.conversion.com.br/blog/brasil-com-br-que-pais-e-esse/>. Acesso em 04 de mai. 2016.

Maes, P.; Shardanand, U. (1995). Social information filtering: Algorithms for automating "word of mouth", In: Human Factors in Computing Systems, New York, p. 210-217.

Maimon, Oded; Rokach, Lior. (2008). Data mining with decision trees: theory and applications. World Scientific Pub Co Inc.

Resnick, P. e Varian, H. R. (1997) Recommender Systems Communications of the ACM, New York, v.40, pub.3, p. 55-58.

Segaran, Toby. (2007). Programming Collective Intelligence - Building Smart Web 2.0 Applications. Sebastopol: O'Reilly Media.

Takahashi, Marcos M. (2015). Estudo comparativo de Algoritmos de Recomendação. USP. São Paulo.