06

# Recomendando cursos

### Transcrição

Já falamos sobre recomendações baseadas na similaridade dos usuários, ao longo das aulas vamos trabalhar justamente com a recomendação de cursos. Vamos elaborar uma pequena simulação utilizando o exemplo anterior, utilizando os cursos de Android, Collections, NHibernate e Cordova disponíveis aqui na Alura. Montando a seguinte tabela:

| - | Android | Collections | NHibernate | Cordova |
|---|---|---|---|---|
| Guilherme | 10 | 9 | | 7 |
| Joana | 7 | | 9 | 10 |
| João | | | | 10 |
| Daniela | 10 | | 7 | |

Vamos recomendar um curso para o Guilherme. Para isso já sabemos que precisamos encontrar alguém similar a ele, alguém próximo. Estes termos são muito vagos e para auxiliar nosso pensamento vamos utilizar um conceito de diferente. Qual a diferença entre o Guilherme e a Joana e entre o Guilherme e os demais? Vamos criar uma coluna auxiliar para nosso acompanhamento.

| - | Android | Collections | NHibernate | Cordova | Diferença |
|---|---|---|---|---|---|
| Guilherme | 10 | 9 | | 7 | |
| Joana | 7 | | 9 | 10 | |
| João | | | | 10 | |
| Daniela | 10 | | 7 | | |

Comparar o Guilherme com ele mesmo ou com alguém que seja muito parecido a ele deve gerar não exatamente uma diferença, pois neste caso a diferença é zero. Já em comparação com a Joana, ocorre uma diferença alta entre as notas. Quando comparamos com João, ocorrem diferenças, mas também ausência de dados, portanto, não é possível afirmar que ela é alta ou baixa, por isso consideramos que a diferença entre eles é a média. No caso do Guilherme e Daniela como existem notas semelhantes, a diferença será baixa. Até aqui consideramos:

| - | Android | Collections | NHibernate | Cordova | Diferença |
|---|---|---|---|---|---|
| Guilherme | 10 | 9 | | 7 | 0 |
| Joana | 7 | | 9 | 10 | Muito alta |
| João | | | | 10 | Média |
| Daniela | 10 | | 7 | | Baixa |

Isso fornece um entendimento básico da situação, mas o que queremos é uma fórmula matemática que nos ajude a calcular a diferença. Uma fórmula na qual comparando o Guilherme com ele mesmo resulte em um valor 0, comparando com a Joana dê um número muito alto e com os demais conforme os resultados de os valores já descritos na tabela. Mas como fazer isso?

Note que se considerarmos apenas as notas, podemos representar o Guilherme como: 10, 9, ? (mistério) e 7. Podemos representar a Joana como: 7, ?, 9, 10. Considerando apenas essas informações não importa se estamos avaliando livros, cursos, restaurantes ou outras coisas, ao menos não nesse ponto.

Se você prestou bem atenção na tabela e na representação que estamos fazendo, vai ficar simples de assimilar que a tabela é uma matriz e que cada usuário nesta matriz pode ser representado com um vetor. Lembra do plano cartesiano? X e Y? X e Y são vetores em duas dimensões. Nossos usuários, por outro lado, estão dispostos em 4 dimensões. Dessa forma, para calcular a distância (*diferença*) entre o Guilherme e Joana teremos:

**F** distância entre os vetores:

- Guilherme e ele mesmo: (10,9,?,7) e (10,9,?,7)
- Guilherme e Joana: (10,9,?,7) e (7,?,9,10)
- Guilherme e João: (10,9,?,7) e (?,?,?,10)
- Guilherme e Daniela: (10,9,?,7) e (10,?,7,?)

Considerando a matemática para cálculo de distâncias, a distância entre um ponto e ele mesmo é 0, então, a diferença entre o Guilherme e ele mesmo, é 0. Mas como resolvemos os outros casos?

Vamos calcular a diferença de cada dimensão separadamente. Primeiro considerando o Guilherme e a Joana. Para o curso de Android temos 10 e 7, neste caso a diferença (*subtração*) é 3. Para o curso de Collections a Joana não tem um valor definido, então vamos descartar essa dimensão. O mesmo acontece para o curso de NHibernate, onde o Guilherme não tem um valor. Para o curso de Cordova, os valores são 7 e 10, que nos dá uma diferença de -3. Somando 3 e -3 temos 0 como resultado. Espera, tá errado! O Guilherme e a Joana são idênticos mesmo avaliando os cursos de forma tão diferente?

Acabamos de observar que a definição de distância não se aplica neste caso, portanto, vamos partir para outra. Seguindo a mesma regra iremos agregar a exponenciação, ou seja, vamos elevar a diferença entre os valores ao quadrado. Para o primeiro valor, temos 9, já que 3 ao quadrado é 9 e no segundo caso teremos 9 também, isso por que -3 ao quadrado também é 9. A exponenciação faz a conversão de sinal e ao final acabaremos somando 9 e 9 que terá como resultado 18. Então, nesta definição a diferença entre o Guilherme e a Joana é 18.

Levando esse raciocínio para os demais usuários chegaríamos ao seguinte resultado.

| - | Android | Collections | NHibernate | Cordova | Diferença |
|---|---|---|---|---|---|
| Guilherme | 10 | 9 | | 7 | 0 |
| Joana | 7 | | 9 | 10 | 18 |
| João | | | | 10 | 9 |
| Daniela | 10 | | 7 | | 0 |

Lembre-se que a diferença neste caso compara apenas o Guilherme com os demais.

A função que descrevemos resolve nosso problema, porém, ela tem um detalhe importante de se fazer nota: ela desconsidera os cursos desconhecidos, o que fará a diferença ser muito grande em alguns casos.

Sabendo disso, podemos utilizar de uma divisão do valor obtido pela soma dos quadrados das diferenças entre os vetores. O valor pelo qual a diferença será dividida é a soma de diferenças calculadas, ou seja, pelo número de cursos que conhecemos nos dois vetores. Entre o Guilherme e a Joana por exemplo, a diferença entre dois vetores puderam ser calculados, o que nos leva a dividir 18 por 2. Entre o Guilherme e o João, apenas um dos vetores pode ser calculado, o que nos levaria a dividir 9 por 1.

> **Importante:** Os casos onde os dois vetores não tem dados presentes são descartados, ou seja, a diferença foi impossibilitada por causa da ausência de valor em um dos dois vetores exclusivamente.

Considerando essa nova abordagem, chegaríamos ao fato de que o Guilherme é diferente da Joana e do João de forma igual.

| - | Android | Collections | NHibernate | Cordova | Diferença |
|---|---|---|---|---|---|
| Guilherme | 10 | 9 | | 7 | 0 |
| Joana | 7 | | 9 | 10 | 18 / 2 = 9 |
| João | | | | 10 | 9 / 1 = 9 |
| Daniela | 10 | | 7 | | 0 |

Note que dependendo da abordagem que utilizamos para calcular a similaridade nos aproximamos ou nos distanciamos das pessoas no mundo. Neste caso, nosso mundo possui 4 dimensões.

Agora que encontramos a fórmula matemática que calcula a distância entre dois usuários baseado nos seus vetores de notas, podemos encontrar qualquer similaridade, sendo elas, primeira, segunda etc. Assim como qualquer número de usuários próximos.

Por fim, para resolver o problema de qual nota o Guilherme daria para o curso de NHibernate, a estimativa a que chegamos é 7. Isso depois de descobrir que dentro os demais usuários, Daniela é a mais parecida com ele e ela deu exatamente nota 7 para o curso.

| - | Android | Collections | NHibernate | Cordova | Diferença |
|---|---|---|---|---|---|
| Guilherme | 10 | 9 | 7 | 7 | 0 |
| Joana | 7 | | 9 | 10 | 9 |
| João | | | | 10 | 9 |
| Daniela | 10 | | 7 | | 0 |

Claro que neste caso, só podemos recomendar um curso para o Guilherme, porque são 4 cursos e o Guilherme já fez 3 deles. No caso do João, poderíamos recomendar 3 cursos, o que também é uma opção configurável.

Essa é a verdadeira lógica por trás dos sistemas de recomendações. Um bom exemplo desse tipo de sistema é o Netflix. Nos próximos passos faremos todo esse processo com linhas de código em Java! Vamos lá?