

Sistemas de Recomendação

Recomendadores Content-based

Representação dos Usuários



PUC Minas

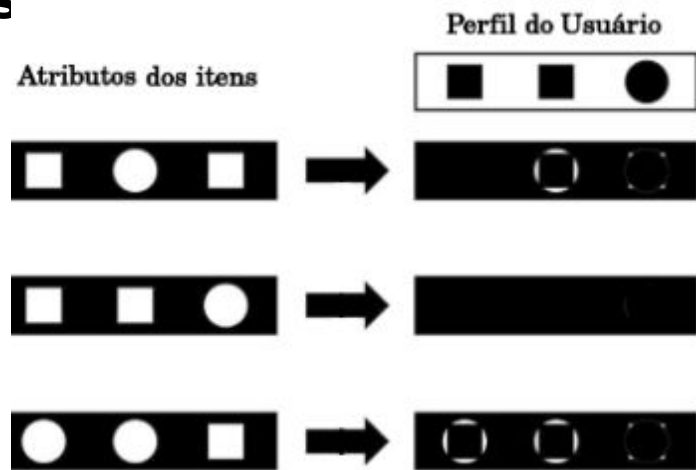
Pós-Graduação Lato Sensu

Nícollas Silva

Representação dos Usuários

O perfil do usuário deve ser modelado sobre as mesmas características dos itens.

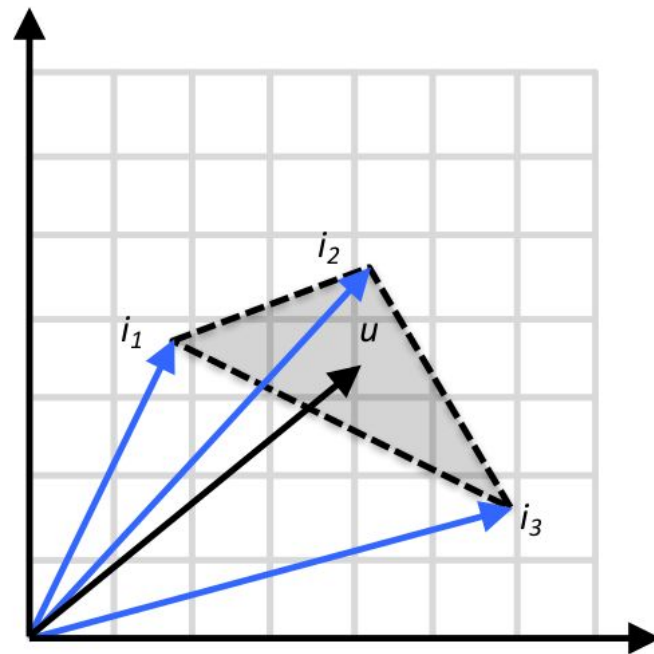
- Vetor de *componentes agregados*



Representação dos Usuários

- Usuário avaliou
 - i_1 : ★★ ★
 - i_2 : ★★ ★★ ★
 - i_3 : ★★ ★
- Modelo do usuário

$$\vec{u} = 3\vec{i}_1 + 5\vec{i}_2 + 3\vec{i}_3$$



Modelo de Rocchio

- Cada item é um vetor de componentes
- Cada usuário é um vetor de itens consumidos

$$\vec{u} = \frac{1}{|R_u|} \sum_{j \in R_u} r_{uj} \vec{j}$$

R_u : itens consumidos pelo usuário u

r_{uj} : rating do usuário u ao item j

- Tarefa de predição:

$$sim(\vec{u}, \vec{i})$$

Vantagens & Desvantagens

- Efetiva para o contexto de *item cold-start* (novos no sistema).
 - Não preciso das informações passadas dos itens.
- Não é tão efetiva em cenários com muitas informações.
 - Outras abordagens conseguem obter melhor desempenho.
- Muito aplicada em recomendadores híbridos.
 - Consegue agregar outras informações (*features*) aos modelos.

Abordagens k -NN

- Um método não paramétrico usado para a tarefa de classificação e regressão.
 - Tenta encontrar os k elementos mais similares ao elemento alvo.
 - Utiliza alguma métrica de similaridade para o cálculo.
- É dependente do número de vizinhos k .
 - Quando $k=1$, o elemento é classificado igualmente ao mais próximo.
 - Com maiores valores de k , a classificação é mais precisa.

Abordagens k -NN

- Novamente, cada item é um vetor de *features*.
 - Temos uma medida de sumarização da importância dessas *features*.
- Procura-se pelos k itens mais similares ao item i aos

$$\vec{u}_i = \frac{1}{|N_{ui}|} \sum_{j \in N_{ui}} r_{uj} \vec{j}$$

N_{ui} : itens vizinhos de i avaliados por u

r_{uj} : rating do usuário u para o item i

- $\text{Pr}_{sim}(\vec{u}, \vec{i})$

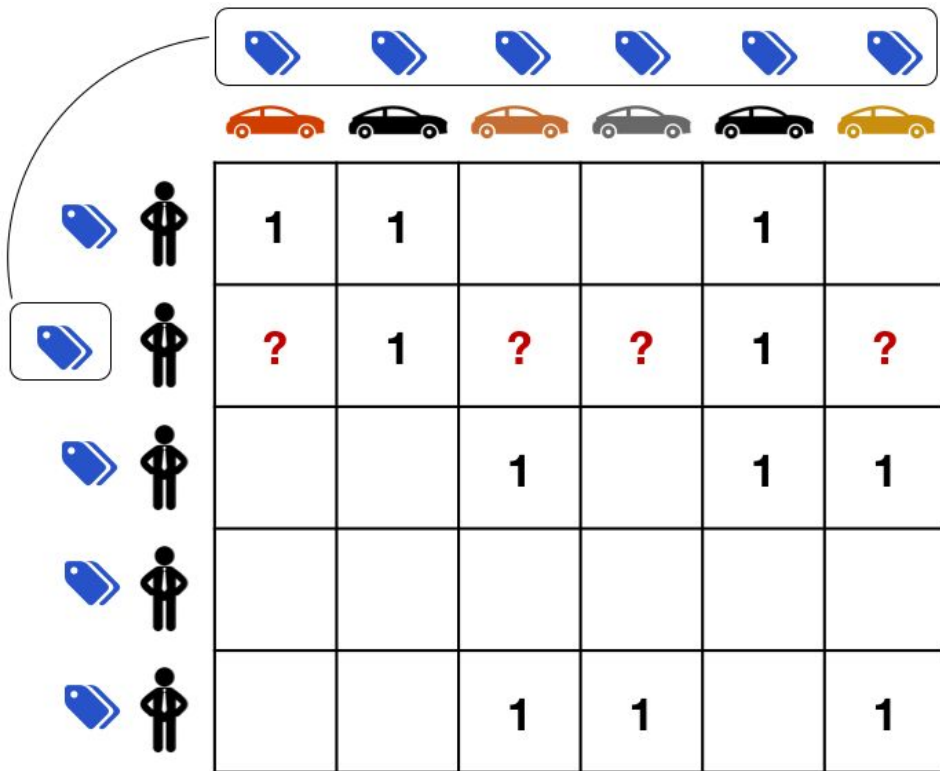
Vantagens & Desvantagens












- É uma abordagem comum em diversos recomendadores.
 - Abordagens de Filtragem Colaborativa.
- Mais efetiva que o modelo de Rocchio.
 - Consegue capturar as informações mais relevantes.
- Mas, ainda não é tão efetiva em cenários tradicionais.
 - Outras abordagens conseguem obter melhor desempenho.

Similaridade entre usuários e itens



Baseado em Conteúdo (CB)



						
	1	1			1	
	?	1	?	?	1	?
			1		1	1
						
			1	1		1

"Mostre-me mais do mesmo que gostei"

Métricas de Similaridade

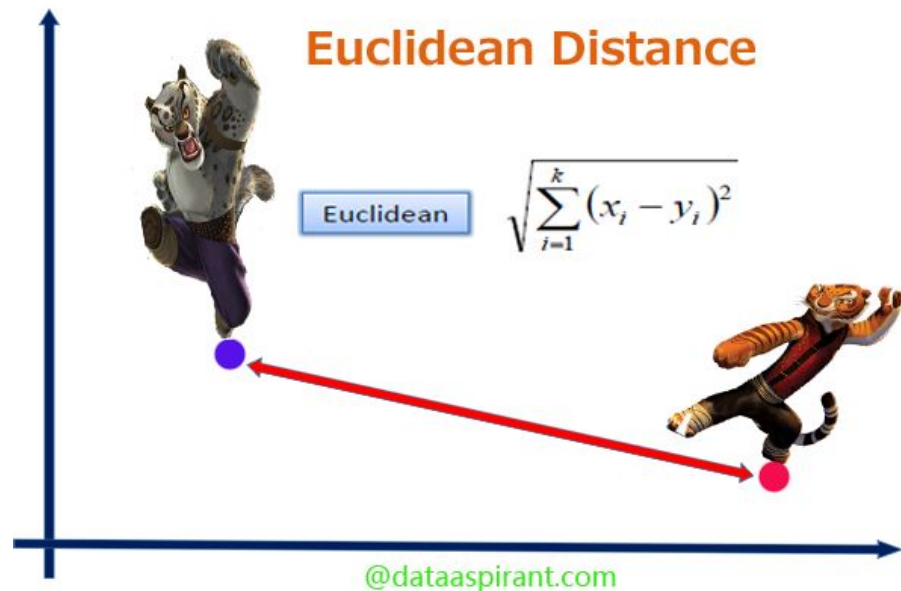
Considerando os usuários e itens como vetores, temos:

- Distância Euclidiana
- Distância de Manhattan
- Similaridade de Cosseno
- Similaridade de Jaccard

Distância Euclidiana

Métrica clássica:

- Diferença par a par de cada elemento do vetor.
- Utilizada quando os dados são densos e contínuos.

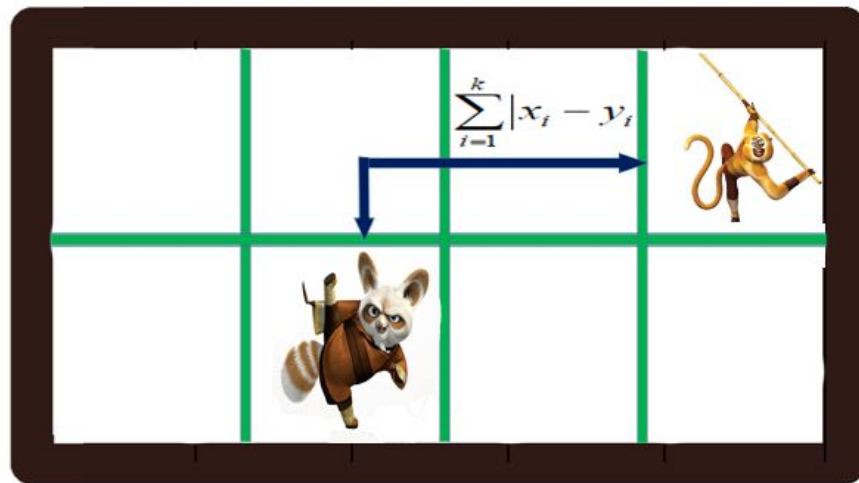


Distância de Manhattan

Baseada na cidade de
Manhattan:

- Soma das diferenças absolutas das coordenadas do vetor.

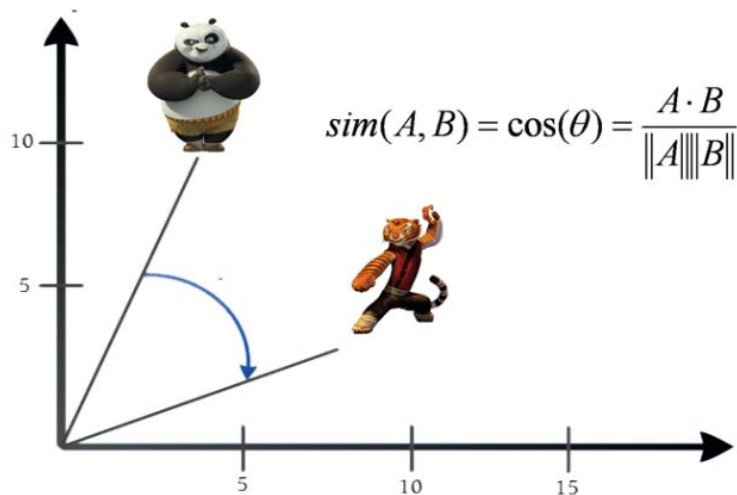
Manhattan Distance



@dataaspirant.com

Similaridade de Cosseno

Cosine Similarity



Métrica extremamente popular.

- Consiste no produto escalar normalizado dos vetores.
- Essa métrica captura a orientação do vetor, mas não sua magnitude.
- Muito usada para vetores esparsos.

Similaridade de Jaccard

- Métrica utilizada para avaliar os elementos do vetor.
- Desconsidera qualquer outra informação do vetor.

@dataaspirant.com

$$\text{Union}(A,B) = \left\{ \begin{array}{c} \text{Panda} \\ \text{Tigre} \\ \text{Macaco} \\ \text{Cavalo} \\ \text{Gato} \\ \text{Elefante} \\ \text{Galo} \end{array} \right\}$$
$$\text{Intersection}(A,B) = \left\{ \begin{array}{c} \text{Panda} \\ \text{Macaco} \end{array} \right\}$$

$|\text{Union}(A,B)| = 7$ $|\text{Intersection}(A,B)| = 2$

Ao aplicar estratégias de CB, devemos:

- Modelar os itens sobre suas características
- Modelar os usuários como um vetor agregado dessas características
- Computar a similaridade entre esses dois componentes