

Mineração de Dados

Aula 5 – parte 1

Especialização em Ciência de Dados e suas Aplicações

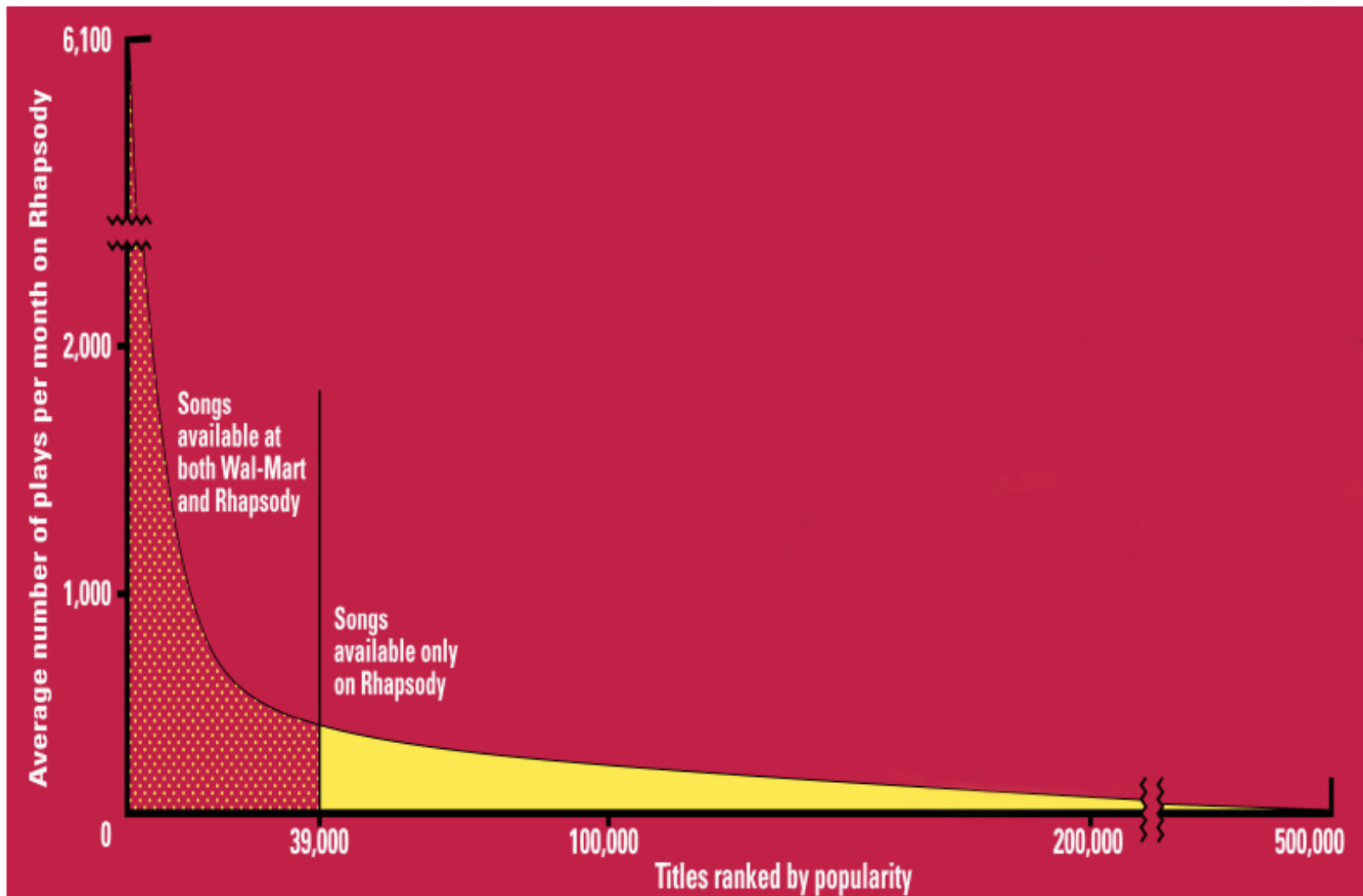
Espaço nas prateleiras é raro nas lojas

A web possibilita custo próximo de zero para disseminar informações sobre produtos

Mais opções demanda melhores filtros

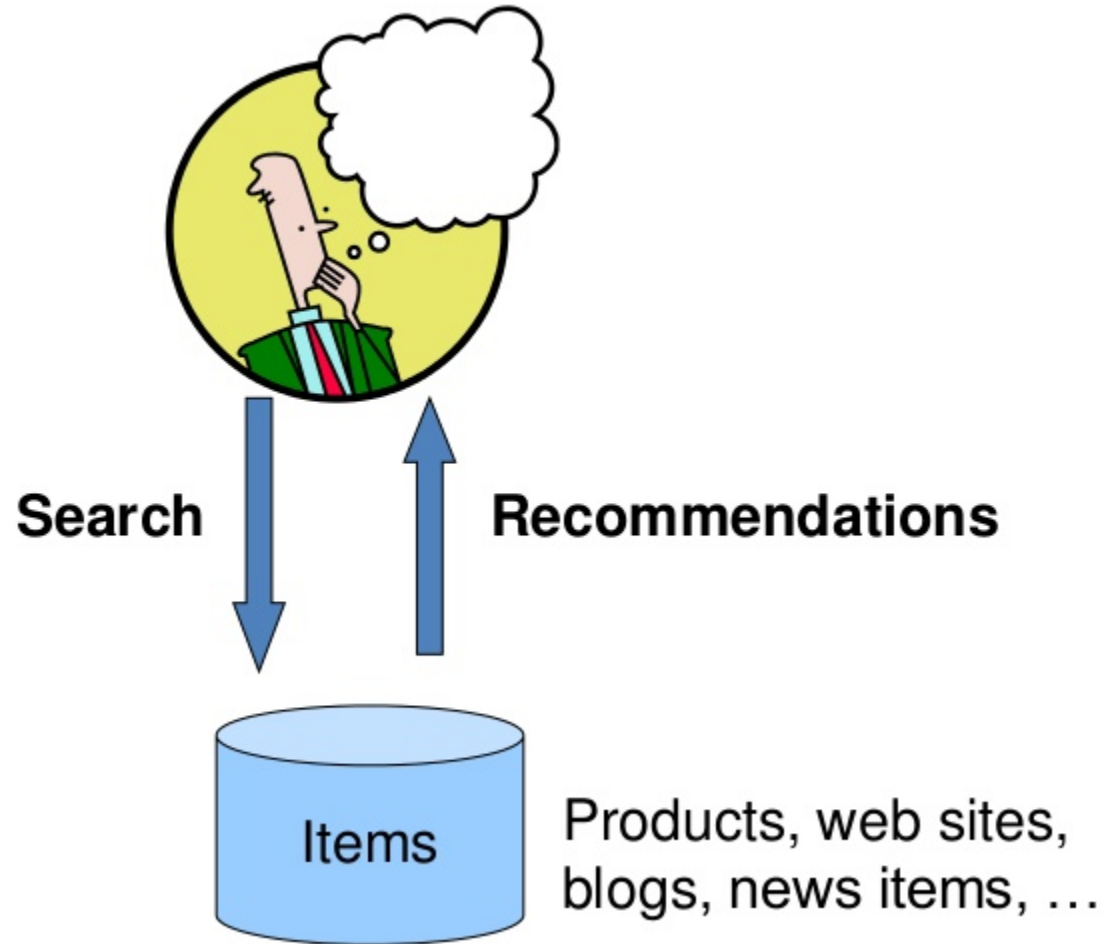
- Sistemas de recomendação

A cauda longa



Source: Chris Anderson (2004)

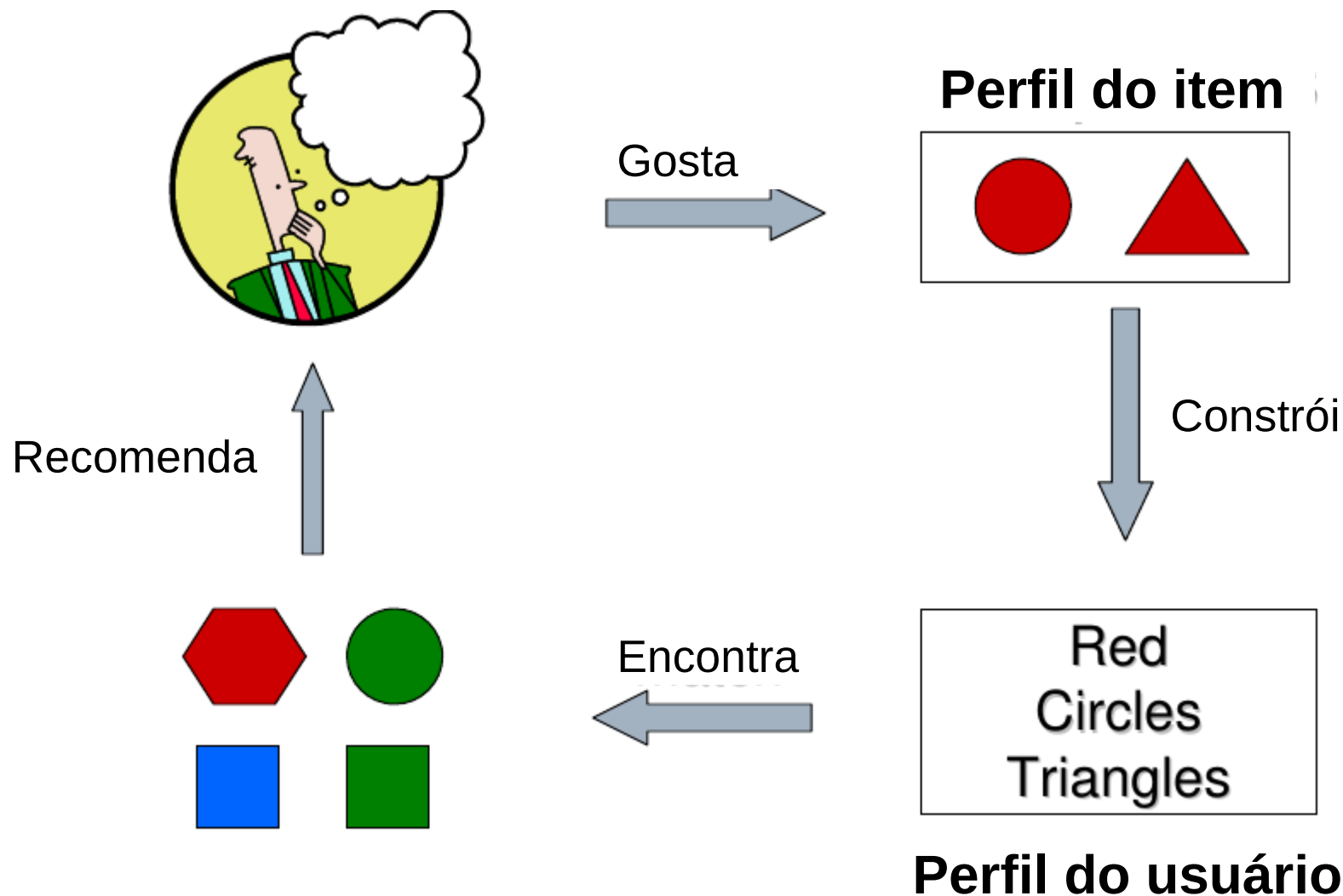
Sources: Erik Brynjolfsson and Jeffrey Hu, MIT, and Michael Smith, Carnegie Mellon; Barnes & Noble; Netflix; RealNetworks



Ideia principal: Recomendar itens ao usuário X similares a itens previamente bem avaliados por X

Exemplo:

- Filmes: mesmo gênero, diretor, atores...
- Pessoas: recomendar pessoas com muitos amigos em comum



Para cada item criar um perfil do item

Perfil é um conjunto de características (*features*)

- Filmes: ator, diretor, título...
- Imagens: metadados, tags...
- Pessoas: conjunto de amigos...

Apesar de ser um conjunto, é conveniente pensar no perfil do item como um vetor

- Uma entrada por característica (ex: cada ator, diretor...)
- Pode ser booleano (tem ou não) ou conter valores reais

Usuário avaliou os itens com perfis i_1, \dots, i_n

Simples: média (ponderada) dos perfis de itens avaliados



Pela avaliação dada em
cada item

Outra possibilidade: Normalizar os pesos usando a média das avaliações do usuário

Outros métodos também são possíveis

Exemplo 1

Itens são filmes

A única característica é “ator”

Perfil do item: vetor com 0 ou 1 para cada ator

Suponha que o usuário X assistiu 5 filmes:

- 2 filmes com o ator A
- 3 filmes com o ator B

Perfil do usuário = média do perfil dos itens

- peso de A = $2/5 = 0,4$
- peso de B = $3/5 = 0,6$

Exemplo 2

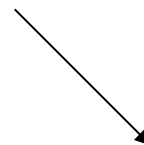
Mesmo exemplo anterior, mas agora o usuário pode dar estrelas (1 a 5)

- 2 filmes com o ator A (3 e 5 estrelas)
- 3 filmes com o ator B (1, 2 e 4 estrelas)



Avaliações ruins, negativas

Passo importante: Normalizar as avaliações subtraindo a média das avaliações do usuário (3 para o exemplo)



Normalização para o ator A: 0, +2

- Peso do perfil: $(0+2)/2 = 1$

Normalização para o ator B: -2, -1, +1

Peso do perfil: $-2/3$

Alguns usuários
podem ser mais
criteriosos do
que outros

De posse de um perfil de usuário e um perfil de item:

- Encontrar os itens mais similares

Uma boa métrica é a similaridade do cosseno

Recomendar os itens mais similares para o usuário

Não é necessário dados de outros usuários para fazer recomendação

Possibilita recomendar para usuários com gostos únicos

Possibilita fazer recomendações para um item novo e/ou não popular

- Não sofre com o problema “first-rater”, pois o perfil dos itens são baseados em características

Explicações para os usuários do porquê de uma recomendação

- Características que causaram a recomendação

Achar características apropriadas pode ser difícil

- Ex: música, pois pode ser difícil categorizar músicas em gêneros específicos

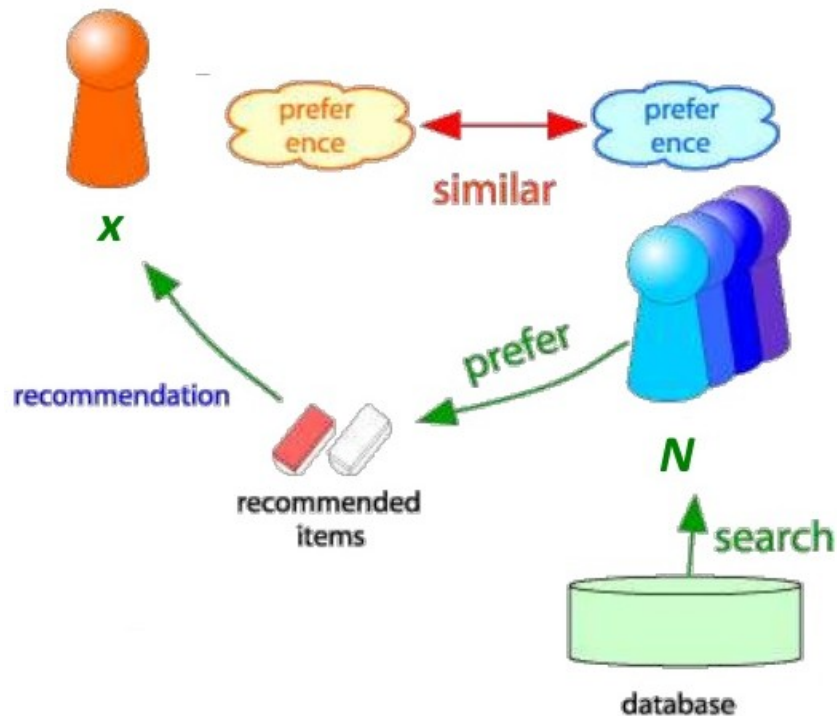
Superespecialização

- Nunca recomenda itens fora do perfil do usuário (depende do que o usuário já avaliou)
- Pessoas podem ter muitos interesses
- Não explora avaliações de outros usuários

Problema de Cold-start para novos usuários

- Como montar um perfil de usuário para um novo usuário?
Média de todos?

Filtragem colaborativa

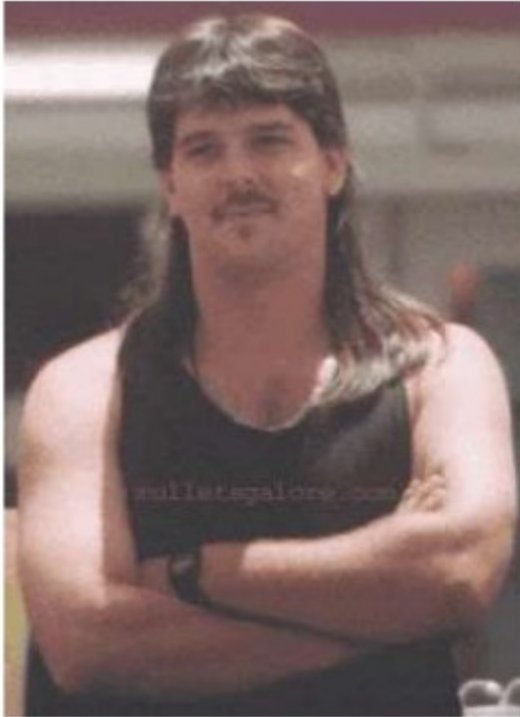


Considerando o usuário X

Achar N usuários que avaliaram filmes de forma **similar** a X

Estimar a avaliação de X baseado nas avaliações de usuários em N

Usuário X

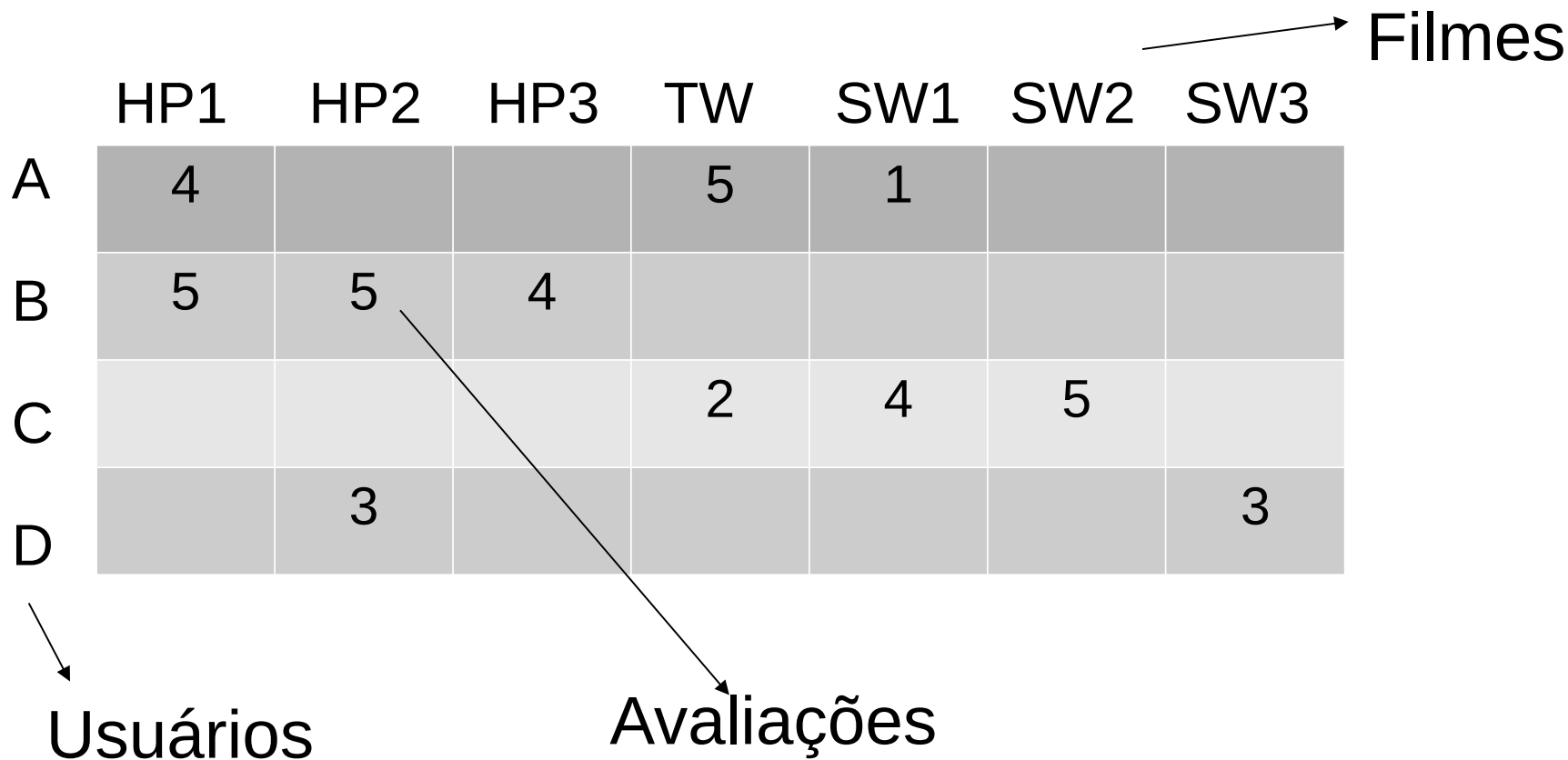


- Comprou um CD do metálica
- Comprou um CD do ACDC

Usuário Y



- Procura por ACDC
- Sistema de recomendação sugere Metálica a partir de dados coletados do usuário X



The diagram shows a 4x7 matrix of movie ratings. The rows are labeled A, B, C, and D, and the columns are labeled HP1, HP2, HP3, TW, SW1, SW2, and SW3. Arrows point from the labels 'Filmes' (Films), 'Usuários' (Users), and 'Avaliações' (Reviews) to their respective parts of the matrix. The cell at row B, column HP2 contains the value 5, and the cell at row D, column SW3 contains the value 3. The cell at row C, column HP3 is empty.

	HP1	HP2	HP3	TW	SW1	SW2	SW3
A	4			5	1		
B	5	5	4				
C				2	4	5	
D		3					3

Filmes

Usuários

Avaliações

	HP1	HP2	HP3	TW	SW1	SW2	SW3
A	4			5	1		
B	5	5	4				
C				2	4	5	
D		3					3

Considere os vetores de avaliações

Precisamos de uma métrica de similaridade $\text{sim}(X,Y)$

Que capture a intuição que $\text{sim}(A,B) > \text{sim}(A,C)$

	HP1	HP2	HP3	TW	SW1	SW2	SW3
A	1			1	1		
B	1	1	1				
C				1	1	1	
D		1					1

Considerar se o item foi avaliado ou não

Medida utilizada: distância de Jacquard $J(A, B) = \frac{|A \cap B|}{|A \cup B|}$

Jacquard-Dist(A,B) = 1/5 Não ok, pois A é mais similar a C do que a B

Jacquard-Dist(A,C) = 2/4

	HP1	HP2	HP3	TW	SW1	SW2	SW3
A	1			1	1		
B	1	1	1				
C				1	1	1	
D		1					1

Considerar se o item foi avaliado ou não

Medida utilizada: distância de Jacquard $J(A, B) = \frac{|A \cap B|}{|A \cup B|}$

Jacquard-Dist(A,B) = 1/5 Não ok, pois A é mais similar a C do que a B

Jacquard-Dist(A,C) = 2/4

Valores das avaliações são ignorados!

Similaridade do cosseno

	HP1	HP2	HP3	TW	SW1	SW2	SW3
A	4			5	1		
B	5	5	4				
C				2	4	5	
D		3					3

$\text{sim}(A,B) = \text{cosseno}(A,B)$

Vazios são tratados como zero

$\text{Cosseno}(A,B) = 0,380$

$\text{Cosseno}(A,C) = 0.322$

$\text{sim}(A,B) > \text{sim}(A,C)$, mas não muito.
Idealmente deveria ser maior

Similaridade do cosseno

	HP1	HP2	HP3	TW	SW1	SW2	SW3
A	4			5	1		
B	5	5	4				
C				2	4	5	
D		3					3

$$\text{sim}(A,B) = \text{cosseno}(A,B)$$

Vazios são tratados como zero

$$\text{Cosseno}(A,B) = 0,380$$

$$\text{Cosseno}(A,C) = 0,322$$

$\text{sim}(A,B) > \text{sim}(A,C)$, mas não muito.

Idealmente deveria ser maior

Trata valores faltantes como zero, o que é negativo

	HP1	HP2	HP3	TW	SW1	SW2	SW3
A	4			5	1		
B	5	5	4				
C				2	4	5	
D		3					3

Normalizar notas: subtraindo de cada nota de um usuário X, a média das notas de X

Média A = $10/3$

Média B = $14/3$

...

Cosseno centralizado

	HP1	HP2	HP3	TW	SW1	SW2	SW3
A	2/3			5/3	-7/3		
B	1/3	1/3	-2/3				
C				-5/3	1/3	4/3	
D		0					0

Centralizamos as avaliações de cada usuário em zero.

Zero é agora a nota média de cada usuário

$$\text{cosseno}(A,B) = 0,09$$

$$\text{cosseno}(A,C) = -0,56$$

$\text{sim}(A,B) > \text{sim}(A,C)$, e reflete melhor o que deveria

Cosseno centralizado

	HP1	HP2	HP3	TW	SW1	SW2	SW3
A	$2/3$			$5/3$	$-7/3$		
B	$1/3$	$1/3$	$-2/3$				
C				$-5/3$	$1/3$	$4/3$	
D		0					0

Centralizamos as avaliações de cada usuário em zero.

Zero é agora a nota média de cada usuário

Essa estratégia é interessante para lidar com usuários que são muito criteriosos e também com os pouco criteriosos por o que deveria

De posse de uma estratégia de similaridade

Como prever avaliação para um item i para o usuário X ?

De posse de uma estratégia de similaridade

Como prever avaliação para um item i para o usuário X ?

Encontrar k usuários mais similares a X que avaliaram o item i

Opção 1: média das avaliações do item i feita pelos k vizinhos

De posse de uma estratégia de similaridade

Como prever avaliação para um item i para o usuário X ?

Encontrar k usuários mais similares a X que avaliaram o item i

Opção 1: média das avaliações do item i feita pelos k vizinhos

Opção 2: média ponderada levando em consideração a similaridade dos usuários

$$\frac{\sum_{y \in N} s_{xy} \cdot r_{yi}}{\sum_{y \in N} s_{xy}}$$

K vizinhos \rightarrow $\sum_{y \in N}$
 $\text{sim}(x,y) \rightarrow s_{xy}$
Avaliação do usuário y para o item $i \rightarrow r_{yi}$

Até então utilizamos filtragem colaborativa usuário-usuário

Outra possibilidade: item-item

- Para cada item i , encontrar outros itens similares
- Estimar a avaliação com base na avaliação de itens similares
- Podemos usar as mesmas métricas de similaridade e função de predição do caso usuário-usuário

Filtragem colaborativa item-item

$|N| = 2$

Usuários

Filmes		1	2	3	4	5	6	7	8	9	10	11	12
	1	1		3			5			5		4	
	2			5	4			4			2	1	3
	3	2	4		1	2		3		4	3	5	
	4		2	4		5			4			2	
	5			4	3	4	2					2	5
6	1		3			3			2			4	

Filtragem colaborativa item-item

$|N| = 2$

Usuários

Filmes

	1	2	3	4	5	6	7	8	9	10	11	12
1	1		3		?	5			5		4	
2			5	4			4			2	1	3
3	2	4		1	2		3		4	3	5	
4		2	4		5			4			2	
5			4	3	4	2					2	5
6	1		3		3			2			4	

Filtragem colaborativa item-item

$|N| = 2$

		Usuários												Centered cosine	
		1	2	3	4	5	6	7	8	9	10	11	12	$\text{sim}(1,m)$	
Filmes	1	1		3		?	5			5		4		1.00	
	2			5	4			4			2	1	3	-0,18	
	3	2	4		1	2		3		4	3	5		<u>0,41</u>	
	4		2	4		5			4			2		-0,10	
	5			4	3	4	2					2	5	-0,31	
	6	1		3		3			2			4		<u>0,59</u>	

Seleção de vizinhos:

Identificar filmes similares ao filme 1, avaliados pelo usuário 5.

Filtragem colaborativa item-item

$|N| = 2$

		Usuários												Centered cosine	
		1	2	3	4	5	6	7	8	9	10	11	12	$\text{sim}(1,m)$	
Filmes	1	1		3		?	5			5		4		1.00	
	2			5	4			4			2	1	3	-0,18	
	3	2	4		1	2		3		4	3	5		<u>0,41</u>	
	4		2	4		5			4			2		-0,10	
	5			4	3	4	2					2	5	-0,31	
	6	1		3		3			2			4		<u>0,59</u>	

Pesos de similaridade: $S_{1,3}=0,41$ e $S_{1,6}=0,59$

Predição com a média ponderada:

$$R_{1,5} = (0,41 \cdot 2 + 0,59 \cdot 3) / (0,41 + 0,59) = 2,6$$

$$\frac{\sum_{y \in N} s_{xy} \cdot r_{yi}}{\sum_{y \in N} s_{xy}}$$

Na prática o resultado de item-item tende a ser melhor na maioria dos casos

Itens são mais “simples” do que usuários

-itens pertencem a um pequeno conjunto de “classes”,
usuários possuem gostos variados

Ex: uma música pode ser da classe MPB

Um usuário pode gostar de MPB e Indie Rock, mas
é mais raro uma música pertencer a várias classes.

-A similaridade de itens é mais significativa do que de
usuários

Avaliando sistemas de recomendação

movies

users

1	3	4			
	3	5			5
		4	5		5
		3			
		3			
2			2		2
				5	
	2	1			1
	3			3	
1					

Avaliando sistemas de recomendação

movies

users

1	3	4			
	3	5			5
		4	5		5
		3			
		3			
2			?		?
				?	
	2	1			?
	3			?	
1					

Conjunto de teste

Comparar as predições com o conjunto de teste (T)

Root-mean-square error (RMSE)

$$\sqrt{\frac{\sum_{(x,i) \in T} (r_{xi} - r_{xi}^*)^2}{N}}$$

Onde: $N = |T|$

r_{xi} é a avaliação predita

r_{xi}^* é a avaliação real

Foco apenas em acurácia as vezes perde o ponto principal:

- **Diversidade de predição** (é interessante apresentar itens diversos que o usuário pode gostar, não só “CDs do ACDC”)
- **Contexto da predição** (um usuário que fez uma viagem pra Paris comprou vários itens sobre o local, mas quando ele volta pode não ter mais interesse no conteúdo)

Na prática, nos importamos apenas em prever boas avaliações

RMSE pode penalizar um método que tem bom desempenho para boas avaliações e desempenho ruim para outras

Alternativa: precisão dos top k

Anand Rajaraman, Jure Leskovec, Jeffrey D. Ullman - Mining Massive Datasets, Cambridge University Press, 2012. – Capitulo 9