CeMEAI
CEPID - Centro de Ciências Matemáticas Aplicadas à Indústria

ICMC USP SÃO CARLOS

Algoritmos baseados em filtragem baseada em conteúdo

Prof. Dr. Marcelo G. Manzato e Arthur Fortes da Costa

Filtragem baseada em conteúdo (FBC)

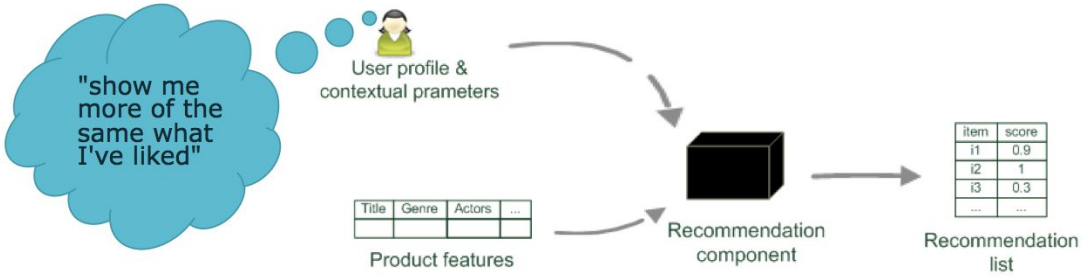
FC não utiliza nenhuma informação sobre os itens

- Apenas interações de usuários
- Problemas de cold-start e esparsidade

FBC calcula recomendações utilizando:

- Descrições sobre os itens (metadados)
- Perfil de usuário contendo o que ele gosta / não gosta

Filtragem baseada em conteúdo (FBC)



Representação de itens

- Utilização de **metadados** estruturados, semi-estruturados e/ou não-estruturados

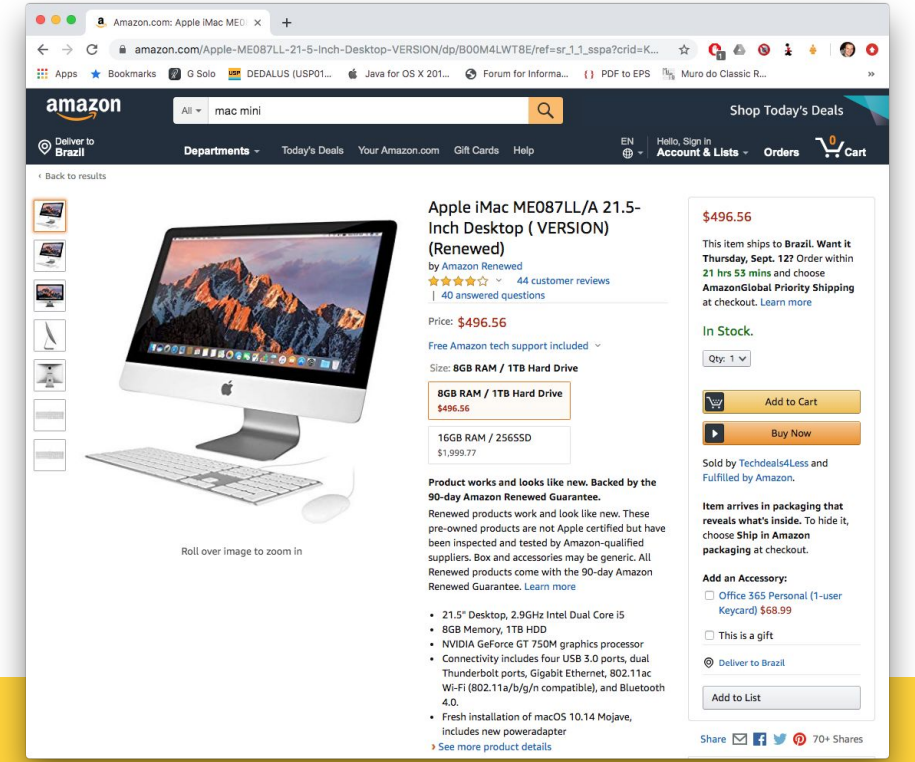
Title	Genre	Author	Type	Price	Keywords
The Night of the Gun	Memoir	David Carr	Paperback	29.90	Press and journalism, drug addiction, personal memoirs, New York
The Lace Reader	Fiction, Mystery	Brunonia Barry	Hardcover	49.90	American contemporary fiction, detective, historical
Into the Fire	Romance, Suspense	Suzanne Brockmann	Hardcover	45.90	American fiction, murder, neo-Nazism

Metadados estruturados, semi-estruturados e não-estruturados

No caso de metadados não estruturados, é necessário realizar um pré-processamento do texto, que pode conter:

- Tokenização
- Remoção de stop words
- Normalização de termos
- Lematização
- Radicalização
- Desambiguação
- TFIDF

Exemplo


Amazon.com: Apple iMac ME0
amazon.com/Apple-ME087LL-21-5-Inch-Desktop-VERSION/dp/B00M4LWT8E/ref=sr_1_1_sspa?crid=K...
Apps
Bookmarks
G Solo
DEDALUS (USP01...
Java for OS X 201...
Forum for Informa...
PDF to EPS
Muro do Classic R...
amazon
All
mac mini
Shop Today's Deals
Deliver to Brazil
Departments
Today's Deals
Your Amazon.com
Gift Cards
Help
EN
Hello, Sign in
Account & Lists
Orders
Cart
Back to results
Roll over image to zoom in
Apple iMac ME087LL/A 21.5-Inch Desktop (VERSION) (Renewed)
by Amazon Renewed
44 customer reviews
40 answered questions
Price: $496.56
Free Amazon tech support included
Size: 8GB RAM / 1TB Hard Drive
8GB RAM / 1TB Hard Drive
$496.56
16GB RAM / 256SSD
$1,999.77
Product works and looks like new. Backed by the 90-day Amazon Renewed Guarantee.
Renewed products work and look like new. These pre-owned products are not Apple certified but have been inspected and tested by Amazon-qualified suppliers. Box and accessories may be generic. All Renewed products come with the 90-day Amazon Renewed Guarantee. Learn more
21.5" Desktop, 2.9GHz Intel Dual Core i5
8GB Memory, 1TB HDD
NVIDIA GeForce GT 750M graphics processor
Connectivity includes four USB 3.0 ports, dual Thunderbolt ports, Gigabit Ethernet, 802.11ac Wi-Fi (802.11a/b/g/n compatible), and Bluetooth 4.0.
Fresh installation of macOS 10.14 Mojave, includes new poweradapter
See more product details
$496.56
This item ships to Brazil. Want it Thursday, Sept. 12? Order within 21 hrs 53 mins and choose AmazonGlobal Priority Shipping at checkout. Learn more
In Stock.
Qty: 1
Add to Cart
Buy Now
Sold by Techdeals4Less and Fulfilled by Amazon.
Item arrives in packaging that reveals what's inside. To hide it, choose Ship in Amazon packaging at checkout.
Add an Accessory:
Office 365 Personal (1-user Keycard) $68.99
This is a gift
Deliver to Brazil
Add to List
Share
70+ Shares

Item Attribute KNN

Jessica	?	2	4	3	2	3
Marta	4	3	?	4	3	2
Jose	1	5	3	4	?	5
Dave	1	?	2	3	4	?

Drama	1	0	1	1	0	0
Ação	0	1	0	1	1	0
Comédia	1	1	0	0	0	1
Sci-Fi	0	0	1	0	1	0

Similaridade baseada em metadados
Usar Cosseno, Pearson, Jaccard, etc.

Predição de nota

$$pred(u,i) = \frac{\sum_{j \in I_{ui}} sim(i,j) * r_{uj}}{\sum_{j \in I_{ui}} sim(i,j)}$$

Ranqueamento

$$pred(u,i) = \sum_{j \in I_{ui}} sim(i,j)$$

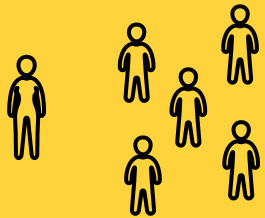
FBC baseada em KNN (FBC-kNN)

Variações

- Alterar o tamanho de vizinhos
- Utilizar limiares superiores e inferiores para similaridade

Vantagens

- Bom para modelar interesses de curtíssimo-prazo (sessão)
- Pode ser usado em combinação com outros métodos para modelar preferências de longo-prazo.



Limitações

- Palavras-chave (características) podem não ser suficientes para julgar a qualidade ou relevância de um item;
- Falta de descrição, semântica, características não textuais, etc;
- Problema do novo usuário;
- Sobre-especialização.