

Algoritmos baseados em filtragem baseada em conteúdo

Prof. Dr. Marcelo G. Manzato



Filtragem baseada em conteúdo (FBC)

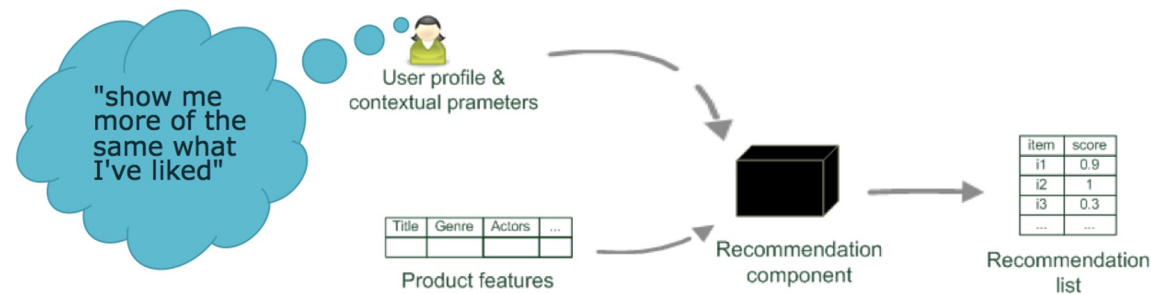
FC não utiliza nenhuma informação sobre os itens

- 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



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.



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

Predição de nota

Ranqueamento

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

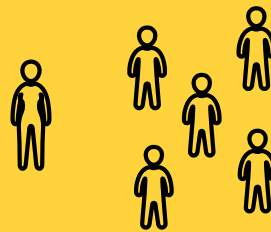
FBC baseada em KNN (FBC-kNN)

Variações

- Representações dos itens baseadas em diferentes características
- Alterar o número de vizinhança

Vantagens

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



Limitações

- 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.