



Social Search

Prof. Luciano Barbosa

(Parte do material retirado dos slides dos livros adotados)



Social Search

- Comunidades de usuários ativamente participando no processo de busca
- Vai além das tarefas clássicas de busca
- Usuários interagem com outros usuários implicitamente ou explicitamente



Web 2.0

- Sites de mídia social
- Conteúdo gerado por usuário
- Usuário pode criar tags
- Ex:
 - Digg, Twitter, Flickr, Facebook, Instagram etc



Tópicos em Social Search

- Tags dos usuários
- Busca dentro de comunidades
- Sistemas de recomendação



Tags de Usuários

- Antes: cartões de biblioteca
 - Termos escolhidos com a busca em mente
 - Especialistas geravam esses termos
 - Termos de alta qualidade (bem explicativos) e escolhidos de um vocabulário controlado

QA76	Clark, Michael
.55	Cultural treasures of the Internet / M. Clark. -- Upper
.C58	Saddle River, N.J. : Prentice Hall, 1995.
1995	xxix, 313 p. : ill. ; 24 cm.
	ISBN 0132096692.
	1. Computers--Cultural Impact. 2. Internet--Cultural Impact.



Tags de Usuários

- Atualmente: Tags
 - Nem sempre criadas pensando em busca
 - Geradas pelos usuários
 - Podem ser escritas de forma não-padrão
 - Chamado de folksonomy





Tipos de Tags

- Baseada em conteúdo
 - Ex: carro, casa
- Baseada em contexto
 - Marco zero, praia
- Atributo
 - Nikon, nacional (filme)
- Subjetividade
 - Bom, ruim etc
- Organizacional
 - Minhas figuras, a fazer



Busca em Tags

- Desafios:
 - A maioria dos itens possui poucas tags
 - Tags são bem curtas

Updated Tips #6: Find a few unique hashtags

Ovoid cramming your posts full of spammy hashtags. Finding a few unique hashtags can be more beneficial than spamming your photo with every basic hashtag you can think of. Most big players hardly use any hashtags and let the content do the talking for them. Accounts with a zillion hashtags tend to look desperate.



Busca em Tags

- Modelos tradicionais de RI não funcionam bem
- Tem que lidar com a diferença entre o vocabulário da consulta e da tag
- Solução: expansão de consulta usando logs de consulta, thesaurus etc



Community Based Question Answering: CBQA

- Engenheiros de busca não conseguem responder consultas mais complexas
 - Informações de muitas fontes
 - Precisa do expertise humano
- CBQA
 - Usuário submete uma consulta
 - Membros da comunidade respondem



Examples de Perguntas

What part of Mexico gets the most tropical storms?

How do you pronounce the french words, coeur and miel?

GED test?

Why do I have to pay this fine?

What is Schrödinger's cat?

What's this song?

Hi...can u ppl tell me sumthing abt death dreams??

What are the engagement and wedding traditions in Egypt?

Fun things to do in LA?

What lessons from the Tao Te Ching do you apply to your everyday life?

Foci of a hyperbola?

What should I do today?

Why was iTunes deleted from my computer?

Heather Locklear?

Do people in the Australian Defense Force (RAAF) pay less tax than civilians?

Whats a psp xmb?

If $C(-3, y)$ and $D(1, 7)$ lie upon a line whose slope is 2, find the value of y ?

Why does love make us so irrational?

Am I in love?

What are some technologies that are revolutionizing business?



CBQA

- Vantagens
 - Encontra respostas para perguntas complexas
 - Respondidas por humanos e não algoritmos
- Desvantagens
 - Geralmente leva tempo para conseguir uma resposta
 - Algumas questões nunca são respondidas
 - Respostas podem estar erradas

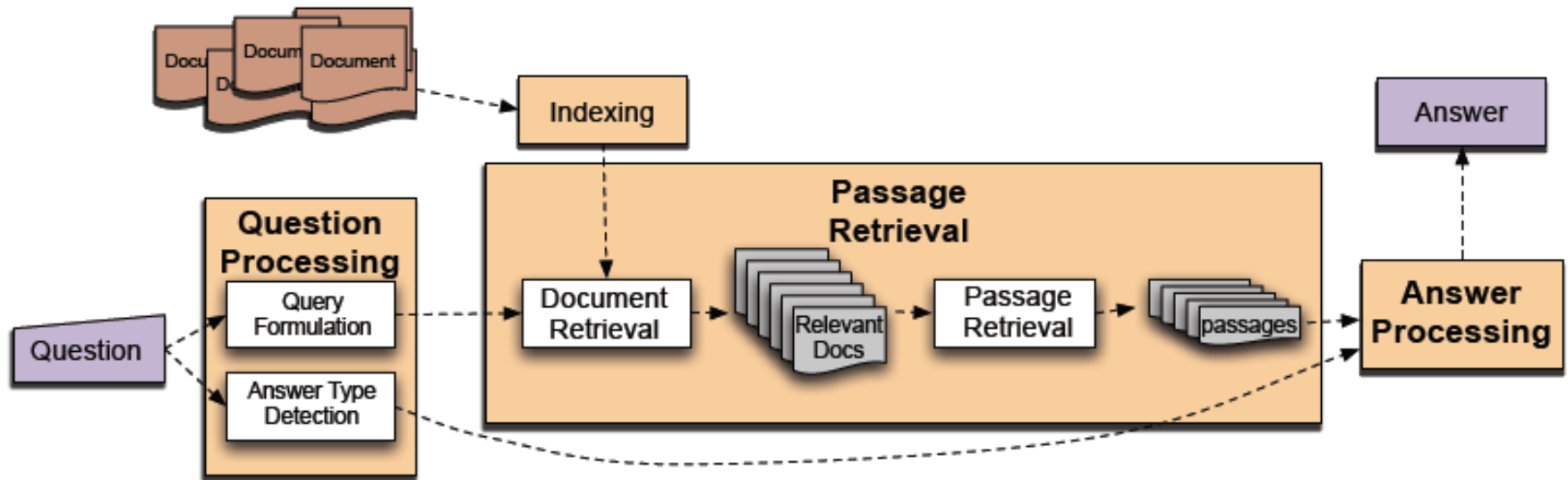


Tipo de Perguntas

- Factóides
 - Quem escreveu a bíblia?
 - Quantas calorias tem um cafezinho?
 - Onde está o Museu do Frevo?
- Complexas
 - Qual a eficácia da aspirina para baixar a febre?
 - O que cientistas pensam do aquecimento global?



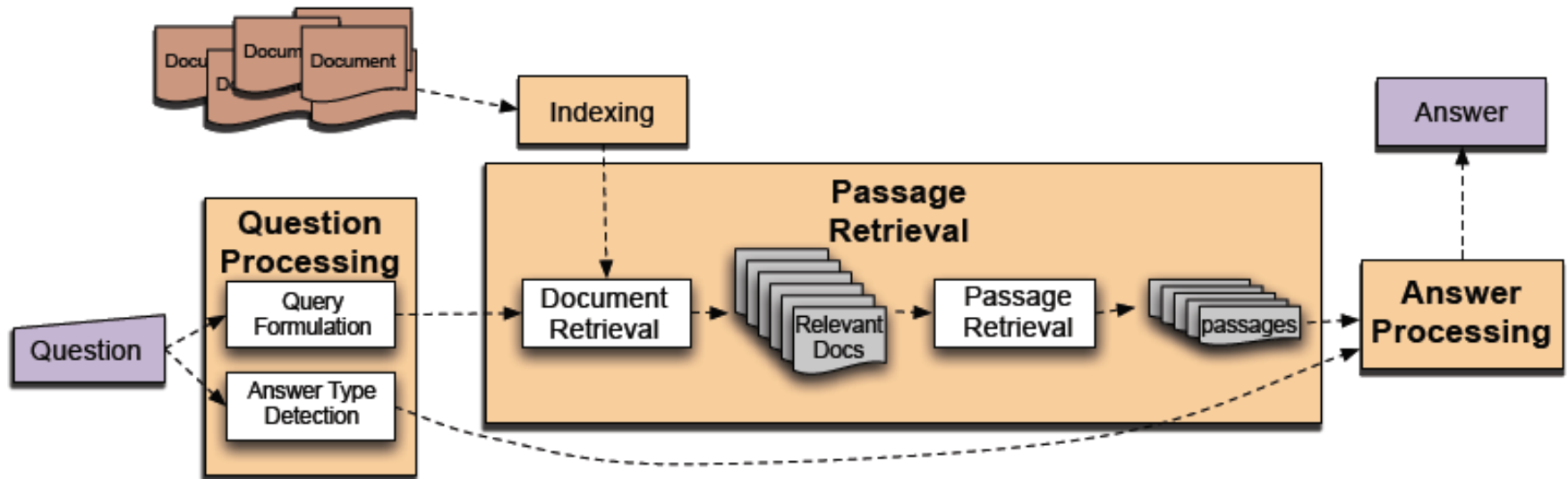
Sistema de Q&A Baseado em RI



- Question processing
 - Detecta tipo da resposta
 - Resposta sobre pessoas, animais, data etc
 - Ex: Quem fundou Brasília?
 - Detecta tipo das perguntas
 - Definição de algo, pergunta matemática, lista etc
 - Constrói consultas para enviar para um engenho de busca



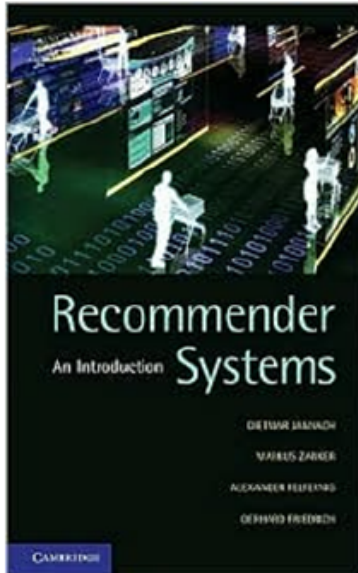
Sistema de Q&A Baseado em RI



- **Passage retrieval**
 - Recupera os documentos ranqueados
 - Quebra em parágrafos
 - Reordena baseado no tipo da resposta (usa um NER)
- **Answer processing**
 - Extraí respostas candidatas
 - Ranqueia candidatos



Sistemas de Recomendação



Recommender Systems: An Introduction

by [Dietmar Jannach](#), [Markus Zanker](#), [Alexander Felfernig](#), [Gerhard Friedrich](#)

AVERAGE CUSTOMER RATING:

★★★★★ ([Be the first to review](#))



Registrieren, um sehen zu können, was deinen Freunden gefällt.

FORMAT:
Hardcover

NOOKbook (eBook) - not available

[Tell the publisher you want this in NOOKbook format](#)

NEW FROM BN.COM

~~\$65.00~~ List Price

\$52.00 Online Price
(You Save 20%)

Add to Cart

NEW & USED FROM OUR

New starting at **\$56.46** (You Save 13%)
Used starting at **\$51.98** (You Save 20%)

See All Prices

[Table of Contents](#)

Customers who bought this also bought





Outros Domínios

You may also like



Jack & Jones
JAMIE - Polo shirt - orange
£21.00

Free delivery & returns

ALTERNATIVE PRODUCTS

Beko Washing Machine

Code: WMB81431LW

£269.99

Zanussi Washing Machine

Code: ZWH6130P

£269.99

Blomberg Washing Machine

Code: WNF6221

£299.99

Related hotels...



Hotel 41

1,170 Reviews

London, England

Show Prices

Read

Commented

Recommended



Germany Just Rejected The
Idea That The European Bailout
Fund Would Buy Spanish Debt

✕



There Is Almost No Gold In The
Olympic Gold Medal

✕

You may also like



★★★★☆ (109)



★★★★☆ (53)



★★★★☆ (33)

MOST POPULAR

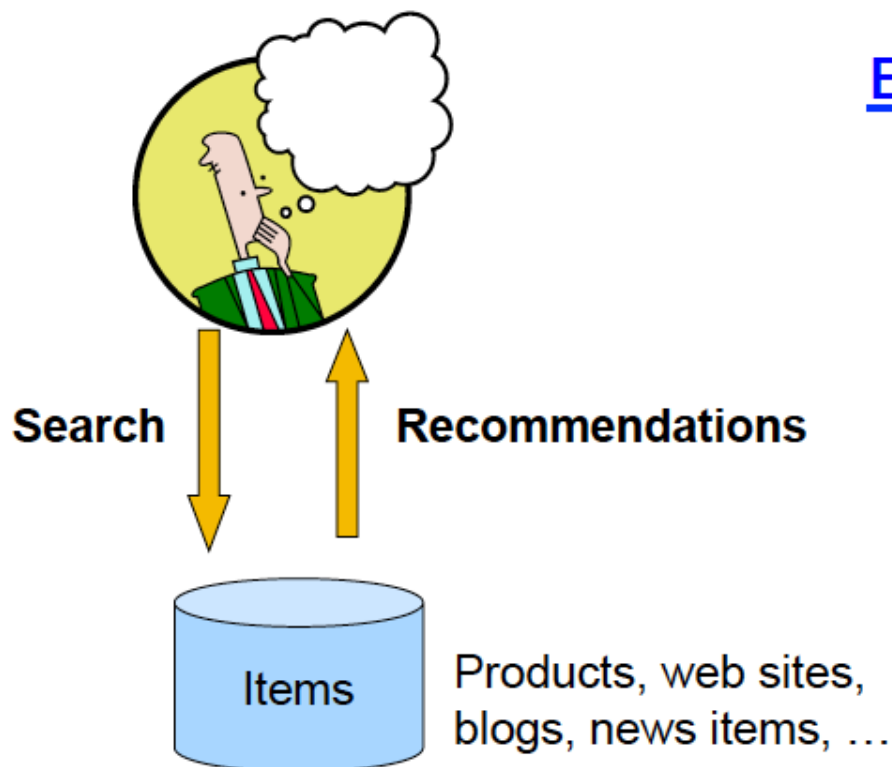
RECOMMENDED

How to Break NRA's Grip on Politics: Michael R. Bloomberg

Growth in U.S. Slows as Consumers Restrain Spending



Visão Geral



Examples:

amazon.com.



StumbleUpon



del.icio.us



m o v i e l e n s
helping you find the right movies

last.fm™
the social music revolution

Google
News

You Tube

XBOX
LIVE



Motivação

- Usuário
 - Encontrar itens interessantes
 - Diminuir o espaço de busca
 - Descobrir novos itens
- Provedor
 - Aumentar fidelidade do usuário
 - Aumentar vendas
 - Obter maior conhecimento do usuário



Definição

- Entrada:
 - Modelo do usuário: preferências, localização etc
 - Itens
- Objetivo:
 - Ranquear itens de acordo com a relevância para o usuário -> apresentar os (mais) relevantes pro usuário

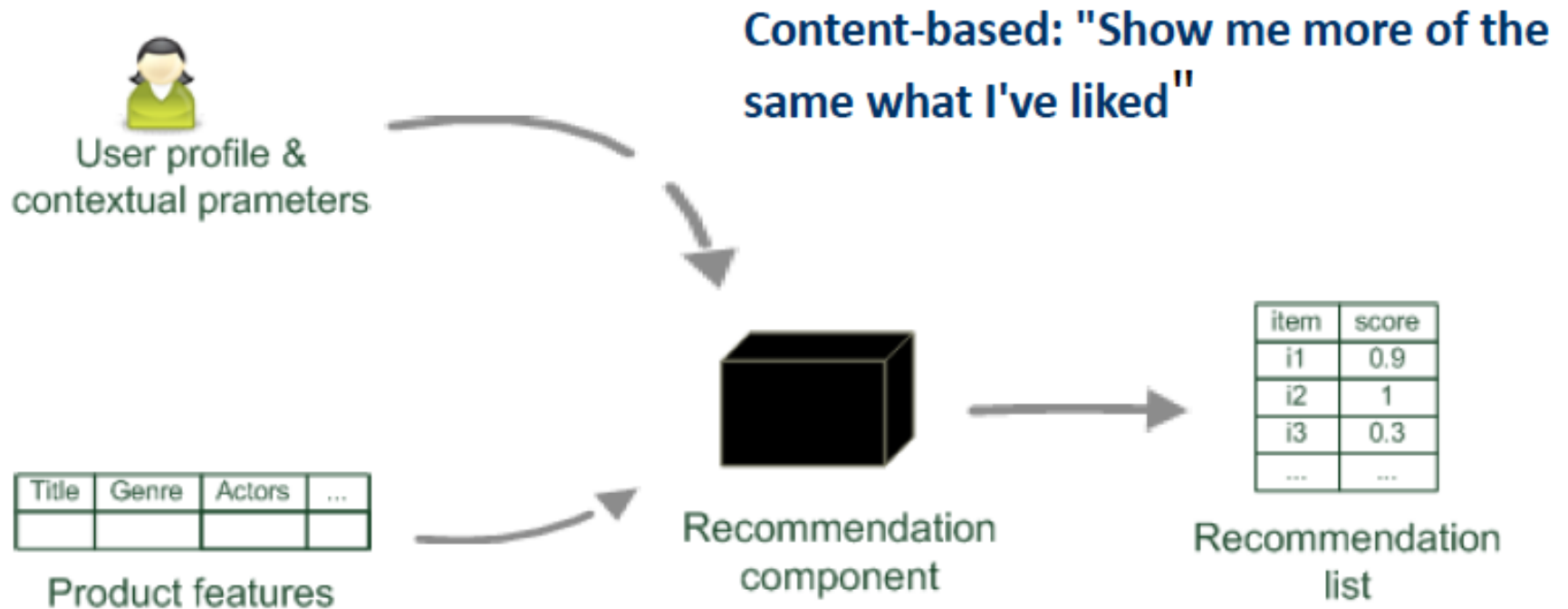


Tipos

- Baseado em conteúdo
- Filtragem colaborativa



Baseado em Conteúdo





Baseado em Conteúdo

- Recomendar itens para um usuário baseado no seu perfil e dos itens: “mostre-me o mesmo daquilo que gosto”
- Exemplo
 - Recomendar filmes com um mesmo ator
 - Recomendar outros sites que tenham conteúdo parecido

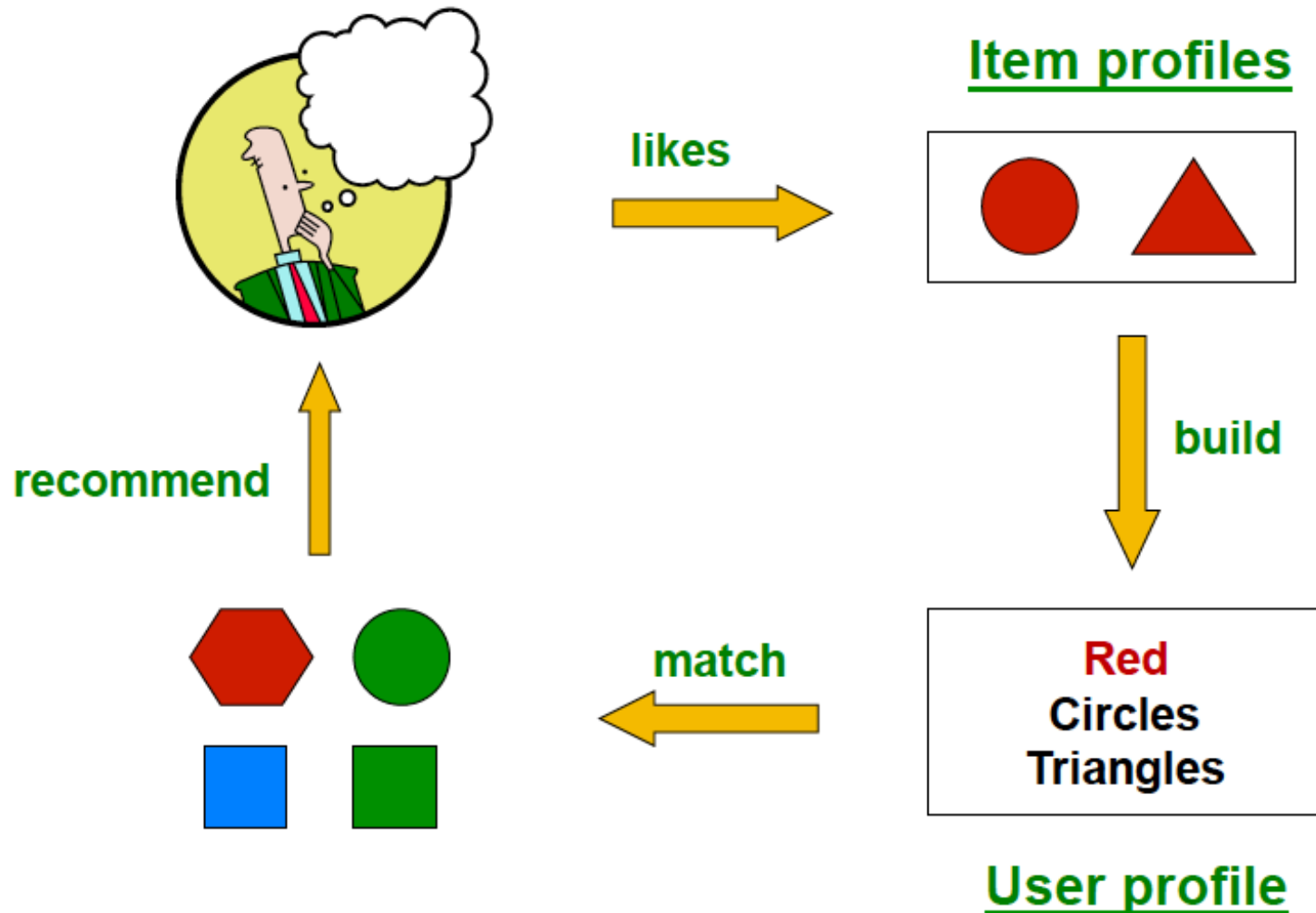


Baseado em Conteúdo

- Perfil de características para um item
 - Filmes: título, gênero, atores etc
 - Documentos: palavras mais importantes (tfidf)
- Perfil do usuário: descrever preferências do usuário



Exemplo





Baseado em Conteúdo

- Métodos:
 - Similaridade entre perfil do item e do usuário
 - Classificador: features dos itens + avaliações do usuários como rótulos



Vantagens

- Não precisa de dados de outros usuários
- Capaz de recomendar itens novos e não populares (não precisa que já tenha sido avaliado)
- Capaz de recomendar para usuários com gosto específico
- Capaz de explicar pelas features usadas

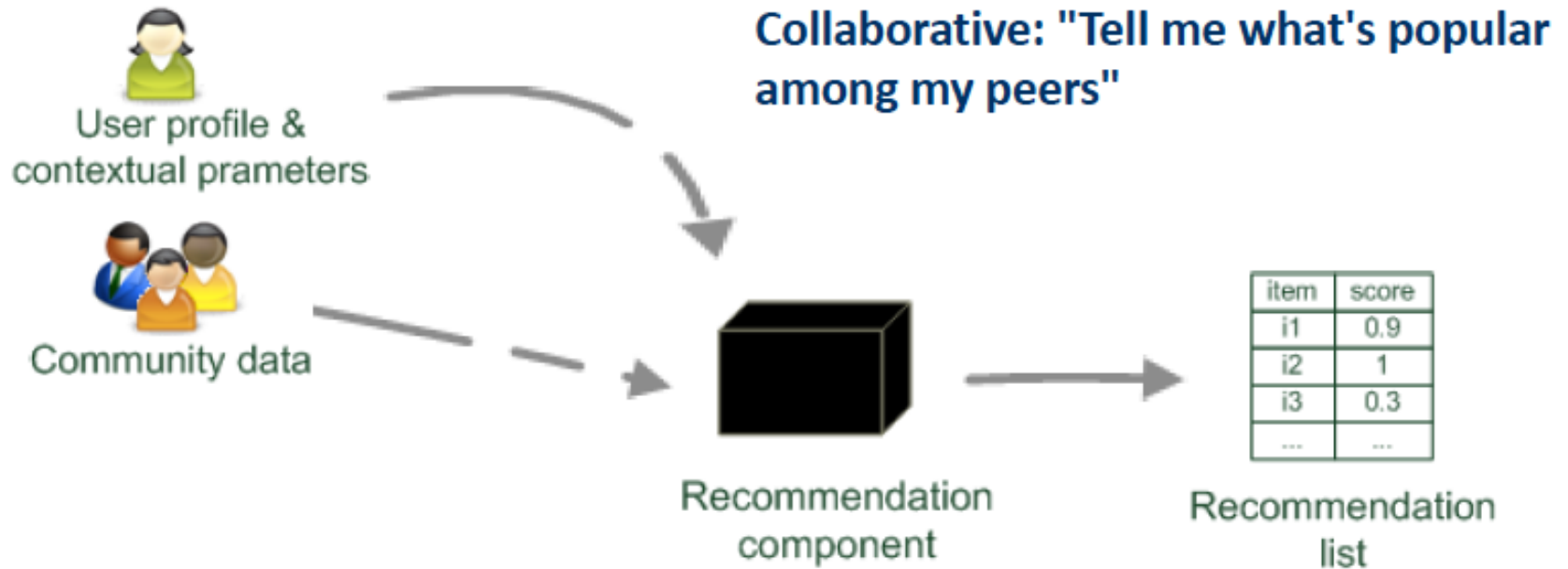


Desvantagens

- Pode ser difícil encontrar as features apropriadas
- Difícil de gerar recomendações para novos usuários
- Super-especializado
 - Nunca recomenda itens fora do perfil do usuário (não explora)
 - Não explora julgamento de outros usuários



Filtragem Colaborativa





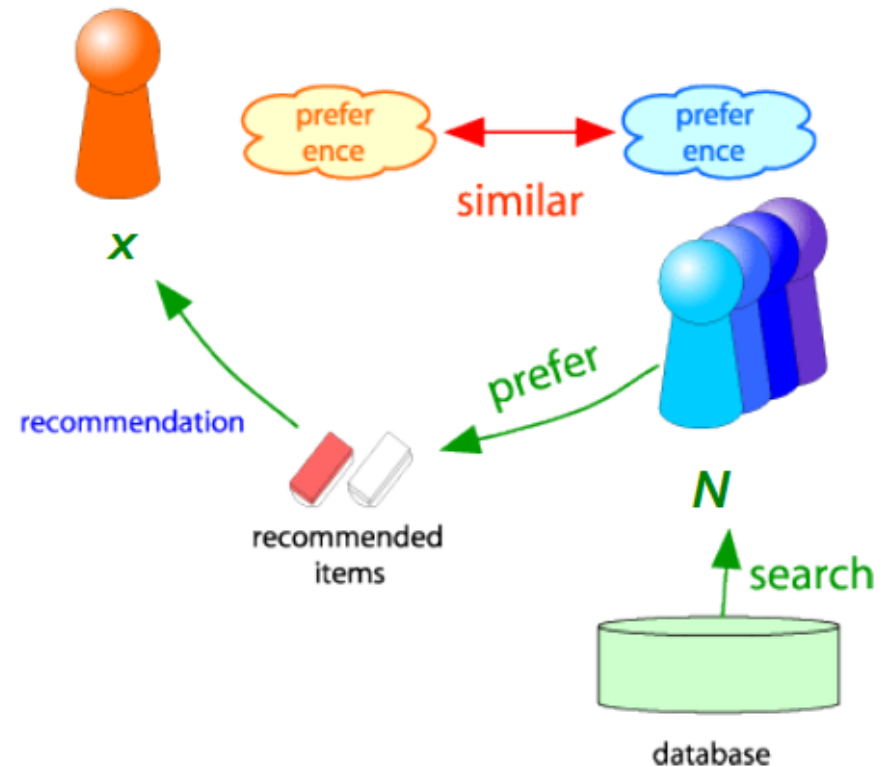
Filtragem Colaborativa

- Abordagem mais utilizada para gerar recomendações
- Usa o conhecimento das preferências dos usuários
- Suposição básica: usuários com preferências parecidas a um dado usuário são usados para inferir recomendações para ele



Ideia Básica

- Dado um usuário x
- Encontrar usuários N com preferências parecidas a x
- Estimar preferências de x para novos itens dado as preferências em N





Matrix de Utilidade

- Duas classes de entidades: usuários e itens
- Representa o par usuário-item com a preferência do usuário

	Item1	Item2	Item3	Item4	Item5
Alice	5	3	4	4	?
User1	3	1	2	3	3
User2	4	3	4	3	5
User3	3	3	1	5	4
User4	1	5	5	2	1



Matrix de Utilidade

- Matrix esparsa
- Populando células vazias
 - Pedi para os próprios usuários
 - Prever a partir de outros



Filtragem Colaborativa (Usuário-Usuário)

- Abordagem básica: vizinhos mais próximos
 - Dado um usuário u e um item i ainda não classificado por u
 - Encontrar conjunto de usuários que tenham classificado outros itens de forma similar a u
 - Combinar preferências de i para prever preferência de u por i



Filtragem Colaborativa: 3 Questões

- Como medir similaridade
- Quantos vizinhos
- Como calcular preferência a partir de outros



Medidas de Similaridade

- Cada usuário representado por um vetor de suas preferências
- Correlação de pearson

$$sim(a, b) = \frac{\sum_{p \in P} (r_{a,p} - \bar{r}_a)(r_{b,p} - \bar{r}_b)}{\sqrt{\sum_{p \in P} (r_{a,p} - \bar{r}_a)^2} \sqrt{\sum_{p \in P} (r_{b,p} - \bar{r}_b)^2}}$$

a, b : usuários

$r_{a,p}$: notas do usuário a para o item p

$r_{b,p}$: notas do usuário b para o item p

P : conjunto de itens, avaliados por a e b

\bar{r}_a, \bar{r}_b : notas médias dos usuários



Exemplo Correlação de Pearson

	Item1	Item2	Item3	Item4	Item5
Alice	5	3	4	4	?
User1	3	1	2	3	3
User2	4	3	4	3	5
User3	3	3	1	5	4
User4	1	5	5	2	1

- Máximo: $\text{cor}(\text{Alice}, \text{User1}) = 0,85$
- Mínimo: $\text{cor}(\text{Alice}, \text{User4}) = -0,79$



Calculando as Notas

- Média das avaliações dos mais próximos
- Baseada na similaridade

$$pred(a, p) = \bar{r}_a + \frac{\sum_{b \in N} sim(a, b) * (r_{b,p} - \bar{r}_b)}{\sum_{b \in N} sim(a, b)}$$

- Combina as diferenças de preferência
- Usa similaridade como peso
- Usuários mais similares têm mais peso



Filtragem Colaborativa (Item-Item)

- Similar ao usuário-usuário
- Encontram-se itens similares ao item dado
- Calcula-se preferência baseado nos vizinhos mais próximos
- Na prática item-item funciona melhor
- Itens são mais simples que usuários



Exemplo

users

movies

	1	2	3	4	5	6	7	8	9	10	11	12	$\text{sim}(1,m)$
1	1		3		?	5			5		4		1.00
2			5	4			4			2	1	3	-0.18
<u>3</u>	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
<u>6</u>	1		3		3			2			4		<u>0.59</u>



Exemplo: N=2

		users											
		1	2	3	4	5	6	7	8	9	10	11	12
movies	1	1		3		2.6	5			5		4	
	2			5	4			4			2	1	3
	<u>3</u>	2	4		1	2		3		4	3	5	
	4		2	4		5			4			2	
	5			4	3	4	2					2	5
	<u>6</u>	1		3		3			2			4	

$$r_{1.5} = (0.41 \cdot 2 + 0.59 \cdot 3) / (0.41 + 0.59) = 2.6$$



Filtragem Colaborativa: Pros/Cons

- Pros: não precisa de features de itens e usuários
- Cons:
 - Precisa de usuários suficientes para encontrar similares
 - Difícil encontrar usuários que avaliam mesmos itens
 - Não pode recomendar um item recém-criado
 - Tende a sugerir itens populares
 - Solução: método híbrido



Redução de Dimensionalidade

- Problemas com filtragem colaborativa
 - Esparsividade da matriz de utilidade
 - Escalabilidade: problema com matrizes grandes
 - “Sinônimos”: itens diferentes podem representar objetos semelhantes



Redução de Dimensionalidade

- Captura fatores importantes escondidos nos dados
- Suposição: menores dimensões capturam sinais e removem ruído



Latent Semantic Indexing

- Objetivo: mapear usuários e itens a um espaço dimensional de conceitos
- M: matrix usuário-item
- Pode ser decomposta usando SVD

$$M = U \times \Sigma \times V^T$$






- Considerar somente os maiores valores singulares em Σ
- Calcular: $M_k = U_k \times \Sigma_k \times V_k^T$
- k é o valor de redução de dimensionalidade



Exemplo: K=2

$$M_k = U_k \times \Sigma_k \times V_k^T$$

U_k	Dim1	Dim2
Alice	0.47	-0.30
Bob	-0.44	0.23
Mary	0.70	-0.06
Sue	0.31	0.93

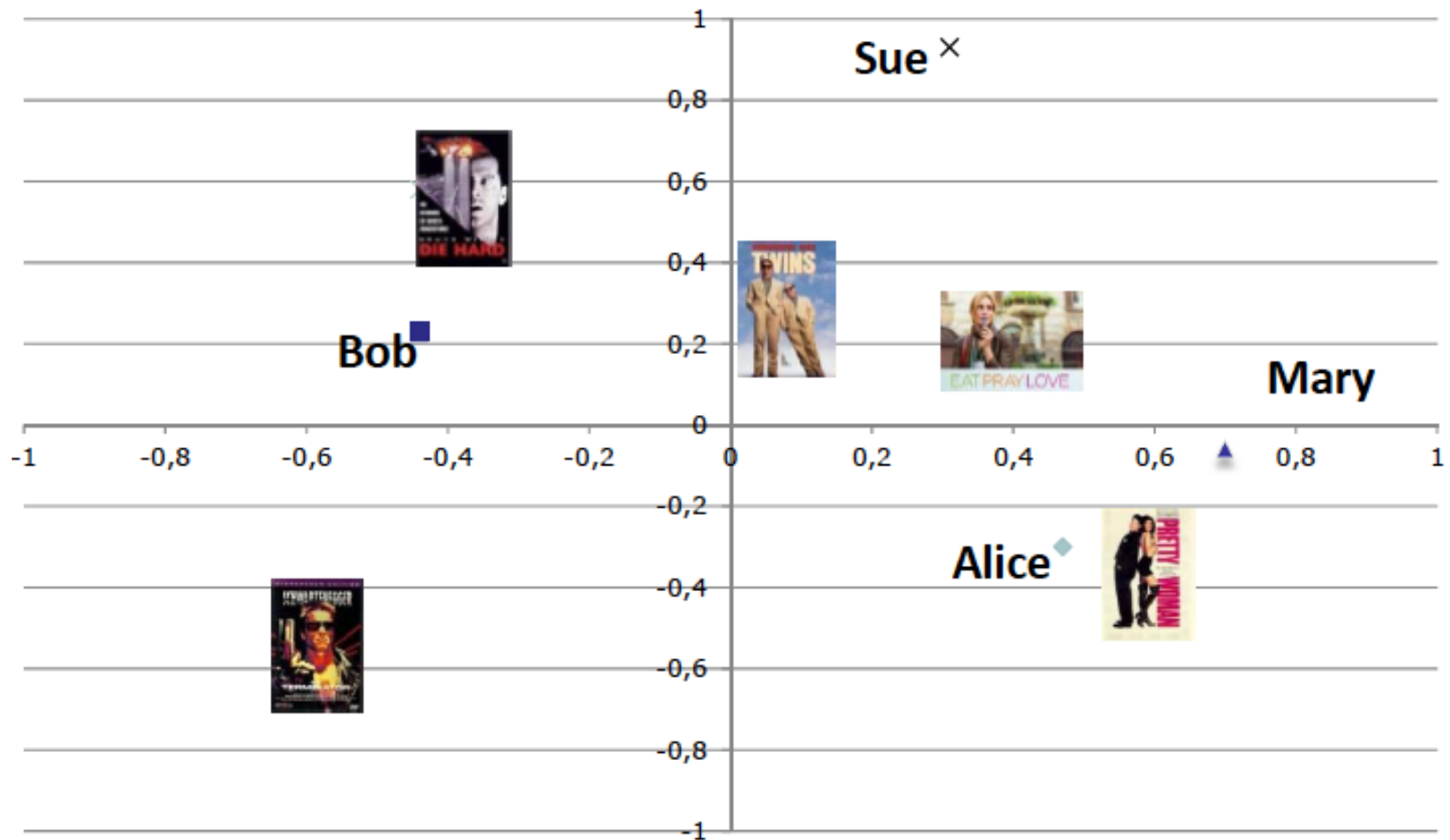
V_k^T					
Dim1	-0.44	-0.57	0.06	0.38	0.57
Dim2	0.58	-0.66	0.26	0.18	-0.36

Σ_k	Dim1	Dim2
Dim1	5.63	0
Dim2	0	3.23

$$\begin{aligned}\hat{r}_{ui} &= \bar{r}_u + U_k(\text{Alice}) \times \Sigma_k \times V_k^T(\text{EPL}) \\ &= 3 + 0.84 = \mathbf{3.84}\end{aligned}$$



Projetando em 2D





Redução de Dimensionalidade

- Pros
 - Modelo menor e mais rápido (tempo de consulta)
 - Vizinhança mais rica (compacta)
- Cons
 - Difícil de entender a semântica das dimensões
 - SVD tem alta complexidade (feito offline)