

## 2 Sistemas de Recomendação

Sistemas de recomendação são utilizados em portais de vendas e conteúdo com a finalidade de prever a preferência dos usuários. Predições com boa precisão podem resultar em margens de venda maiores e aumentar a satisfação dos usuários. As principais funções dos sistemas de recomendação são analisar os dados dos usuários e extrair informações úteis para futuras predições [4]. Existem diversas técnicas para implementação de sistemas de recomendação, incluindo as baseadas em conteúdo, as colaborativas, entre outras. Para melhorar a performance das predições, estes métodos podem ser combinados em sistemas híbridos.

Existem vários sistemas deste tipo acessíveis pela Internet que visam gerar recomendações para diversos tipos de produtos como músicas, filmes, livros, etc. Por exemplo, sistemas de recomendação são agora parte integral de diversos portais de comércio eletrônico como a Amazon.com e a iTunes Store [5]. De uma forma geral, estes sistemas buscam adquirir opiniões ou preferências sobre itens de um grupo de usuários, e usar estas opiniões para apresentar itens que possam fazer sentido para outros usuários.

A partir desta descrição geral vemos que os sistemas de recomendação precisam de, basicamente, duas coisas para poder funcionar de maneira apropriada:

- 1 Informações sobre as preferências dos usuários
- 2 Um método para determinar se um item é interessante para um usuário

Normalmente, as preferências dos usuários compreendem informações externas como suas características pessoais (idade, sexo, localidade, etc.), seu histórico de interações com o portal em questão, e suas avaliações em cima dos produtos [6]. A forma para determinar se um item é interessante para um usuário ou não depende do tipo de sistemas de recomendação. Neste capítulo, discutiremos algumas técnicas de recomendação comumente utilizadas.

## 2.1 O Processo de Recomendação

Em geral, pode-se dizer que todo sistema de recomendação segue um processo bem definido a fim de criar recomendações. Pensando no processo de recomendação como uma caixa preta, como mostrado na Figura 2.1, podemos identificar duas fontes de informação necessárias como entrada do processo. Estas fontes de informação correspondem aos dados dos usuários e as informações a respeito dos itens e dos usuários. Idealmente, estes dados relacionados ao perfil dos usuários deveriam ser fornecidos explicitamente pelos próprios usuários. No entanto, estas informações também podem ser extraídas de outras fontes como a navegação pelas páginas, itens consumidos ou comprados, forma como os itens foram consumidos, etc.

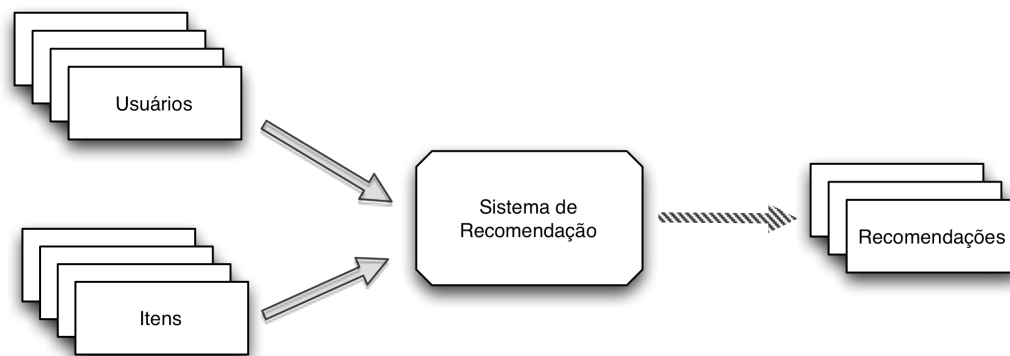


Figura 2.1 - Processo de recomendação

## 2.2 O Problema de Recomendação

Métodos de recomendação tem sido utilizados informalmente há anos. Recomendações positivas e negativas ajudam as pessoas a descobrir coisas novas que eles vão gostar ou evitar alternativas ruins que eles não vão gostar. Isto resulta em economia de tempo e reduz a sobrecarga de informações, que é um problema sério hoje em dia.

Sistemas de recomendação surgiram como uma área de pesquisa independente em meados de 1990 quando os pesquisadores passaram a se concentrar em problemas de recomendação que contavam explicitamente com a estrutura das avaliações [62, 63, 64]. Durante a última década, muito tem sido feito tanto na indústria quanto no meio acadêmico no desenvolvimento de novas

abordagens para sistemas de recomendação e o interesse nesta área ainda permanece.

Informalmente, o problema de recomendação pode ser reduzido a um problema de estimar avaliações para os itens que não foram experimentados por um usuário. Intuitivamente, esta estimativa é geralmente baseada na avaliação dada por este usuário para outros itens e em algumas outras informações relacionadas ao contexto. Uma vez que as avaliações para os itens ainda não avaliados podem ser calculadas, os itens de maior avaliação estimada podem ser recomendados para o usuário.

Formalmente, o problema de recomendação pode ser representado da forma a seguir [13]. O modelo formal de um sistema de recomendação consiste em três itens, um conjunto  $C$  com os usuários, um conjunto  $S$  com os itens passíveis de recomendação, tais como livros, filmes ou restaurantes e uma função de utilidade  $u$ . O espaço de itens possíveis  $S$  pode ser muito grande, com todos os filmes ou livros que poderão ser recomendados. Da mesma forma, o espaço dos usuários também pode ser muito grande, chegando a casa dos milhões, em alguns casos. A parte fundamental do modelo formal é a função de utilidade, que mede a utilidade de um item  $c$  para um usuário  $s$ , isto é,  $u: C \times S \rightarrow R$ . Aqui,  $R$  é um conjunto ordenado, podendo conter inteiros não negativos, ou números reais em um determinado intervalo, o que depende do esquema de representação utilizado no sistema de recomendação.

Usando o modelo descrito, o problema de recomendação é reduzido a escolher um item  $s_i \in S$ , para um usuário  $c_j \in C$  que maximiza a função de utilidade. Em sistemas de recomendação a utilidade de um item é geralmente representada por uma avaliação indicando o quanto um usuário em particular gosta de um item em particular.

Cada usuário do espaço  $C$  pode ser definido usando um perfil. Isso é o chamado perfil de usuário, que contém informações sobre os gostos do usuário para o domínio de recomendação. A representação do perfil do usuário depende da abordagem e do domínio, mas, por exemplo, ela pode incluir características de vários usuários, como idade, sexo, renda, estado civil, entre outros. No caso mais

simples, o perfil de usuário pode conter apenas um único elemento, como o ID de usuário.

Da mesma forma, cada item do espaço  $S$  pode ser definido com um conjunto de características, dependendo do esquema de representação escolhido. Por exemplo, em uma aplicação de recomendação de filmes  $S$  pode ser uma coleção de filmes. Cada um dos filmes em  $S$  pode ser representado não só pelo seu identificador único, mas também por sua informação de conteúdo, como seu título, gênero, diretor, ano de lançamento, atores principais, entre outros, o que é pensado para ter efeito sobre o processo de recomendação.

	<b>Matrix</b>	<b>Rocky</b>	<b>Hannibal</b>	<b>Ghost</b>
João	1	1	1	⊗
Joana	⊗	-1	⊗	1
Jaime	1	-1	-1	-1
Bernardo	⊗	1	1	-1

Tabela 2.1 - Matriz de avaliações para recomendação de filmes

A principal preocupação de um sistema de recomendação é que a função de utilidade  $u$  geralmente não é definida no espaço todo de  $C \times S$ . A utilidade é representada tipicamente por avaliações e é inicialmente definida apenas nos itens previamente avaliados pelos usuários. Por exemplo, em um sistema de recomendação de filmes, os usuários inicialmente avaliam algum subconjunto de filmes que eles já viram. Um exemplo de matriz de avaliação dos usuário para os itens para uma aplicação de recomendação de filmes é apresentado na Tabela 2.1, onde as avaliações são especificadas como 1, positiva, ou -1, negativa. O "⊗" significa que os usuários não avaliaram os filmes correspondentes. Portanto, o mecanismo de recomendação deve ser capaz de estimar ou prever as avaliações dessas combinações usuário-filme e, em seguida, fazer recomendações apropriadas com base nessas previsões. Se considerarmos o espaço  $C \times S$  como uma matriz de avaliações, muitas das avaliações não são fornecidas inicialmente, por isso a matriz é dita esparsa. Assim, o trabalho do sistema de recomendação é extrapolar  $u$  para todo o espaço  $C \times S$ . O caminho que é seguido para fazer estas

extrapolações define a função de utilidade. Esta tarefa de extrapolação tem sido abordada de diversas maneiras [64, 65, 66]:

1. Especificando heurísticas que irão definir a função de utilidade empiricamente através da validação de seu desempenho.
2. Estimar a função de utilidade através da otimização de critérios de desempenho pré-definidos, como o erro médio quadrático [13].

### **2.3 Estratégias de Recomendação**

As técnicas de recomendação têm algumas possíveis classificações [5, 7, 8].

Especificamente, os sistemas de recomendação possuem:

- dados do contexto, as informações que o sistema tem antes do processo de recomendação iniciar;
- dados de entrada, as informações que os usuários precisam comunicar ao sistema visando a geração das recomendações
- um algoritmo que combina os dados do contexto e os dados de entrada para produzir as sugestões.

Desta forma, as técnicas de recomendação mais comuns podem ser agrupadas em duas categorias distintas. A abordagem de filtragem de conteúdo busca criar um perfil para cada usuário ou produto de forma a caracterizar sua natureza. Por exemplo, para o caso de o produto ser um vídeo, pode-se usar o canal onde o mesmo é exibido, o programa ao qual ele pertence, os atores que participam e outros dados para a formação deste perfil para então recomendar outros vídeos que se encaixem dentro deste perfil. O mesmo acontece na criação dos perfis dos usuários, onde podem ser utilizadas informações como idade, sexo, localização geográfica, ou até mesmo respostas fornecidas em um questionário dedicado. Estratégias baseadas em conteúdo demandam informações adicionais dos produtos que, muitas vezes, não estão disponíveis ou são difíceis de se coletar [10].

Uma outra estratégia é baseada no comportamento prévio dos usuários sem a necessidade de derivarmos um perfil para o mesmo [10]. Esta abordagem é conhecida como filtragem colaborativa. Não há a necessidade de explicitar perfis para cada item ou usuário. A idéia é analisar as relações entre os usuários e as interdependências entre os produtos para poder identificar novas associações entre os produtos e os usuários.

### Filtragem de Conteúdo

A principal idéia por trás dos sistemas de recomendação baseados em conteúdo é sugerir itens para os usuários que sejam similares ao itens que estes usuários já tenham avaliado positivamente em algum momento antes. Recomendações baseadas em filtragem de conteúdo são uma extensão das pesquisas relacionadas a filtragem de informação [11]. Em um sistema baseado em conteúdo, os objetos de interesse são definidos pelas suas características. Por exemplo, o NewsWeeder [12], um sistemas de recomendação de notícias, usa as palavras dos textos para caracterizar os mesmos.

Algoritmos de recomendação baseados em conteúdo aprendem os perfis de interesse dos usuários com base nas características presentes em cada um dos itens que esses usuários avaliaram previamente. Sendo assim, eles constroem os perfis dos usuários a partir do perfil dos itens. O tipo de perfil derivado depende do método de aprendizado empregado. Árvores de decisão, redes neurais, e representações vetoriais são alguns exemplos destes métodos que tem sido comumente utilizados. Estes perfis de usuários são modelos que demandam algum tempo de observação e são atualizados a medida em que novas evidências relacionadas às preferências dos usuários são observadas.

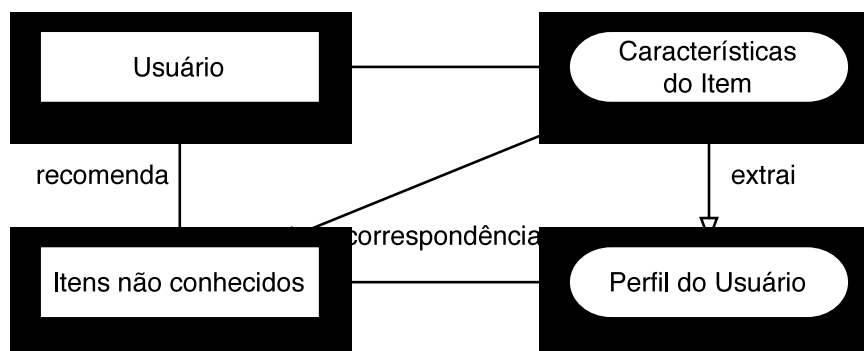


Figura 2.2 - Recomendação baseada em filtragem de conteúdo

Em termos formais, sistemas de recomendação baseados em filtragem de conteúdo estimam a utilidade  $u(c,s)$  de um item  $s$  para um usuário  $c$  baseado nas utilidades  $u(c,s_i)$  indicadas pelo usuário  $c$  para os itens  $s_i \in \mathcal{S}$  que são “similares” ao item  $s$  [13].

Um exemplo de sucesso da abordagem de filtragem de conteúdo é o Projeto *Music Genome* [9], usado no serviço de rádio online Pandora.com [35]. Uma pessoa especializada neste tipo de identificação classifica as músicas da rádio levando em consideração centenas de características musicais. Estes atributos esclarecem não apenas a identidade musical de cada faixa como também os fatores relevantes para entender as preferências musicais dos ouvintes.

### Filtragem Colaborativa

Filtragem colaborativa é uma técnica que utiliza o histórico de interações entre itens e usuários visando encontrar relacionamentos entre os mesmos. Um exemplo de relacionamento pode ser dois vídeos sendo visualizados por muitos usuários, indicando que eles são similares. Nenhuma informação contextual a respeito do item em questão é levada em consideração, ou seja, os algoritmos não têm ciência de sobre quais itens ou que tipo de itens eles estão processando.

Em termos formais, sistemas de recomendação baseados em filtragem colaborativa estimam a utilidade  $u(c,s)$  de um item  $s$  para um usuário  $c$  baseado na utilidade  $u(c_j,s)$  do item  $s$  indicadas pelos usuários  $c_j \in \mathcal{C}$  que são “similares” ao usuário  $c$  [13]. Por exemplo, num sistema de recomendação de vídeos, para recomendar um vídeo para o usuário  $c$ , o sistema tentará encontrar os “pares” do usuário  $c$ , ou seja, outros usuários com gosto parecido ou que tenham avaliado os mesmos vídeos de forma parecida.

É comum nos sistemas de filtragem colaborativa que os perfis dos usuários sejam armazenados como vetores de itens e suas avaliações para aqueles itens. Esses vetores tendem a aumentar continuamente a medida em que os usuários interagem com o sistema. Alguns sistemas levam em consideração dinâmicas temporais para descontar os desvios no padrão de interesse dos usuários com o passar do tempo [14]. As avaliações podem ser binárias, tais como

gostei ou não gostei, ou valoradas de acordo com o nível de preferência. Dentre os sistemas mais conhecidos que utilizam esta abordagem temos o da Netflix [15], da MovieLens [16], da Amazon [17], entre outros.

#### Best Value

Buy **Gran Turismo 5 (Prima Essential Track Guide): Prima Essential Game Guide** and get **Gran Turismo 5** at an **additional 5% off** Amazon.com's everyday low price.



**Buy Together Today: \$54.84**

[Add both to Cart](#)

[Show availability and shipping details](#)

#### Customers Who Bought This Item Also Bought

Page 1 of 12





 Logitech PlayStation 3 Driving Force GT Racing Wheel by Logitech ★★★★★ (278) \$104.00	 PlayStation 3 DualShock 3 Wireless Co... by Sony Computer Enterta... ★★★★★ (929) \$41.54	 PlayStation 3 160GB System by Sony ★★★★★ (341) \$249.99	 Pro Evolution Soccer 2011 by Konami ★★★★★ (66) \$29.99
--	---	--	---

Figura 2.3 - Filtragem colaborativa na Amazon.com

As duas formas mais comuns de implementar uma solução de filtragem colaborativa são os algoritmos de vizinhança ou grafos, e modelos de fatoração latente [3].

Métodos de vizinhança visam computar relacionamentos entre itens ou usuários e construir um modelo baseado num grafo que descreva a vizinhança:

- Métodos item-item constroem os grafos de vizinhança com os vértices conectando itens similares;
- Métodos usuário-usuário constroem os grafos de vizinhança com os vértices conectando usuários com gosto parecido.



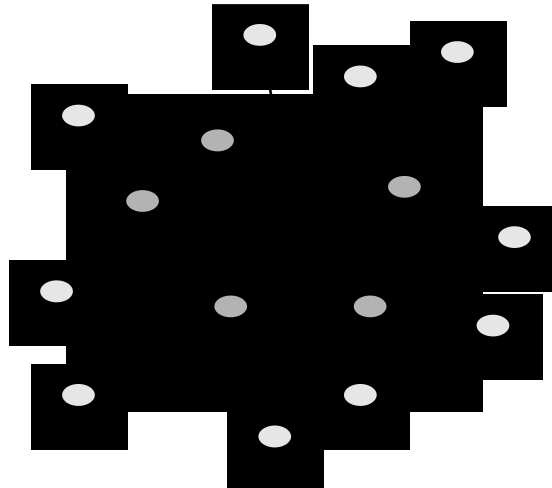


Figura 2.4 - Processo de formação de vizinhança

Estes métodos tem obtido sucesso em muitos domínios, mas seus algoritmos possuem algumas limitações:

1. **Esparsidade dos dados.** Métodos de vizinhança se baseiam em correspondências exatas, fazendo com que os algoritmos sacrifiquem a precisão dos sistemas de recomendação [61]. Uma vez que o coeficiente de correlação é definido apenas entre usuários que avaliaram, pelo menos, dois itens em comum, muitos pares de usuários não têm nenhuma correlação entre si. Na prática, muitos sistemas de recomendação são usados para operar grandes massas de dados, por exemplo, a Amazon.com recomendando livros e o eBay com produtos usados. Nestes sistemas, mesmo os usuários altamente ativos podem ter interagido ou avaliado bem menos que 1% dos itens. Assim, os algoritmos de vizinhança podem ser incapazes de fazer boas recomendações de itens para um determinado usuário.
2. **Escalabilidade.** Algoritmos de vizinhança demandam um esforço computacional que seja capaz de crescer tanto com o número de usuários quanto com o número de itens. Com milhões de clientes e produtos, um sistema de recomendação *web* padrão executando os algoritmos existentes acabam enfrentando problemas de escalabilidade.
3. **Sinônimos.** Num cenário real, nomes de produtos diferentes podem se referir a objetos semelhantes. Sistemas de recomendação baseados em

correlação não conseguem encontrar esta associação latente e tratam estes itens como produtos diferentes.

Modelos de fatoração latente tentam explicar o interesse dos usuários por certos itens caracterizando os mesmos em, digamos, 20 a 100 fatores latentes aprendidos nessas relações de interesse. Eles constroem uma aproximação da matriz de interações entre os usuários e os itens e usam essa aproximação para estimar a probabilidade de outras interações. Focamos este trabalho numa especialização do problema de Fatoração de Matrizes Não-Negativas (*NMF*, do inglês *Non-Negative Matrix Factorization*) [41]. Descrevemos em detalhes o funcionamento deste método no capítulo 3.

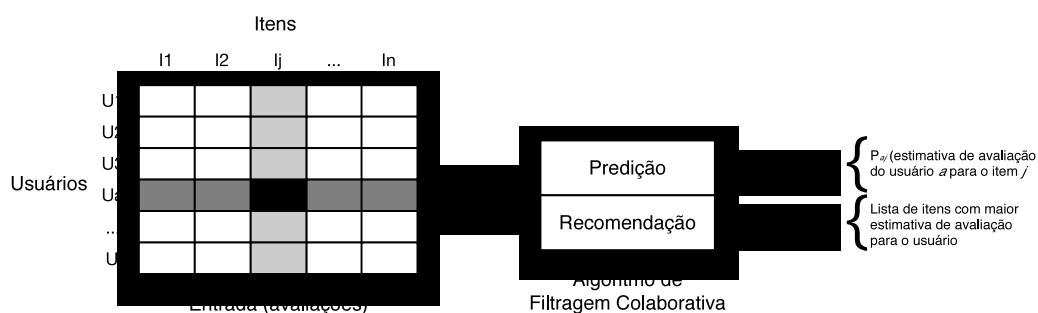


Figura 2.5 - Processo de filtragem colaborativa

Existem muitos desafios para os sistemas de filtragem colaborativa. Seus algoritmos precisam ter performance em conjuntos de dados altamente esparsos, escalar com o aumento do número de usuários e itens, conseguir conceber boas recomendações num curto período de tempo e ainda lidar com outros problemas como o ruído no conjunto de dados e de privacidade.

### Comparação das Estratégias

Há vantagens e desvantagens de ambos os tipos de abordagens para sistemas de recomendação.

A principal desvantagem das abordagens baseadas em filtragem de conteúdo em comparação com as baseadas em filtragem colaborativa é que elas dependem muito dos metadados que descrevem os itens, o que implica que essas informações devem ser obtidas e inseridas no sistema. Assim, a manutenção de um sistema de recomendação baseado em conteúdo requer um esforço significativo, já que existe a necessidade constante de manter as informações

atualizadas ou acrescentar novas informações quando novos itens são adicionados ao sistema.

Uma outra questão é que nem todas as informações podem ser expressas facilmente ou de forma consistente. Algo simples como o gênero de um filme não pode ser utilizado diretamente, pois esta informação nem sempre é bem definida. Os usuários podem ter interpretações diferentes a respeito do gênero ao qual um filme pertence, além disso um mesmo filme pode estar situado entre dois gêneros distintos. Usar os dados de gênero pode, assim, causar problemas quando o sistema se baseia neste tipo de informações.

Abordagens de filtragem colaborativa não dependem destes tipos de informação, o que as torna mais útil em domínios onde não está claro quais atributos dos itens implicam nas preferências dos usuários. A principal deficiência de algoritmos de filtragem colaborativa, porém, é o que se chama de problema de partida a frio. Uma vez que estes algoritmos dependem de informações de avaliação, é impossível recomendar itens que ainda não foram avaliados pelos usuários, ou dar recomendações a usuários que não forneceram qualquer informação de preferência.

Além disso, uma vez que esses sistemas contam com informações de preferência, é necessário que os usuários compartilhem seus gostos. Usuários que estão muito preocupados com a questão de privacidade pode considerar essa necessidade de compartilhamento um problema. Em sistemas de recomendação baseados em conteúdo, não há necessidade de compartilhar as informações entre os usuários, pois os algoritmos só utilizam as avaliações pessoais e semelhanças entre os usuários em termos de meta-informação, e não comparando os padrões de avaliação.

Levando essas vantagens e desvantagens em consideração a conclusão é que diferentes abordagens são mais adequadas em diferentes domínios. Existem tentativas de combinar filtragem colaborativa e filtragem baseada em conteúdo em sistemas de recomendação de filmes visando contornar as deficiências de cada um dos métodos [65, 67], mas por conta da facilidade de manutenção os algoritmos de filtragem colaborativa têm sido mais utilizados na maioria dos

sistemas de recomendação hoje em dia.