4
RECOMENDADORES BASEADOS EM CONTEÚDO

4.1
Introdução

Como pôde ser visto no capítulo 2, todo produto planejado para dar valor a um nicho de mercado possui características baseadas em um posicionamento. Para a criação de um posicionamento para um produto, necessita-se conhecer os interesses e as preferências de um grupo de consumidores e transformá-los em números, ajustando a quantidade de aspectos de valor com o custo máximo que esses consumidores estariam dispostos a suportar por esse produto.

O método de filtragem colaborativa ignora esse processo que vai desde a criação do posicionamento do produto até a formação de suas características individuais. A filtragem colaborativa apenas leva em conta a proximidade entre usuários ou itens, causando problemas quando o usuário não possui suficientes avaliações.

O capítulo 3 apresentou como solução para os problemas de filtragem colaborativa o uso de uma abordagem híbrida com recomendadores baseados em conteúdo. Neste capítulo, discutem-se os recomendadores baseados em conteúdo que, ao contrário dos filtros colaborativos, levam em conta as informações tanto de usuário quanto dos itens que irão recomendar. .

Ainda neste capítulo, será apresentado um método criado por Mohammad Javad Hosseinpour [7] que executa um processo análogo ao utilizado em marketing para a criação de características de produtos. Enquanto o objetivo do processo em marketing é de criar características de produtos a partir de posicionamentos, no processo de Hosseinpour, dado um conjunto de características de um produto, criam-se posicionamentos com números fuzzy de interesse para consumidores potenciais. O método permite que, por meio de uma interface de usuário simples, um consumidor identifique quais os posicionamentos que lhe dão valor para que sejam recomendados produtos.

A transformação de características de produtos em fatores de posicionamento é feita no algoritmo por meio de opiniões de especialistas (por exemplo, os próprios vendedores ou os fabricantes dos itens a serem recomendados), que são tratadas como números fuzzy.

A seguir são apresentadas as características dos recomendadores não baseados em filtragem colaborativa, suas falhas e suas potencialidades de interação com os outros algoritmos estudados nesse documento, de forma a melhorar a recomendação de produtos em sites de e-commerce.

4.2
Recomendações Baseados em Conteúdo

Um sistema de recomendação baseado em conteúdo faz uso de processamento de texto para efetuar suas recomendações. Os textos podem vir de diversas formas: documentos, *URLs*, mensagens de notícias, *logs* de sites, descrição de itens ou usuários, preferências de usuários, etc. Os recomentadores buscam padrões nesses textos que permitam a recomendação [66].

Recomendadores baseados em conteúdo utilizam métodos heurísticos e algoritmos de classificação para fazer suas recomendações [5], por exemplo analisando em dois documentos a freqüência das palavras em cada um para buscar uma similaridade, ou descobrindo características em comum de itens que o usuário comprou no passado [79].

Recomendadores baseados em conteúdo tem forte aplicação para criação de filtros de spam onde o conteúdo das mensagens é processado pelo algoritmo de forma a recomendar aquelas que são definidas como spams indesejados. Um *survey* da aplicação destes algoritmos para filtro de spam pode ser obtido em [80]. Da mesma forma que no caso de spams os algoritmos analisam os textos das mensagens, no caso de recomendação de itens, o algoritmo terá de analisar as especificações do item.

As técnicas baseadas em conteúdo possuem um problema de inicialização que difere do problema de mesmo nome em filtros colaborativos. Enquanto na filtragem colaborativa o problema é de um usuário novo que tenha avaliado poucos itens (ou itens novos pouco avaliados), o problema de inicialização em técnicas baseadas em conteúdo ocorre quando há poucas informações de conteúdo

para ser analisadas. Eles também são limitados à presença de características explicitamente associadas aos objetos que recomendam, muitas vezes necessitando de opiniões de especialistas ou de informações não presentes. Essas técnicas baseadas em conteúdo também podem apresentar problemas de super-especialização, recomendando apenas itens altamente direcionados a certos tipos de usuários [56].

Diversos dos problemas que existem nos recomendadores baseados em conteúdo tendem a ser cada vez mais superadas em aplicações que utilizam a internet. As redes sociais adicionaram muitos dados sobre usuários disponíveis na internet, dados estes que podem ser utilizadas pelo algoritmo baseado em conteúdo. O artigo [81] explora algoritmo baseado em conteúdo com teoria sociológica para fazer recomendações de conteúdo de mídia.

Outros sistemas de recomendação incluem recomendadores baseados em demografia, que usam dados do usuário tais como sexo, endereço, ocupação etc. [82]. Sistemas de recomendação baseados em conhecimento e em utilidade utilizam conhecimento de como um objeto particular satisfaz a suas necessidades [83].

4.3
Recomendador Baseado em Conteúdo por Números Fuzzy

Nas sessões seguintes, apresenta-se um método de recomendação baseado em conteúdo que associa itens a usuários levando em consideração características de produtos e avaliações dos usuários quanto a estes itens. Tal método foi proposto por Hosseinpour [7] e utiliza números fuzzy e características do produto como bases de recomendação. A razão para a apresentação deste algoritmo em particular está na sua relação direta com os conceitos de marketing e com a sua facilidade de hibridização com um filtro colaborativo. Essas ponderações serão detalhadas no capítulo 5, quando será apresentado o modelo de recomendação proposto neste trabalho.

4.3.1
Modelagem

O recomendador baseado em conteúdo proposto por Hosseinpour (2008) faz uso de números fuzzy e características do produto como bases de recomendação. Os números fuzzy, aqui considerados como triangulares, conforme mostrado na Figura 11, podem ser vistos como distribuições de possibilidade e são denotados por, $\tilde{p} = (p_1, p_2, p_3)$ com função de pertinência $\mu_{\tilde{p}}(x)$, onde p_1, p_2 e p_3 são números reais tais que $p_1 \leq p_2 \leq p_3$.

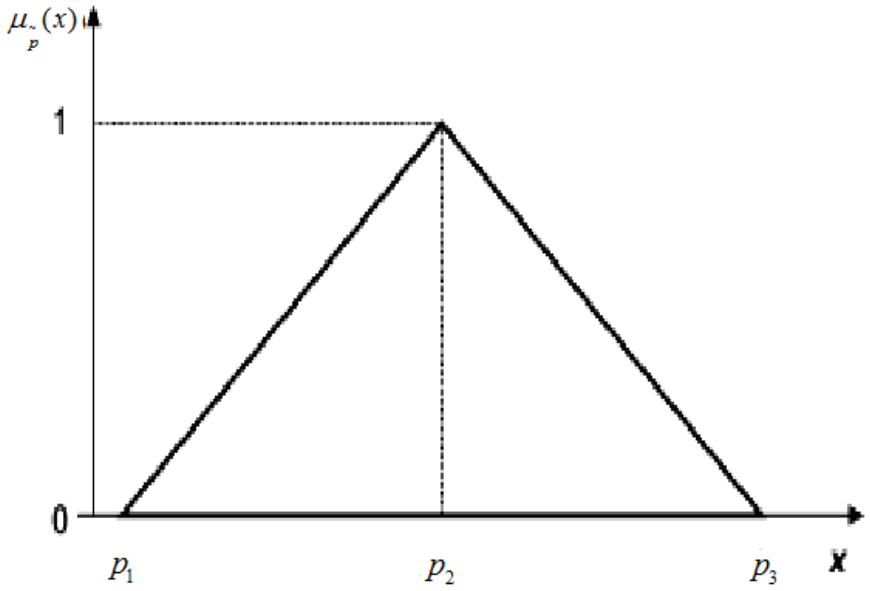

Figura 11 – Número fuzzy triangular

Todo item a ser recomendado pode ser definido como possuindo um conjunto de características técnicas que o diferencia unicamente de outros itens da mesma categoria de produto. O método de Hosseinpour considera apenas itens da mesma categoria. Exemplos de categorias seriam televisores, geladeiras e laptops. Geladeiras podem ter como características o número de portas, o volume interno e gasto energético, por exemplo.

Na metodologia proposta, as diversas *características técnicas* que distinguem itens entre si numa mesma categoria devem ser transformadas em *componentes* que tenham algum valor para os usuários. Para a avaliação dos componentes definem-se sete números fuzzy, representados na Figura 12, aos

quais são associados os termos lingüísticos muito baixa (MuB), baixa (B), média baixa (MeB), média (Me), média alta (MeA), alta (A) e muito alta (MuA).

Por exemplo, uma televisão pode ter como característica a sua imagem, que pode ser constituída por atributos numéricos, como o tamanho da tela em polegadas, assim como atributos categóricos, como qualidade digital ou não e se ela é de Plasma, LCD ou OLED. Essas características técnicas geram uma avaliação única para o componente imagem da televisão.

No caso de atributos numéricos, os números fuzzy podem ser igualmente distribuídos ao longo do valor normalizado ou não. A Figura 12 e a Tabela 5 podem exemplificar o uso de 7 números fuzzy ao longo de um valor numérico que varia de 0 a 8. Da mesma forma, caso a especificação seja categórica, utiliza-se um número fuzzy por categoria. Supondo o caso da tela da televisão ser de Plasma, LCD ou LED, pode-se dividir como:

<Muito Baixo> = Plasma

<Médio> = LCD

<Muito Alto> = LED

A ordem vem pelo conhecimento de que a tecnologia Plasma é inferior as de LCD e LED. A forma como as categorias se dividem em números fuzzy passam pelo conhecimento de quem está modelando podendo ser subjetiva e necessitando, portanto, de alguém com bom conhecimento do produto ou de como o produto é visto por consumidores.

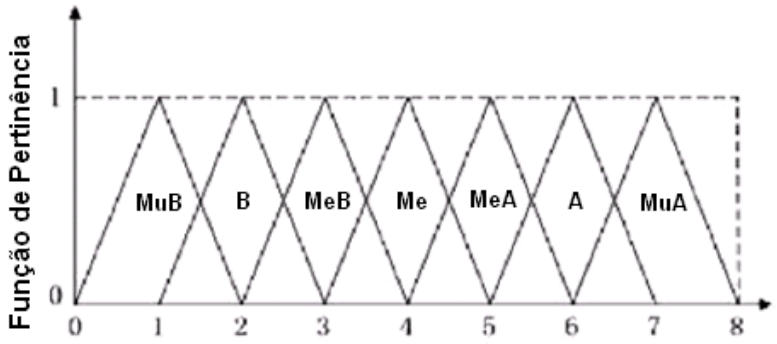

Figura 12 - Função de Pertinência para números fuzzy triangulares

Tabela 5 – Termos Lingüísticos dos números fuzzy triangulares

Termos Lingüísticos Relacionados aos Números Fuzzy	Números Fuzzy Triangular
Muito Baixo (MuB)	(0,1,2)
Baixo (B)	(1,2,3)
Médio Baixo (MeB)	(2,3,4)
Médio (Me)	(3,4,5)
Médio Alto (MeA)	(4,5,6)
Alto (A)	(5,6,7)
Muito Alto (MuA)	(6,7,8)

O método sugere que os componentes sejam obtidos diretamente a partir das características físicas do produto. Esse processo conta com a opinião de especialistas (pessoas que conheçam o produto como vendedores ou fabricantes), tanto para transformar as características físicas em números fuzzy, quanto para dar pesos a estes números com relação aos componentes.

Considera-se que cada item I_i é representado por um vetor $\widetilde{P}_i$ com n componentes relativos àquele item. O mesmo item I_i possui outro vetor E_i composto pelas especificações técnicas que o diferenciam de outros itens da mesma linha. Cada componente $p_i \in \widetilde{P}_i$ é composto por um vetor de especificações $\widetilde{E}_p = (\widetilde{e}_p^1, \widetilde{e}_p^2, ..., \widetilde{e}_p^k)$, onde cada especificação funcional $\widetilde{e}_p^i$ é um número fuzzy triangular que representa a capacidade de esta especificação do produto afetar o componente. Como cada especificação técnica pode influir de forma distinta sobre o componente, pode-se considerar também um vetor de pesos $W=(w_1,w_2,...,w_k)$. É possível, então, calcular o valor de um componente como um número fuzzy triangular a partir de especificações técnicas:

$$p_i = \sum_{j=1}^{k} (e_i^j \times \quad w_i^j) \qquad 15$$

O vetor de componentes $\widetilde{P}_i = (\widetilde{p}_1, \widetilde{p}_2, ..., \widetilde{p}_n)$ é formado por números fuzzy triangulares que representam cada qual um componente do produto e que podem ser utilizados para comparações dos interesses de valor do cliente.

A conversão de especificações funcionais e técnicas de um produto em um número fuzzy, bem como a definição e composição dos componentes, podem ser obtidas de diversas formas: opiniões de especialistas, pesquisas de opinião, pesquisas científicas, etc.

Para exemplificar esse processo, utiliza-se um estudo sobre os desejos de consumo de celulares em mulheres [84]. A partir de pesquisas de opinião, chegou-se a cinco componentes de valor para as consumidoras: Design, Preço, Tamanho, Tecnologia e Marca. Estes formam o vetor de componentes para qualquer item do tipo "celular".

A mesma pesquisa analisa cada um destes componentes quanto às características físicas do celular. No componente Design, por exemplo, há o atributo (especificação) Flip. A pesquisa compara o flip tradicional ao deslizante e celulares sem flip.

A Tabela 6 apresenta 5 colunas ordenadas da seguinte forma: (1) Atributos; (2) Níveis; (3) Utilidade parcial retirada da pesquisa; (4) utilidade normalizada e (5) a coluna da importância que também foi retirada da pesquisa.

A utilidade normalizada é calculada da seguinte forma: Utilidade normalizada = (utilidade parcial – valor mínimo das utilidades parciais) dividido pela amplitude. A amplitude é a diferença entre o valor máximo e o valor mínimo da utilidade parcial (1,0000 – (-1,3743) = 2,3743).

A partir da coluna “Importância Relativa” na Tabela 6, é possível obter a importância dada por mulheres aos diversos atributos do celular, sendo o “flip” aquele de maior importância, com 30,88%. Em segundo lugar ficou a “tecnologia”, com 27,20%, em terceiro, o “preço”, com 21,32% e, em quarto, as “funções/aparência”, com 20,60%. Podemos notar que os atributos “preço” e “funções/aparência” tiveram seus percentuais muito próximos. Isso demonstra que, provavelmente, estes atributos são igualmente importantes na hora da tomada de decisão. Ou seja, a mulher avalia o custo x benefício (o preço em relação às funções que o telefone tem).

Fazendo uma analogia dos resultados da pesquisa com a Eq.15, o Design seria um componente (p) sendo este descrito por especificações do celular:

e_1=Presença de Flip [sem = muito baixo (0), tradicional = muito alto(0.91) , deslizante= alto (0.82)]

e_2 = Personalização da Carcaça [não = muito baixo, sim = médio alto (0.72)]

Os valores de *w* vem da importância relativa: 30.88% para o Flip e 20.60% para a carcaça personalizada. Em um celular com carcaça personalizada e flip deslizante o cálculo seria:

$$p_{Design} = médioalto * 20.60\% + alto * 30.88\% =$$
$$p_{Design} = (4,5,6) * 0.206 + (5,6,7) * 0.3088$$
$$p_{Design} = (2.36,2.88,3.39)$$

Tabela 6 – Tabela de comparação de utilidade de celulares para mulheres (pesquisa Mattiota)

ATRIBUTO	NÍVEL	UTILIDADE PARCIAL	UTILIDADE NORMALIZADA	IMPORTÂNCIA RELATIVA
Preço	R$250-R$400	0,4912	0,79	21,32%
	R$401-R$600	(0,0526)	0,56	
	R$601-R$900	(0,4386)	0,39	
Flip	Com Flip Tradicional	0,7953	0,91	30,88%
	Sem Flip	(1,3743)	0,00	
	Flip Deslizante	0,5789	0,82	
Tecnologia	Toques de chamada em MP3	(0,2164)	0,49	27,20%
	Com Câmera VGA 4X Zoom	1,0000	1,00	
	Com Rádio	(0,7836)	0,25	
Funções / Aparência	Com espelho	(0,1404)	0,52	20,60%
	Personalização de carcaça	0,3392	0,72	
	Com tabela de calorias e ciclo menstrual	(0,1988)	0,50	

4.3.2
Similaridade entre números Fuzzy

Todo algoritmo de recomendação necessita gerar uma similaridade entre os usuários e os itens para comparação.

No método proposto por Hosseinpour, o usuário define o conjunto de componentes que ele busca no item que lhe será recomendado, sendo cada componente um número fuzzy, conforme descrito anteriormente. Ao mesmo tempo, os diversos itens possuem seus componentes definidos como números fuzzy a partir das especificações, de forma a poderem ser comparados com o conjunto de componentes definido pelo usuário. A comparação entre dois números fuzzy é realizada calculando-se a compactação próxima de ambos.

Suponha-se que um usuário defina o componente para um item de seu interesse como $\tilde{q}_B = (q_B^1, q_B^2, q_B^3)$ e que exista um item que possua como valor para o

mesmo componente o número fuzzy $\tilde{q}_A = (q_A^1, q_A^2, q_A^3)$. O cálculo da similaridade entre ambos (por meio da compactação próxima [7]) é dado por:

$$N_E(\tilde{q}_A, \tilde{q}_B) = 1 - \frac{1}{\sqrt{3}} \left(\sum_{j=1}^{3} \left| q_A^j - q_B^j \right|^2 \right)^{1/2} \qquad 16$$

Por exemplo, se um cliente necessita de uma televisão com tamanho de tela muito pequena $\tilde{q}_A = (0,1,2)$ e uma dada televisão possui uma tela de tamanho grande $\tilde{q}_B = (5,6,7)$, utiliza-se a Eq. 2 acima para resolver a similaridade entre os interesses do consumidor e os valores do produto para um componente. Porém, em geral, um produto pode apresentar vários componentes relevantes para a decisão do cliente. O cálculo de similaridade deve considerar então a proximidade de vários números fuzzy relacionados.

Considerem-se os números fuzzy triangulares $\tilde{X} = (\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n)$, representando os interesses do cliente para cada componente de um produto, e $\tilde{Y} = (\tilde{y}_1, \tilde{y}_2, ..., \tilde{y}_n)$, representando o valor que um produto oferece a cada um dos respectivos componentes. Considerando que um item pode possuir componentes mais importantes para o usuário do que outros (por exemplo, o custo do produto pode ser mais relevante para um produto popular que outros componentes), faz-se uso de um vetor de pesos $\tilde{V} = (\tilde{v}_1, \tilde{v}_2, ..., \tilde{v}_n)$, cujos valores são normalizados, i.e. $\sum_{i=1}^{n} v_i = 1$. O vetor de pesos V é definido pela opinião de especialistas, podendo ser este o próprio vendedor ou fabricante do item. A similaridade entre os interesses e os valores dos itens, para todos os componentes, por compactação próxima, é dada por:

$$N_E(\tilde{X}, \tilde{Y}) = \sum_{i=1}^{n} \left(N_E(\tilde{x}_i, \tilde{y}_i) \times v_i \right) \qquad 17$$

A partir da Eq. 17, é possível comparar os interesses de um cliente por todos os produtos disponíveis à venda, com base no valor que cada produto oferece por meio dos componentes. Quanto menor o valor de $N_E(\tilde{X}, \tilde{Y})$, mais próximo o item analisado está dos interesses do cliente.

Em *websites* utilizados massivamente por diversos usuários, pode-se obter o componente dos produtos a partir de avaliações quanto a características relevantes. A Figura 13 apresenta um exemplo de um sistema de votação para televisões que pode ser utilizado no método proposto por Hosseinpour, onde cada valor de avaliação é representativo de um dos sete números fuzzy associados ao componente.

Figura 13 – Interface com o usuário.

4.3.3
Arquitetura do Sistema Hosseinpour

O sistema de recomendação fuzzy apresentado nesse capítulo pode ser utilizado para propor recomendações a usuários com base no valor que estes atribuem a cada componente de produto. A figura 14 apresenta uma visão geral da arquitetura desse sistema de recomendação.

Essa arquitetura considera a possibilidade de o consumidor oferecer ao sistema informações sobre seus interesses pessoais com relação a uma categoria de produtos. Na interface com o usuário, este define quanto de valor cada componente do produto lhe oferece. O sistema considera a existência de um banco de dados de componentes pré-calculados para cada item disponível ao usuário. O cálculo dos componentes é feito a partir de votos e das características técnicas do produto, conforme explicado anteriormente.

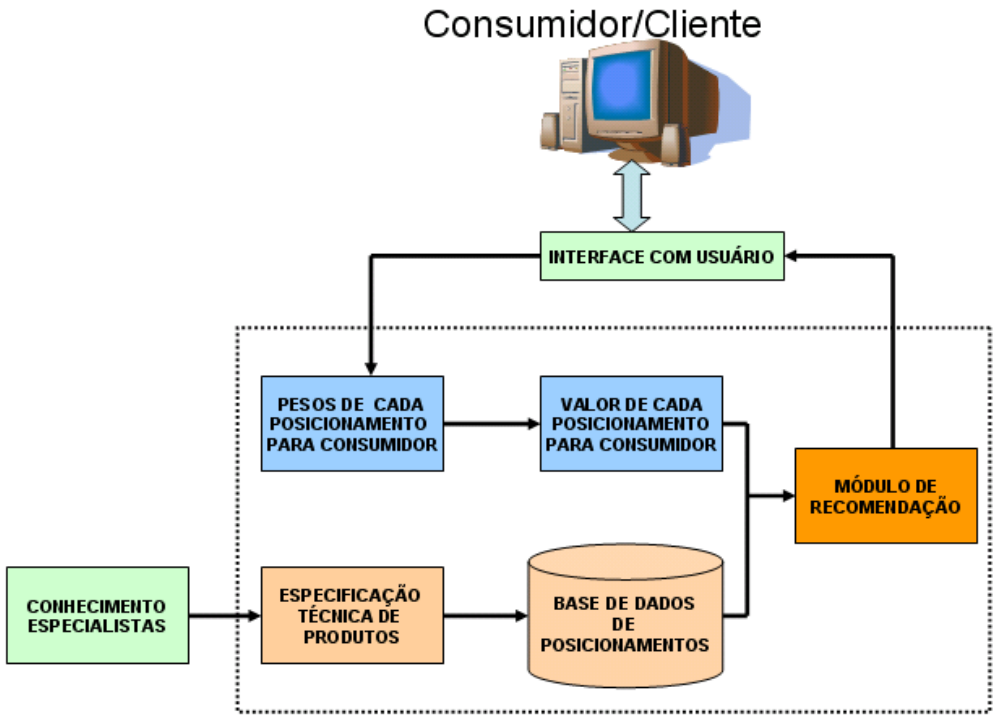

Figura 14 – Sistema de recomendação por números fuzzy.

O módulo de recomendação calcula a similaridade entre os componentes de cada item à disposição para compra pelo usuário e aqueles passados pelo usuário através da interface. Os valores de similaridade de cada item são base para comparação entre eles; os que tiverem menor similaridade são os que mais possuem maior valor para o usuário; estes são apresentados como recomendação de compra.

O sistema de recomendação, portanto, cumpre os seguintes passos:

Passo 1 (Definições de Componente) – Para cada segmento de produto presente na base de dados de produtos à venda, devem ser definidos os componentes para a categoria de produto, como também as especificações técnicas e os pesos a elas atribuídos.

Passo 2 – Devem ser definidos para cada item presente no segmento de produto os números fuzzy relativos a cada especificação técnica relevantes aos componentes. Esse processo é feito por meio da opinião de especialistas ou então a partir de pesquisas como em [84]. .

Passo 3 (Pré-Processamento) - Devem ser calculados os valores de cada componente, com base na Eq.17. Deve ser criado o vetor de componentes fuzzy para cada item e armazenado em uma base de dados.

Caso existam componentes de maior nível, devem ser calculados primeiramente os componentes de nível inferior e, em seguida, os de nível superior.

Passo 4: Por meio de respostas a um questionário, conforme visto na Figura 13, o consumidor gera um vetor de componentes de valores R, que quantifica suas necessidade.

Passo 5: É calculada a similaridade entre os valores fuzzy de R e dos componentes C para cada item, utilizando-se as equações 15 e 16. Os valores de similaridade de cada item são comparados e os dez melhores são apresentados como recomendação ao usuário.

4.3.4
Vantagens e Desvantagens

A principal vantagem do sistema proposto por Hosseinpour é que, conforme será explicado no próximo capítulo, ele possui similaridades com conceitos de marketing de componente de produtos para dar valor ao consumidor. As recomendações feitas pelo sistema expressam necessidades passadas pelo consumidor e, portanto, podem despertar interesse de compra.

As desvantagens deste sistema são muitas. A primeira é que ele só analisa uma categoria de produtos por vez. Serve para recomendar a compra, por exemplo, da melhor televisão dentre as televisões disponíveis, mas, se o usuário quiser outro tipo de produto (um DVD Player, por exemplo), ele terá de preencher novamente uma lista de componentes referentes à nova categoria.

Outra desvantagem do sistema é necessitar de uma grande quantidade de opiniões de especialistas para formar sua base de dados de componentes. No caso de um sistema web, os componentes podem ser gerados a partir de votos dos usuários nos produtos, sendo a opinião de especialistas necessária apenas para novos produtos, ainda não votados.

Essas duas desvantagens serão novamente discutidas no capítulo 5 onde, conforme os objetivos deste estudo, será proposto um algoritmo novo capaz de

superar as desvantagens tanto de algoritmos de filtragem colaborativa quanto os baseados em conteúdo, aproveitando a sinergia de ambos para uma melhor recomendação de itens a usuários.

Por fim, há a desvantagem de que o usuário deve não apenas se identificar, mas preencher um questionário para definir o que ele deseja. Muitos usuários não possuem tempo para preencher questionários. A partir de um processo de segmentação de usuários em nichos com comportamentos de compra semelhantes essa desvantagem do algoritmo pode, potencialmente, ser reduzida, o que pode ser motivo de outros estudos sobre o algoritmo. O algoritmo que será proposto no capítulo 5 resolve o problema da identificação do usuário com a hibridização com filtros colaborativos.