5

FILTRO HÍBRIDO FUZZY

5.1
Introdução

Nos capítulos anteriores foram apresentados diversos algoritmos e conceitos que são atualmente empregados no varejo para auxiliar suas vendas, quer seja na definição de novos produtos, em sistemas de segmentação de usuários, ou na recomendação de produtos.

O objetivo deste trabalho foi buscar um sistema de recomendação de desempenho superior ao dos utilizados comercialmente, que possa ser utilizado por sites de comércio virtuais, onde as bases de dados são esparsas e de grande escala, e que seja particularmente útil para o tratamento de itens de múltiplas categorias. Para tanto, foi desenvolvido um algoritmo de recomendação híbrido entre um filtro colaborativo e um sistema baseado em conteúdo, sendo este uma das contribuições criadas nesta tese.

A seguir, será apresentado o modelo proposto criado para alcançar esse objetivo, explicitando-se o motivo do uso de cada algoritmo e os métodos que serão utilizados para avaliar o desempenho do novo modelo.

Inicialmente, uma nova interpretação para o método de recomendação apresentado no capítulo 4 será feita, conectando-o ao conceito de posicionamento apresentado no capítulo 2. Em seguida, será explicado o motivo pelo qual o método de Hosseinpour foi escolhido em detrimento de outros algoritmos baseados em conteúdo. A interpretação do método de Hosseinpour com os conceitos de posicionamento de marketing mais uma contribuição relevante da tese para a comunidade científica.

Posteriormente, será apresentado o algoritmo de filtragem colaborativa mesclado ao método baseado em conteúdo escolhido, apresentando os argumentos de escolha. Será também descrito a forma como ambos algoritmos são unidos.

Por fim, será apresentada a modelagem de uma base de dados sintética a ser utilizada para testar o modelo proposto e compará-lo aos modelos tradicionais de

filtragem colaborativa, bem como escolha e detalhamento da base de dados real. A base de dados sintética é uma contribuição da tese para a comunidade científica, uma vez que ela foi criada de forma a se aproximar do comportamento de compra real de consumidores brasileiros, podendo ser utilizada por outros algoritmos de recomendação para análise de sua performance, sobretudo com relação a esparsidade da base de dados.

5.2
Recomendador baseado em conteúdo

No capítulo 4 foram apresentados os algoritmos de recomendação baseados em conteúdo, tendo sido dada uma atenção especial ao método de recomendação de itens de uma mesma categoria pelo uso de avaliações feitas por usuários sobre componentes de itens com base em números fuzzy.

Esse método foi exposto detalhadamente devido à sua aproximação conceitual com a idéia de posicionamento de produtos que existe em marketing, aproximação esta que representa uma contribuição desta tese para com a comunidade científica. Essa proximidade, se explorada, pode ser utilizada para aperfeiçoar o método original, o que será explicado com maior profundidade nessa seção.

Segundo apresentado no capítulo 2, o processo de geração de especificações de produtos pelas grandes empresas passa pela segmentação de pessoas em nichos de mercado, seleção de um nicho, geração de posicionamentos para o nicho e, então, criação de especificações técnicas com base no posicionamento. Por exemplo, para uma televisão, pode existir um nicho de mercado para o qual o posicionamento *imagem* é fundamental, e isso gera a criação de televisores onde a especificação *tamanho da tela* e *HDMI* são prioridade.

Enquanto os analistas de marketing partem dos posicionamentos para gerar especificações, o método proposto por Hosseinpour faz o procedimento inverso: gera, a partir de valores de especificações técnicas dos produtos, um número fuzzy denominado por ele “componente” que, a partir deste momento, será interpretado a partir desta tese como posicionamento do produto. Mais do que isso, o método cria uma forma de comparar posicionamentos entre produtos diferenciados e

propõe que o usuário defina o seu posicionamento para que o sistema faça recomendações.

O fato de o conceito de posicionamento em marketing ser bem definido e explorado por inúmeros fabricantes de produtos no mundo inteiro desde a sua concepção dá uma credibilidade ao método de recomendação por posicionamentos que não foi explorado por Hosseinpour em seu artigo.

Além disto, as relações entre o método original proposto por Hosseinpour e o conceito de posicionamento permitem que se aperfeiçoe o método original com novos conceitos importantes que serão apresentados a seguir.

5.2.1
Posicionamentos Compostos

O desejo de compra de um consumidor por vezes esconde posicionamentos mais complexos do que simplesmente o custo ou a qualidade de um produto. Quanto mais próximo o produto chega ao real desejo do consumidor, mais provável será a compra, porém alguns posicionamentos não são de fácil obtenção a partir apenas das especificações técnicas do produto. Um exemplo seria a necessidade de status social. Diversos atributos diferentes de um celular podem trazer status, porém seria mais simples definir status por um conjunto de posicionamentos. No caso do celular, quanto maior a presença de especificações que gerem percepção de design, de tecnologia e quanto maior o seu preço, potencialmente maior será o status por ele gerado. Da mesma forma, um computador pode se posicionar como "o melhor para jogos de videogame", e este posicionamento pode ser calculado diretamente dos posicionamentos de gráfico, som e capacidade de processamento.

Posicionamentos podem ser unidos para se chegar a novos posicionamentos de maior grau de abstração e valor para o cliente. Dado um posicionamento de nível k $\tilde{p}^{k}$ composto por um conjunto de *t* posicionamentos de nível menor $\tilde{P}^{k-1} = (\tilde{p}_1^{k-1}, \tilde{p}_2^{k-1}, ..., \tilde{p}_t^{k-1})$, cada qual sendo um número fuzzy triangular e pertencente ao mesmo item *I*, pode-se calcular o valor de $\tilde{p}^{k}$ por:

$$p^k = \sum_{j=1}^{k} (p_j^{k-1} X \quad w_i) \qquad 18$$

onde w_i é o peso de cada posicionamento de nível inferior a ser multiplicado pelo valor do posicionamento de nível inferior. Os pesos e a combinação de posicionamentos podem ser definidos por um especialista ou por pesquisas de opinião. A equação 5.1 é similar a que permite gerar componentes por meio de especificações, presente no artigo de Hosseinpour, porém a composição de posicionamentos para gerar posicionamentos novos de nível maior de abstração é uma contribuição desta tese.

Por meio deste cálculo, pode-se potencialmente criar posicionamentos complexos seguindo uma linha de pensamento semelhante às dos projetistas dos produtos ao criar suas especificações, indo de níveis mais altos de valor ao cliente para os mais baixos.

À medida que são criados posicionamentos mais complexos a partir de posicionamentos mais simples utilizando este método, existe um risco de os números fuzzy se tornarem ainda mais fuzzy, com um domínio grande demais. Por este motivo, não é recomendável subir muito o nível de abstração dos posicionamentos.

5.2.2
Posicionamentos de Marcas

Um ponto relevante a considerar em posicionamento de produtos é a sua marca. Empresas investem milhões para se diferenciar dos competidores e isso gera um posicionamento implícito para todos os produtos por elas produzidos. Por exemplo, a empresa Adidas financia esportes e remunera atletas para utilizarem seus produtos. Ela buscou desde cedo se posicionar como uma marca esportiva, para jovens que almejam destaque em esportes. Por mais que a característica de um tênis da Adidas não favoreça atividades físicas, é pouco provável que um consumidor não o associe a esporte.

Uma fonte de posicionamentos são os das próprias marcas dos itens, que podem ser considerados quando for criado o vetor de posicionamento. Além disto, ao se calcular os valores dos posicionamentos, faz-se uma média ponderada do

valor final calculado com o valor de posicionamento base da marca do produto ou do posicionamento do próprio estabelecimento comercial onde ele está sendo vendido. Uma camisa sendo vendida em uma loja de produtos populares dificilmente será vista pelo consumidor como um produto de elite.

Da mesma forma, a loja onde um produto costuma ser vendido dá uma boa base para definir o posicionamento do próprio cliente com relação ao produto. Um cliente que compra roupas apenas em lojas populares está se posicionando para itens de vestuário como uma pessoa em busca do menor custo e que valoriza pouco a qualidade do tecido ou o design da roupa. Tanto a marca do produto quanto a marca da loja dizem muito sobre os posicionamentos do produto e dos usuários.

A partir deste conceito, pode-se usar a média dos posicionamentos dos produtos vendidos por lojas pertencentes a uma marca para se criar um posicionamento global da marca. A partir deste posicionamento global, pode-se recomendar as lojas em que o consumidor terá interesse de comprar.

5.2.3
Posicionamentos de Usuários

O método originalmente proposto por Hosseinpour considera a geração de componentes apenas para os itens, deixando os usuários definirem os seus componentes próprios individualmente por meio de uma interface gráfica.

Uma das implicações da interpretação do método original utilizando o conceito de posicionamento de marketing é que se entende que o usuário tem também posicionamentos que são definidos a partir de uma segmentação por nichos de mercado.

Existem diversas formas de segmentação de usuários em nichos, algumas utilizando métodos de *clusterização* ou classificação. Esses métodos tendem a se aproveitar de características dos próprios usuários, como sexo, idade, ou classe social, para separá-los. Atualmente com acesso a redes sociais existem muitos dados de usuários disponíveis para serem processados de forma a fazer tal segmentação.

Da mesma forma que se pode gerar posicionamento de itens com base em suas especificações técnicas, também é teoricamente possível utilizar as

informações do usuário para descobrir como ele se comporta diante de certos posicionamentos. Desta forma, remove-se a necessidade de o usuário utilizar uma interface gráfica para se definir quanto aos itens.

É importante notar que não estamos buscando um sistema de recomendação que necessite destes posicionamentos de usuários para funcionar, apesar de que o uso de redes sociais para obter tais dados ser uma área de pesquisa nova e que pode ser explorado em trabalhos futuros. Tais posicionamentos estão sendo comentados aqui para compreensão posterior do processo com o qual será feita a criação artificial de usuários e a compra de itens por tais usuários na base dados sintética.

5.2.4
Definindo a escolha do recomendador

Dentre os algoritmos apresentados no capítulo 3 e 4 que não compunham algoritmos de filtragem colaborativa, foram destacados os baseados em conteúdo e os baseados em regras de associação.

Os algoritmos baseados em regras de associação foram os primeiros algoritmos desenvolvidos para recomendação e foram amplamente estudados e aplicados. Eles apresentam o problema de gerar muitas regras a um custo computacional enorme e com baixa acurácia, o que sugere a propensão de interferir negativamente nos algoritmos de filtragem colaborativa.

Os algoritmos de recomendação baseados em conteúdo são aplicados de forma híbrida com algoritmos de filtragem colaborativa, conforme explicado no capítulo 3, geralmente utilizando freqüência de palavras existentes em web sites ou especificações técnicas dos itens. O uso constante de algoritmos baseados em conteúdo com filtros colaborativos com bons resultados leva a crer que a abordagem de hibridização pode trazer sucesso ao modelo proposto.

Ao utilizar o algoritmo de posicionamento fuzzy baseado na interpretação do método de recomendação proposto por Hosseinpour e o conceito de posicionamentos de produtos de marketing, hibridizado com o filtro colaborativo, espera-se obter uma melhora na recomendação de itens e uma maior facilidade de definição dos posicionamentos dos produtos com base na opinião de especialistas,

uma vez que o conceito de posicionamento de produtos é de simples análise para aqueles que projetam os itens recomendados.

5.3
Algoritmo de Filtragem Colaborativa

Conforme descrito no capítulo 3, existe uma vasta gama de algoritmos de filtragem colaborativa, sendo essa uma área de ampla pesquisa. Ao analisar os diversos artigos e as propostas presentes no estado da arte, percebe-se que em geral os algoritmos propostos buscam aperfeiçoar o algoritmo baseado em memória utilizando *Correlação Pearson*. A Tabela 4, apresentada no capítulo 3, oferece uma visão geral das vantagens e desvantagens dos algoritmos e auxilia na decisão de qual algoritmo deve ser utilizado para o modelo proposto neste trabalho.

Recorde-se que o objetivo deste trabalho é desenvolver um modelo que atue sobre bases de dados de lojas em que os itens são multi-categóricos e os dados são esparsos e de grande escala. Além disto, lojas virtuais contam com a entrada quase diária de itens novos e o algoritmo deve ser capaz de se adaptar a esta condição.

Observando a Tabela 4, percebe-se que os algoritmos baseados em memória têm as vantagens de permitir a entrada de novos itens de forma incremental e boa escalabilidade; o seu principal problema é a degradação do desempenho frente a dados esparsos e à recomendação de itens a usuários novos. Os algoritmos baseados em modelo atuam bem com bases esparsas e de grande escala e apresentam um melhor desempenho de predição, porém são lentos em seu treinamento, desempenham mal na escala e perdem informação caso sejam utilizadas técnicas de redução de dimensões, se a medida não for *lower bounding* [85].

Uma vez escolhido o sistema de recomendação fuzzy, que é um sistema baseado em conteúdo, para hibridização com o algoritmo de filtragem colaborativa, pode-se esperar que, a exemplo de outros sistemas de recomendação baseados em conteúdo, este complemente a filtragem colaborativa na análise de itens e usuários novos e nos problemas de esparsidade.

Outro ponto importante é que boa parte dos algoritmos apresentados no capítulo 3, sobretudo os baseados em modelo, foram analisados com a base de

dados *MovieLens*, que possui cerca de dez mil itens e setenta mil usuários, o que é uma quantidade muito inferior à de uma base de dados real de lojas virtuais com milhões de usuários e vários milhões de itens. O algoritmo de filtragem colaborativa baseada em memória item-para-item, por outro lado, é utilizado diariamente em uma base de dados deste porte, no site da *Amazon*.

Como o algoritmo de filtragem colaborativa baseado em memória item-para-item é utilizado com sucesso em recomendações comerciais de grande porte (*Amazon*) e como serve, em diversos artigos, como base de comparação entre sistemas de recomendação, ele será aqui utilizado na hibridização que formará o algoritmo proposto e também como *benchmark*. Sendo assim, chamaremos o filtro colaborativo de benchmark "FC Item-Item".

5.4
Hibridização

Embora já tenham sido definidos os algoritmos de filtragem colaborativa e baseado em conteúdo que farão parte da hibridização proposta, a forma como esta se dará é fundamental para garantir um melhor resultado no processo de recomendação.

No capítulo 3 foram apresentados diversos modos de hibridização entre algoritmos de filtragem colaborativa e de recomendação baseada em conteúdo, como, por exemplo, a possibilidade de se misturarem avaliações por meio de pesos de ambos os algoritmos e utilizá-los de forma seqüencial. A definição de como esses algoritmos serão utilizados deve levar em conta suas vantagens e desvantagens.

Conforme mencionado anteriormente, o algoritmo de filtragem colaborativa baseado em memória tem como seu principal problema a degradação do desempenho frente a dados esparsos e à recomendação de itens e usuários novos. Os filtros colaborativos baseados em itens foram desenvolvidos para resolver problemas de escalabilidade existentes nos recomendadores baseados em usuários. Sendo M o número de usuários e N o número de itens, o cálculo off-line da similaridade entre itens é intenso em tempo, com ordem de grandeza $O(N^2M)$ no pior cenário. Como muitos usuários compram poucos itens, o tempo tende a ser de uma ordem de grandeza $O(NM)$, mas isto acarreta problemas de esparsabilidade.

O algoritmo conta com a vantagem de ser incremental (adicionar itens novos é simples) e com a vantagem de não depender de informações dos itens para recomendações, apenas contando com os votos dos usuários.

O método de recomendação por posicionamento fuzzy, por sua vez, possui algumas desvantagens. A primeira é que necessita de um especialista para gerar os posicionamentos com base nas especificações dos itens. O segundo problema está no fato de ter sido desenvolvido para itens de uma mesma categoria, o que impede recomendações cruzadas entre itens de categorias diferentes. O terceiro problema é exigir que o usuário defina o tipo de posicionamento para fazer a recomendação. Por fim, esse algoritmo necessita gerar a similaridade de cada usuário para cada item, apresentando uma função $O(NM)$. Por outro lado, o algoritmo está apoiado em conceitos de marketing que torna simples e intuitiva a definição dos posicionamentos tanto para usuários como para especialistas. O algoritmo também conta com uma taxa de acerto elevada quando testado em itens de uma mesma categoria e resolve o problema de itens novos e usuários novos.

Dada a desvantagem do algoritmo de recomendação por posicionamento fuzzy de não prever recomendações entre categorias distintas, a possibilidade de utilizar combinação de similaridades entre ambos os algoritmos se torna mais difícil. Poder-se-ia tentar criar uma relação entre os posicionamentos de categorias diferentes, mas esta não seria global o suficiente para todas as categorias possíveis, de forma que se descarta de pronto a possibilidade de combinação entre similaridades.

Outra possibilidade que se apresenta na literatura para hibridizar é a troca [59], onde os algoritmos se alternam para obter maior ganho de suas vantagens. Como o filtro colaborativo não apresenta problemas de recomendação entre categorias, a opção de utilizá-lo primeiramente apresenta-se como a mais viável. Assim, propõe-se que o filtro colaborativo baseado em memória categoria-a-categoria (FC Categórico) faça a escolha da categoria a ser comprada e o método de recomendação fuzzy (Método Fuzzy) tome a decisão final do item a ser comprado, em uma hibridização em cascata, conforme visto na Figura 15.

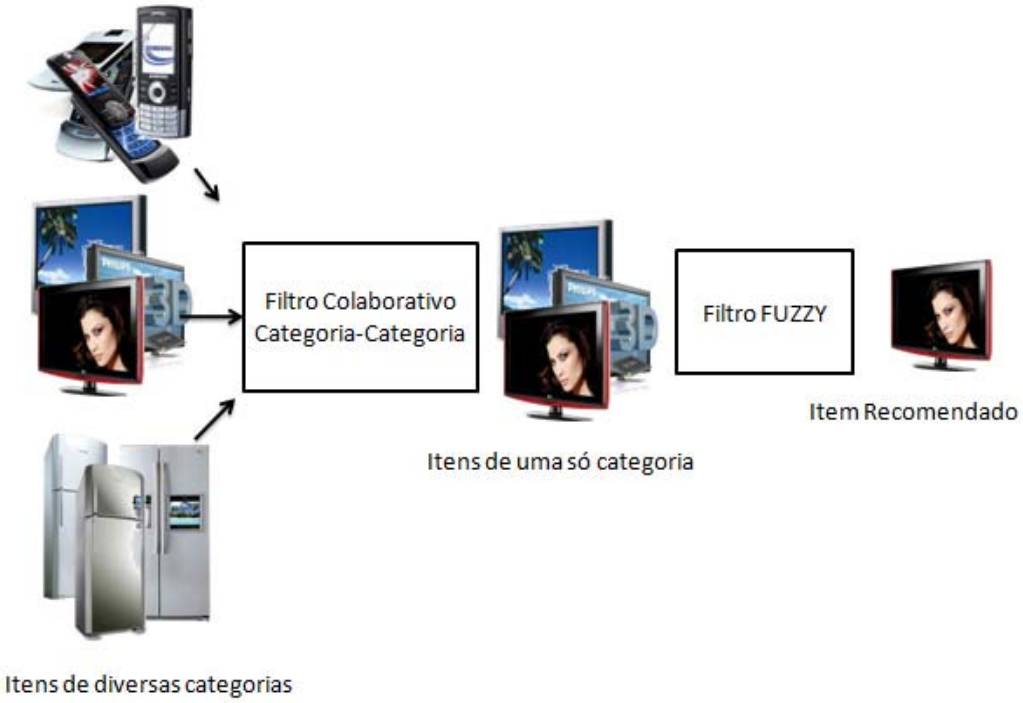

Figura 15 – Processo híbrido de filtragem em duas etapas.

A vantagem deste modelo está em aproveitar ao máximo as características positivas de ambos os algoritmos. Ao manter o filtro colaborativo atuando apenas em categorias, reduz-se em muito o problema de esparsidade e escalabilidade. Em bases de dados varejistas, pode ser que um usuário jamais avalie uma porcentagem significativa dos itens, porém ele certamente avaliará itens de categorias distintas em quantidade suficiente com relação ao total de categorias disponíveis, de forma que a reduzir os problemas de esparsidade. Desta forma, é possível descobrir relações do tipo "quem compra televisão também compra DVD". Da mesma forma, ao adicionar um novo usuário, já nas primeiras compras haverá boas avaliações sobre as suas categorias preferidas na loja em comparação a outros usuários, e novas categorias poderão ser avaliadas rapidamente também.

A ordem de grandeza do tempo de treinamento do FC Categórico, assim como o do FC Item-Item é $O(C^2M)$, onde C é o número de categorias. Como no caso do FC Categórico a base de dados é menos esparsa, a ordem de grandeza se mantém.

Uma vez que o FC Categórico recomende a categoria mais apropriada para o usuário, é acionado o Método Fuzzy (Figura 15), baseado em posicionamento de números fuzzy. Este apresenta ao usuário os posicionamentos disponíveis para a

categoria para que ele se posicione e, a partir destes posicionamentos, recomenda o item mais apropriado dentre todos da categoria. Assim, o Método Fuzzy não sofre com o problema de ter de escolher a categoria a apresentar, podendo aproveitar sua elevada taxa de acerto para dar a recomendação final do item ao usuário, porém o Método Fuzzy ainda apresenta um problema para hibridização que é a necessidade do usuário se posicionar.

5.5
Aperfeiçoando o Método Fuzzy

O método de recomendação baseado em números fuzzy apresentado no capítulo 4 considera a apresentação ao usuário de um questionário para a escolha de seus "posicionamentos" (Figura 13). Essa abordagem traz problemas uma vez que o usuário pode não dispor de tempo ou ter interesse em preencher tal questionário. No tópico 5.3.2 deste capítulo, foi proposto utilizar os dados do próprio usuário para posicioná-lo, porém mesmo essa técnica é pouco eficaz em lojas virtuais, que não contam com muitos dados além do nome e do cartão de crédito do comprador. Utilizar dados de redes sociais pode ser uma solução a ser estudada para o problema, porém nesta tese buscamos uma solução que apenas os dados de compra do cliente.

O que as lojas possuem é uma lista grande de itens comprados anteriormente pelo cliente, o que é em parte o motivo pelo qual o filtro colaborativo baseado em memória funciona bem para este tipo de problema, pois usa apenas listas de compras para fazer sua recomendação.

O Método Fuzzy apresenta a vantagem de se aproveitar da quantidade de dados disponível para gerar melhores recomendações. Se o usuário se dispuser a preencher um questionário, a recomendação deverá alcançar os 80% de acurácia previstos por Hosseinpour. Caso haja apenas dados de itens de compras, cujos posicionamentos são calculados a partir das especificações como explicado anteriormente, pode-se calcular um posicionamento médio dos itens comprados e considerar que este posicionamento médio corresponde ao posicionamento do usuário. Por outro lado, itens de categorias diferentes podem ter posicionamentos diferentes. Deste modo, propõe-se para o cálculo do posicionamento do usuário o algoritmo exposto no Quadro 1:

Quadro 1 – Definição de posicionamento do usuário baseado em itens comprados.

Dada uma categoria cat escolhida:

Para cada item comprado Ic:

Armazenar posicionamentos P_{com} em comum a todos itens

Armazenar posicionamentos P_{cat} específico da categoria

Fazer média aritmética fuzzy dos posicionamentos em comum $\overline{P_{com}}$

Fazer média aritmética fuzzy dos posicionamentos da categoria $\overline{P_{cat}}$

$P_{Final} = \overline{P_{com}} + \overline{P_{cat}}$

Gerar similaridade entre P_{Final} e os posicionamentos de todos os itens disponíveis a recomendação

Por exemplo, digamos que um usuário tenha comprado no passado os seguintes itens (com seus posicionamentos representados por números fuzzy):

Televisão (P_{preco}=alto P_{imagem}=alto P_{design}=baixo)

Geladeira (P_{preco}=muito alto $P_{tamanho}$=médio P_{design}=baixo)

Laptop (P_{preco}=alto P_{imagem}=baixo P_{design}=baixo $P_{tamanho}$=médio)

Este mesmo usuário deseja comprar um Videogame, cujos posicionamentos são: P_{preco}, P_{imagem} e P_{design}:

Então:

Posicionamentos em comum a todos os itens comprados: Preço e Design

O posicionamento de Imagem não é comum a todos, mas é comum a Televisão e Laptop.

Logo, faz-se uma média aritmética dos números fuzzy de posicionamento da televisão, geladeira e laptop para gerar o posicionamento comum de preço e design. Faz-se também a média aritmética dos posicionamentos de imagem do laptop e da televisão para formar o posicionamento categórico de imagem.

O conjunto final de posicionamento do usuário incluirá posicionamentos de Preço, Design e Imagem médios para comparar com outros itens da categoria videogame.

De forma similar com que o filtro colaborativo baseado em memória utiliza a Correlação Paerson como similaridade entre itens, pode-se utilizar os posicionamentos do usuário (que é a média dos posicionamentos de todos os itens

comprados por ele) para, através da equação de compactação próxima (Eq. 17), calcular a similaridade entre um novo item a ser recomendado e o posicionamento do usuário. Para este método, que é uma contribuição relevante desta tese, chamamos de Filtro Fuzzy.

Propõe-se então que ao invés de utilizar o Método Fuzzy, onde o usuário define os seus posicionamentos manualmente, utilize-se um Filtro Fuzzy, no qual os posicionamentos do usuário são calculados com base nas compras passadas e os itens a serem recomendados tem sua similaridade calculada com a equação de compactação próxima (Eq. 17).

O Filtro Fuzzy requer a comparação dos posicionamentos de cada item em uma dada categoria com a média dos posicionamentos das compras feitas pelo usuário. O cálculo da média de posicionamentos tem uma ordem de grandeza O(MP) onde P é o número total de posicionamentos possível para o item, um número muito baixo comparado a M. Em seguida, deve ser feita a operação de compactação próxima entre o posicionamento médio de cada usuário e cada item na base de dados. Essa operação tem uma ordem de grandeza de, no máximo, $O(MNP_m)$ onde P_m é a quantidade de posicionamentos máxima de uma categoria.

Ao filtro colaborativo híbrido, composto pelo FC Categórico em cascata com o Filtro Fuzzy, dá-se o nome de Filtro Hibrido Fuzzy. Há de se notar que o Filtro Fuzzy é uma adaptação do método de Hosseinpour para ser utilizado como um filtro colaborativo. O método de Hosseinpour apenas compara posicionamentos selecionados por usuários em uma interface de usuário com os posicionamentos dos itens, obtidos a partir de suas especificações. O Filtro Fuzzy usa o conceito do Filtro Colaborativo baseado em itens para comparar itens comprados com outros itens comprados pelo usuário, utilizando a medida de compactação próxima (Eq 17) como similaridade fuzzy entre seus posicionamentos no lugar da similaridade cosseno comumente utilizada. O Filtro Hibrido Fuzzy é uma contribuição importante da tese que será testada no próximo capítulo com o FC Item-Item para gerar conclusões quanto a sua eficácia com relação a esparsidade e acurácia.

A própria hibridização com o Filtro de Categoria necessita de uma atenção especial uma vez que o método de Hosseinpour não leva em conta comparação entre itens de categorias distintas e uma adaptação no algoritmo, a ser explicada mais a frente, se faz necessária para que a hibridização seja possível.

O Filtro Híbrido Fuzzy, por ser uma composição de dois algoritmos, tem um tempo computacional sendo a soma de ambos. A ordem de grandeza do tempo computacional passa a ser $Max(O(N^2P) + O(MP)+ O(MNP_m))=O(MNP_m)$, sendo um pouco maior do que o do FC Item-Item.

5.6
Redução do Espaço de Busca

A hibridização proposta possui uma ineficiência quanto ao uso do recomendador fuzzy nos itens, uma vez que este requer que se tenha a similaridade entre os posicionamentos entre itens para todos os itens. Se, por um lado, esse processamento é realizado off-line, por outro ele toma um tempo elevado devido à grande quantidade de itens.

Uma solução para este problema advém do estudo de comportamentos orçamentários do consumidor. Em tabelas como a apresentada no capítulo 2 (Tabela 1) determina-se um limite superior de gastos do consumidor em categorias de consumo. Cruzando essa informação com informações do consumidor, como sua classe social e compras passadas, é possível filtrar itens que simplesmente estão fora da capacidade de compra presente de um consumidor, reduzindo-se, portanto, o espaço de busca e a necessidade de cálculo de todos os itens.

Obter dados orçamentários do consumidor, ou sua classe social, só é possível após um usuário do banco de dados ter feito muitas compras, necessitando, portanto, da acumulação de informações de compra do usuário em um longo espaço de tempo. A compra do usuário é uma pista de sua classe social e orçamento. Uma pessoa que compra em lojas caras pode ser classificada como de uma classe social mais alta do que aquela que só compra em casas populares. Há de se lembrar que o Marketing de Relacionamento considera a formação de bancos de dados deste tipo (CRM) onde o máximo de informação do usuário é armazenada.

Por exemplo, considere-se um cliente de classe média baixa (IBGE: orçamento entre R$ 2.150,00 a R$ 3.821,00). Seguindo a Tabela 1, esse cliente gasta 1.9% de seu orçamento com eletrônicos, o equivalente a 72 reais mensais no

máximo. Sendo assim, apresentar uma televisão ou DVD com custo mensal de 300 reais por mês teria pouca chance de sucesso na venda.

Tendo menos itens a avaliar, pode-se potencialmente reduzir o tempo de processamento na busca por recomendações, sem sacrificar a qualidade da recomendação.

5.7
Avaliando o Modelo

Ao se analisar o estado da arte, percebe-se que a grande maioria dos algoritmos é avaliada em comparação com o algoritmo de filtragem colaborativa baseado em memória. Outra característica de grande quantidade dos documentos é o uso da base de dados *MovieLens* como padrão para comparar modelos. Por fim, percebe-se que os sistemas de recomendação são avaliados por meio de métricas de acurácia preditiva, tais como o *Erro Médio Absoluto* (MAE) e suas variações, métricas de acurácia de classificação, como *precisão*, *medida-F1* e *sensitividade ROC*, e métricas de acurácia de ranking, tal como *Correlação Produto-Momento de Pearson*, *Tau de Kendall e Precisão de Média* (MAP).

Para avaliar o modelo de recomendação híbrido fuzzy proposto neste trabalho, será utilizado como base de comparação o algoritmo de filtragem colaborativa baseado em memória item-para-item, acompanhando o padrão visto na literatura.

Quanto às medidas de desempenho, serão utilizadas métricas de precisão (*precision*) e revocação (*recall*). Para utilizar essas métricas, as compras de cada usuário foram divididas de forma que parte delas gere um conjunto de treinamento (CTr) e as últimas, o conjunto de teste (CTe). Cria-se, a partir do conjunto de treinamento, um conjunto de recomendações (top-N) de tamanho variável. Por fim, os itens que pertencem tanto a top-N e a CTe formam o conjunto de acertos (CA). Assim, definem-se as métricas:

$$revocação(\mathrm{Re}) = \frac{|CTe \cap top-N|}{|Cte|} \qquad 19$$

$$precisão(\Pr) = \frac{|CTe \cap top - N|}{N} \qquad 20$$

A avaliação final é dada pela métrica F1:

$$F1 = \frac{2 * \Pr * \mathrm{Re}}{\Pr + \mathrm{Re}} \qquad 21$$

Para os testes, serão utilizadas duas bases de dados. A primeira será a *MovieLens*, bastante utilizada para avaliar sistemas de recomendação, e que possui 100.000 avaliações para 1682 itens (filmes), feitas por 943 usuários. A base *MovieLens* possui muito poucas informações sobre os filmes (data de lançamento e gênero) e, portanto, será utilizada para testar o Filtro Fuzzy (seção 5.4). Como o gênero é uma das poucas informações disponíveis para treinar o Filtro Fuzzy, foi decido que essa informação não poderia ser utilizada também para testar um filtro de categorias. Consideramos os gêneros como especificações dos filmes e que Filmes é uma categoria por si só, o que impede o emprego do filtro de categorias nestes experimentos.

Com base nos gêneros dos filmes, são formados os posicionamentos (Figura 16):

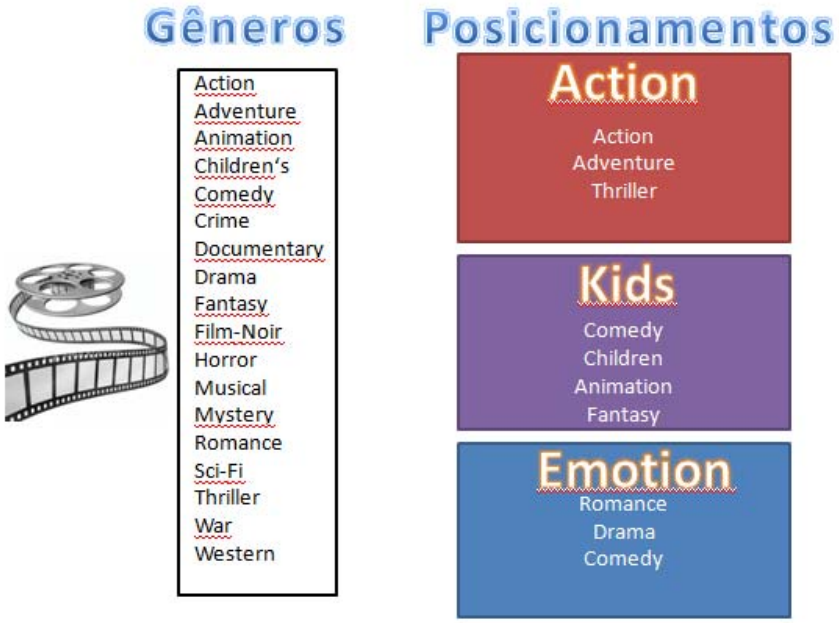

Figura 16 – Filmes e Posicionamentos

A segunda é uma base de dados sintética, criada para testar o modelo proposto quanto à esparsidade e escala. A base de dados sintética foi criada com um número de usuários e de itens comparáveis aos utilizados nas lojas virtuais. Todos os usuários e itens da base de dados sintética foram criados com posicionamentos baseados em suas especificações técnicas (no caso dos itens) ou características demográficas (no caso de usuários). As compras foram feitas usando a similaridade entre os posicionamentos dos usuários e os posicionamentos dos itens, de acordo com o que se espera de usuários reais segundo a teoria de marketing.

Conforme foi visto no capítulo 3, a base sintética é uma aproximação da realidade, cujos resultados servem para comparar características de recomendadores, mas que podem ser tendenciosos quanto a comparação de sua acurácia. Deste modo, essa base, apesar de ser uma importante contribuição desta tese, teve sua análise voltada principalmente a resultados quanto a esparsidade da base de dados, pois a comparação de acurácia entre algoritmos pode ser considerada tendenciosa.

5.8
Esquema das Bases de Dados

O esquema do banco de dados é uma coleção de objetos e de suas correlações. Os objetos do esquema são estruturas lógicas que se referem diretamente aos dados do banco. Incluem estruturas tais como tabelas, visões, seqüências, procedimentos armazenados, sinônimos, índices, agrupamentos e links de bancos de dados. O esquema é o mesmo para a base de dados sintética e para a base de dados real, sendo que na primeira, os dados são criados artificialmente e na segunda são criados por interação real com clientes e itens reais.

Três objetos estão presentes em bancos de dados de lojas virtuais:

- Cliente – entradas no banco de dados representativas dos clientes cadastrados que farão compras. O cliente possui dados que os diferenciam, como idade, sexo, moradia e situação civil.
- Itens – entradas no banco de dados de cada produto ou serviço à disposição para ser vendido. Os itens são categorizados com base em

características semelhantes e costumam ter descrições, especificações técnicas e marca.

- Compras – tabelas que conectam usuários aos produtos comprados, constituindo-se em um histórico de compras do usuário. Podem incluir informações como o preço de venda, as unidades vendidas, o modo de compra, o vendedor, etc.

A esses objetos presentes em quaisquer lojas virtuais são adicionados alguns objetos que dão suporte às regras de compra:

- Categoria de Item – grupamentos de itens com o mesmo tipo de funcionalidade. Ex: Televisão, Gravador, Sofá, Eletrodoméstico, etc. As categorias podem ser hierárquicas, com algumas pertencentes a outras mais abrangentes. Por exemplo, a categoria televisão e gravador pertencem a "eletrodomésticos".
- Posicionamento de Item – objetos que armazenam, para cada item, os números fuzzy com os posicionamentos do item com relação à categoria a que ele pertence. Esse objeto foi descrito detalhadamente no capítulo 4.
- Posicionamento de Cliente - objetos que armazenam os interesses de compra do cliente com relação a uma categoria de produto.

O esquema apresentado na Figura 17 apresenta as principais entidades descritas acima e as ligações entre elas, incluindo suas quantidades. Por exemplo, a relação entre o cliente e a compra é tal que aquele pode fazer múltiplas compras (representadas por *) e cada uma se relaciona com um único usuário (1). Há de se notar que o banco de dados final pode possuir diversos outros objetos, buscando otimização de consultas. O importante para este estudo é que tais objetos existam e que os posicionamentos de itens clientes sejam criados segundo os algoritmos estudados nos capítulos 3 e 4, respectivamente.

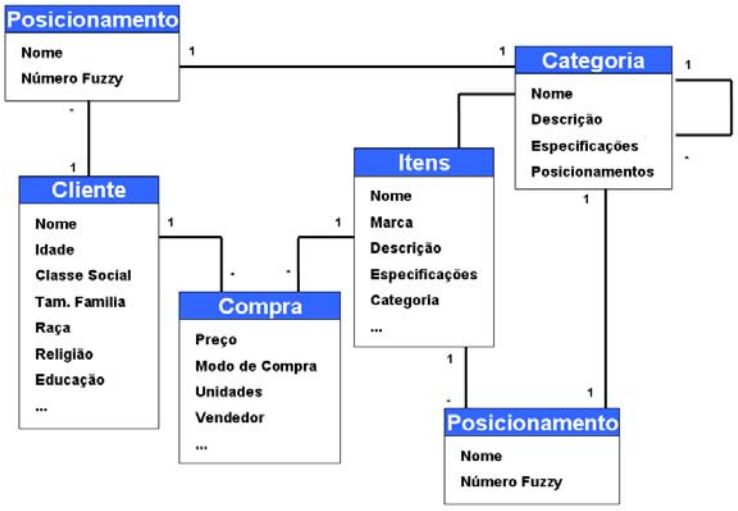

Figura 17 – Esquema básico do banco de dados sintética

5.9
Definição das Bases de Dados

Ambas as bases de dados, sintética e real, utilizadas para testar o modelo proposto, são criadas utilizando o mesmo esquema apresentado na seção anterior. A base de dados *MovieLens* foi adicionada como tendo todos os itens em uma única categoria, de nome *Movies*. O motivo de uso de apenas uma categoria para a base de dados *MovieLens* é devido a pouca quantidade de atributos existente para cada filme nessa base, o que levou a opção de "Gênero" do filme ser utilizado como especificação do filme e não categoria. As bases de dados nada mais são do que uma simulação de uma loja virtual, com clientes comprando itens. A diferença básica entre as duas bases de dados é que, na base sintética, as regras de compra são controladas e os usuários e itens são artificiais, enquanto que na base real os usuários são pessoas reais fazendo compras de itens (filmes) reais. Outra diferença fundamental é que na base sintética pode-se criar um número de usuários e itens voltados a testar efeitos de escala e esparsidade, o que não é possível na base real criada pela *MovieLens,* onde estas quantidades são fixas.

Uma base de dados real, como a *MovieLens*, contém regras de compra desconhecidas, baseadas no interesse de cada cliente que entra na loja virtual e compra um produto (no caso, filmes). As compras efetuadas na base de dados serão divididas em um conjunto de treinamento e um de teste. A avaliação da

acurácia do algoritmo refletirá sua capacidade de prever uma compra no conjunto de testes apenas com as informações disponíveis no processo de treinamento.

A base de dados sintética é composta também pela compra de diversos clientes e produtos, mas as regras de compra são previamente definidas. É composta pela criação aleatória de clientes com características demográficas diferentes, o que gera perfis de compra diferentes. Cada cliente compra de acordo com o seu perfil, definido por regras pré-estabelecidas, baseados no seu posicionamento (oriundo dos dados demográficos) e dos posicionamentos dos itens (oriundo de suas especificações técnicas), de forma similar ao que espera-se da teoria de marketing.

Como na base sintética os usuários são criados artificialmente, é possível tê-los em quantidade semelhante ao de bases de dados reais, de forma a produzir resultados suficientes para uma avaliação do modelo proposto. No caso, utilizamos inicialmente razões entre usuários e itens de forma similar ao de bases de dados reais como das gigantes varejistas: *Submarino* e *Americanas.com*. A razão entre usuários e itens variaram de acordo com os experimentos para causar esparsidade na base de dados e analisar a performance dos algoritmos ante a estas esparsidades.

A seguir serão apresentados detalhamentos da construção da base de dados sintética, para que seja possível a avaliação e repetição deste processo em futuros estudos.

5.9.1
Regras de Criação

A base de dados sintética é preenchida por dados criados a partir de regras definidas. Tais regras poderão ser mais tarde utilizadas para comparar resultados de algoritmos de recomendação diversos quanto à sua capacidade de aprendizado.

Cada objeto presente no esquema do banco de dados deverá ter sua regra própria para criação. Isso permite que todos os objetos sejam criados automaticamente e que se tenha uma base extensa para testes de desempenho dos algoritmos e para treinamento.

5.9.1.1
Regras para Clientes

As regras para clientes só existem no banco de dados sintético. Foram criadas de modo a preencher todas as propriedades demográficas do objeto de forma aleatória dentro de limites pré-estabelecidos. Por exemplo, o valor da idade foi definido como um número aleatório entre 14 e 100 anos. Algumas regras especiais foram feitas de forma a correlacionar propriedades, como, por exemplo, a idade e o nível de escolaridade. Também foram empregadas pesquisas estatísticas regionais para garantir que a amostra criada tenha um perfil realista da população de uma região. Por exemplo, a Tabela 7 obtida na ABEP (Associação Brasileira de Empresas de Pesquisa) apresenta a divisão demográfica da população brasileira em classes sociais. Estados foram considerados na base de dados para definir a classe social e, portanto, o orçamento dos usuários. Diversos outros estudos estatísticos foram utilizados a partir de informações do IBGE e da ABEP, podendo a base de dados ser criada, inclusive, para apresentar características regionais. Por exemplo, pode-se usar dados da região nordeste para tratar o sistema de recomendação voltado a esta região.

A quantidade de Clientes criados é variável, podendo ser usado para aumentar a escala do banco de dados, assim como a relação entre clientes e itens afeta a esparsidade do banco. Para a base de dados sintética usada nos experimentos, foram criados entre 5000 e 20.000 usuários, dependendo da esparsidade desejada.

Tabela 7 – Classes sociais no Brasil – ABEP 2010.

Classe	Renda Média (por pessoa)		Renda Família (4 pessoas)		% da populaç
Classe A 1	R$ 9.733,47		R$ 38.933,88		1%
Classe A 2	R$ 6.563,73	R$ 9.733,47	R$ 26.254,92	R$ 38.933,88	4%
Classe B 1	R$ 3.479,36	R$ 6.563,73	R$ 13.917,44	R$ 26.254,92	9%
Classe B 2	R$ 2.012,67	R$ 3.479,36	R$ 8.050,68	R$ 13.917,44	15%
Classe C 1	R$ 1.194,53	R$ 2.012,67	R$ 4.778,12	R$ 8.050,68	21%
Classe C 2	R$ 726,26	R$ 1.194,53	R$ 2.905,04	R$ 4.778,12	22%
Classe D	R$ 484,97	R$ 726,26	R$ 1.939,88	R$ 2.905,04	25%
Classe E	R$ 276,70	R$ 484,97	R$ 1.106,80	R$ 1.939,88	3%
Média	**R$ 1.500,00**	**R$ 2.500,00**	**R$ 6.000,00**	**R$ 10.000,00**	**Média**

5.9.1.2
Regras para Categorias de Item

As categorias representam agrupamentos de itens com características semelhantes. Cada categoria nova criada deverá possuir as especificações técnicas relevantes de todos os itens a ela pertencentes, bem como limites superiores ou inferiores para essas especificações, e os números fuzzy relativos a essas especificações. Exemplo: todos os televisores possuem especificação de tamanho da tela, podendo variar entre 10 a 100 polegadas. Caso o tamanho da tela seja inferior a 20, atribui-se o número fuzzy *muito baixo*; para valores entre 20 e 30, *baixo*, e assim por diante.

Cada categoria também deve possuir uma lista de posicionamentos comum a todos os itens que ela representa, assim como os pesos dados para cada especificação técnica na formação do número fuzzy relativo ao posicionamento (Capítulo 4).

As categorias dos itens podem ser criadas de forma automática ou manualmente, considerando-se itens reais com suas especificações técnicas e posicionamentos. No caso da base sintética criada, foram criadas quatorze categorias gerais e uma global.

Para cada categoria foram criados posicionamentos e especificações características dos produtos que fariam parte da mesma. Por exemplo, para geladeira existe a especificação de potência e posicionamento de capacidade.

A categoria global possui os posicionamentos e especificações pertinentes a todos as categorias e permite comparar itens entre categorias diferentes. Foram criadas para ela os seguintes posicionamentos:

- Preço – Define o efeito que o preço do produto possui no interesse de compra do cliente
- Qualidade – Define o nível de risco de defeito que o cliente aceita ao comprar um item.
- Design – Define o quanto o cliente se preocupa com marcas e com o status social que o item oferece.
- Funcionalidade – Define a força que as funcionalidades do produto exercem na decisão de compra.

A forma como esses posicionamentos globais são criados a partir das especificações de cada item dependem de categoria para categoria. Mais categorias podem ser criadas em outras bases de dados de forma a englobar as características específicas dos produtos recomendados, porém sempre criando as especificações e posicionamentos dos itens que serão gerados nessas categorias.

5.9.1.3
Regras para Itens

Cada item criado deve pertencer a uma categoria de item, o que implica na herança de um conjunto de especificações técnicas e posicionamentos definidos pela categoria na qual ele está inserido.

Cada especificação técnica do item criado é preenchida por um valor aleatório com limites superior e inferior definidos pela categoria à qual o item pertence. O valor aleatório resultante definirá também o número fuzzy relativo à especificação técnica, conforme definido na categoria do item.

Exemplo: todos os armários possuem especificação de número de portas, podendo variar entre 1 a 10. Caso o número de portas seja de 1, atribui-se o número fuzzy *muito baixo*; para valor de portas igual a 2, *baixo*, e assim por diante.

Foram criados na base de dados sintética um total de 1.000 itens por categoria, num total de 7.000 itens.

5.9.1.4
Regras para Posicionamento de Itens

Para cada item criado, um conjunto de posicionamentos definido pela sua categoria deve ser igualmente adicionado ao banco de dados. Os valores dos posicionamentos são calculados segundo os pesos de cada especificação técnica do item (obtidos da categoria do item) e dos números fuzzy referentes a cada especificação técnica do item (obtidos no item).

Como foi dito anteriormente, para categoria um conjunto de especificações e posicionamentos são definidos e devem estar presentes em todos os itens

pertencentes aquela categoria. Os valores de cada especificação são calculados aleatoriamente dentro de limites estabelecidos e transformados em números fuzzy.

A partir da equação 15 do Método Fuzzy de Hosseinpour, usando os pesos definidos pela própria categoria, os números fuzzy relativos aos posicionamentos são calculados para cada item criado em cada categoria.

Além disto, alguns posicionamentos podem ser criados a partir de outros posicionamentos, conforme foi discutido na seção 5.2.1 (Posicionamentos Compostos). No caso, o posicionamento de Preço dos itens foi criado com base nos posicionamentos de Qualidade, Design e Funcionalidade, com pesos definidos pela categoria, usando a Eq. 18. Esta modelagem segue a lógica de que produtos de maior qualidade, design e funcionalidade devem possuir preço maior.

5.9.1.5
Regras para Posicionamento de Clientes

Cada cliente terá um conjunto de posicionamentos com relação a cada categoria de itens com base em suas características demográficas individuais, definidas durante sua criação. As regras de criação desses posicionamentos são semelhantes às dos itens, sendo que se consideram as características demográficas dos clientes como especificações técnicas relevantes para o posicionamento. Esta modelagem segue o conceito introduzido nesta tese na seção 5.2.3 (Posicionamento de Usuários).

Por exemplo, para qualquer usuário na base sintética existe a característica demográfica “classe social”. Essa classe social varia, conforme a Tabela 7 de A1 até E. A classe social E é modelada com o número fuzzy “Muito Baixo” enquanto que A1 recebe o número fuzzy “Muito Alto” e assim por diante. As outras características demográficas (sexo, idade, escolaridade, etc) são modeladas de forma semelhante em números fuzzy de maneira semelhantes as especificações dos itens.

Cada categoria, no caso do item, possui as informações de especificações que compõem os posicionamentos, assim como seus pesos. Igualmente para os usuários, as categorias definem quais características demográficas, usando respectivos pesos, formam, usando a equação 15 do Método Fuzzy de Hosseinpour, os posicionamentos dos clientes. Por exemplo, a categoria “celular”

considera que o sexo e a classe social tem relevância na alta no posicionamento de design.

5.9.1.6
Regras de Compra

As regras de compra definem que item um cliente irá comprar, sendo que o mesmo usuário poderá fazer diversas compras diferentes, formando uma cesta de compra. As compras são a base para o aprendizado do filtro colaborativo (Capítulo 3); predizer a compra é o objetivo final do algoritmo de recomendação.

A decisão da compra no caso da base de dados sintética é feita a partir da categoria. São definidos primeiramente vetores de categorias V que são compradas juntas, no caso da base sintética criada para avaliar os filtros, foram criadas as seguintes regras:

- V1={Televisores, Gravadores, Sofás}
- V2={Camas, Armários, Televisores}
- V3={Carros, Celulares, Refrigeradores}
- V4={Gravadores, Sofás, Camas}
- V5={Geladeira, Armários, Sofás}

Também é definido o número de compras N por cliente, sendo que para a base sintética utilizada para avaliação nesta tese foi usado N=7, podendo esse número ser variável. Em seguida, as compras são feitas seguindo o algoritmo presente no Quadro 2.

As regras de compra são baseadas tanto em heurísticas quanto no sistema de posicionamento por números fuzzy, conforme pode ser visto no algoritmo exposto. Primeiro o algoritmo define a categoria que será comprada, segundo um dos vetores de categoria (V1 a V5). Em seguida, dado que se sabe a categoria que vai ser comprada, o algoritmo escolhe o item da base de dados na categoria selecionada usando a Equação 17 de compactação próxima proposto no Método de Hosseinpour, tendo como parâmetro os posicionamentos do cliente e do item para a categoria selecionada.

O fato de se utilizar os posicionamentos fuzzy para gerar as compras, por um lado se aproxima do comportamento de compra real do cliente, dado que considera a teoria de marketing que busca exatamente compreender como os clientes definem que produto comprar. Por outro lado, essa abordagem pode tornar o teste com o Filtro Fuzzy tendencioso. Há de se notar que o Filtro Fuzzy não conhece os dados demográficos do usuário e, portanto, não calcula o posicionamento do usuário real, apenas tentando auferir esse posicionamento por meio das compras feitas anteriormente pelo usuário. Essa diferença é crucial e há de se notar que as semelhanças entre o algoritmo de criação de compras e do Filtro Fuzzy é coerente com a realidade, afinal, a teoria de marketing basicamente diz que você é o que você compra.

Ainda assim, os resultados da base sintética podem ser considerados tendenciosos para o Filtro Fuzzy e, portanto, seus resultados devem ser analisados com parcimônia, focando em analisar a capacidade dos algoritmos de recomendação de atuar sobre dados esparsos e extensos. Para testar a eficácia do Filtro Fuzzy em uma base não-tendenciosa, será feito também testes na base de dados *Movielens* cujas regras de compra são desconhecidas.

Quadro 2 - Algoritmo de Regra de Compra

```
Para cada cliente c_l na tabela Cliente {
    De n=0 a N {
        Se n=0{
            Selecione aleatoriamente uma categoria Cat da tabela
            Categoria de Item;
        }
        Else{
            Liste Vetores de Categorias V que contenham um dos itens
            comprados anteriormente;
            Selecione a categoria Cat de item que não foi comprado
            anteriormente nas listas;
        }
        Obtenha os posicionamentos p_cat do cliente com a categoria;
        Para cada item I_cat pertencente a categoria Cat {
            Obtenha os posicionamentos Ip_cat do item I_cat com a
            categoria;
            Calcule a compactação próxima (Sim) entre Ip_cat e p_cat.
        }
        Compre item com maior compactação próxima (Sim)
    }
```