UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE INFORMÁTICA
PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

BYRON LEITE DANTAS BEZERRA

"UMA SOLUÇÃO EM FILTRAGEM DE INFORMAÇÃO PARA SISTEMAS DE RECOMENDAÇÃO BASEADA EM ANÁLISE DE DADOS SIMBÓLICOS"

ESTE TRABALHO FOI APRESENTADO À PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO DO CENTRO DE INFORMÁTICA DA UNIVERSIDADE FEDERAL DE PERNAMBUCO COMO REQUISITO PARCIAL PARA OBTENÇÃO DO GRAU DE MESTRE EM CIÊNCIA DA COMPUTAÇÃO.

ORIENTADOR: Francisco de Assis Tenório Carvalho

RECIFE, FEVEREIRO/2004

BANCA EXAMINADORA

Prof. Dr. Marcílio Carlos Pereira de Souto
Profa. Dra. Teresa Bernarda Ludermir
Prof. Dr. Francisco de Assis Tenório Carvalho

Agradecimentos

Em primeiro lugar agradeço a Deus por me iluminar sempre em todos os momentos difíceis da minha vida.

Agradeço sobretudo a minha esposa Elaine, minha filha Brícia (que está a caminho neste momento), meus pais Dantas e Graça e minha irmã Lorena por sempre compreenderem os momentos ausentes em vista das atividades exigidas para o êxito do presente trabalho.

Agradeço ao meu orientador e amigo, o Professor Dr. Francisco Carvalho. Graças a ele obtive a direção correta para as pesquisas científicas realizadas nesse trabalho.

Resumo

Sistemas de Recomendação permitem que sites de Comércio Eletrônico sugiram produtos aos consumidores provendo informações relevantes que os ajudem no processo de compra. Para isso é necessária ❶ a aquisição e ❷ a adequada utilização do perfil do usuário. O processo de aquisição pode ser implícito (comprar um livro ou consultar um item em uma loja on-line) ou explícito (dar uma nota a um filme ou recomendar um artigo a um amigo). Já as soluções propostas para o segundo problema podem ser classificadas em duas categorias principais com relação ao tipo de filtragem adotada: Filtragem Baseada em Conteúdo (baseia-se na análise da correlação entre o conteúdo dos itens com o perfil do usuário) e Filtragem Colaborativa (baseada na correlação de perfis de usuários). Tais técnicas possuem limitações, como escalabilidade na primeira abordagem e latência na segunda. Contudo, elas são complementares, o que impulsiona o surgimento de filtragens híbridas, cujo foco é aproveitar o melhor de cada método. Todavia, as filtragens híbridas não superam completamente os problemas principais de ambos os métodos.

A motivação deste trabalho surge do desafio de superar os problemas principais existentes nos métodos de Filtragem Baseada em Conteúdo. Para isso, o trabalho concentra-se no domínio de recomendação de filmes, caracterizado por atributos complexos, como sinopse, e no qual predomina uma aquisição explícita do perfil do usuário. Diante disso, o presente trabalho apresenta um novo método de filtragem de informação baseado nas teorias de Análise de Dados Simbólicos.

Na abordagem proposta o perfil é modelado através de um conjunto de descrições simbólicas modais que sumarizam as informações dos itens previamente avaliados. Uma função de dissimilaridade que leva em conta as diferenças em posição e em conteúdo foi criada a fim de possibilitar a comparação entre um novo item e o perfil do usuário. Para avaliar o desempenho deste novo método foi modelado um ambiente experimental baseado no EachMovie e definida uma metodologia para avaliação dos resultados. Para fins de comparação é utilizada a filtragem de informação por conteúdo baseado no algoritmo dos k Vizinhos Mais Próximos (kNN).

A construção de um ambiente experimental de avaliação do modelo permitiu diagnosticar estatisticamente o melhor desempenho da filtragem baseada em dados simbólicos modais, tanto em velocidade quanto em memória, com relação ao método baseado no kNN.

Palavras-chave: Comércio Eletrônico, Personalização, Sistemas de Recomendação, Filtragem de Informação, Análise de Dados Simbólicos.

Abstract

Recommender systems allow E-commerce websites to suggest products to their costumers, providing relevant information to help them in shopping tasks. In order to do, it is need to ❶ acquire and ❷ to use adequately the user profile. The process of acquiring user preferences can be implicit (buying a book or searching an item in a virtual store) or explicit (giving a grade to some movie or suggest a paper to a friend). The proposed solutions for the 2nd problem could be classified in two main groups concerning the kind of filtering approach, e. g., Content Based Filtering (which is based on the correlation between the user profile and items content) or Collaborative Filtering (which is based on the users profiles correlation). These techniques have restrictions, such as scalability in the first approach and latency in the second one. However, they are complementary, which has motivating hybrid filtering approaches trying to mix the better characteristics of the previous ones. Nevertheless, these ones do not solve the main problems of both filtering methods.

The motivation of this work is the challenge of overcoming the main problems inherent of Content Based Filtering. Therefore, the work focuses on the movie recommendation domain, which is characterized by complex attributes, such as synopsis, and by an explicit acquisition of user profile. For that reason, this work presents a new information filtering method based on Symbolic Data Analysis theories.

In the proposed approach, each user profile is modeled using a set of modal symbolic descriptions that summarize the information taken from a set of items the user has previously evaluated. The comparison between a new item and a user profile is accomplished by way of a new suitable dissimilarity function that takes content and position differences into account. In order to evaluate the performance of this new technique, an experimental environment was designed based on the EachMovie database. Additionally, it was defined a appropriate methodology to interpret the results. The proposed approach has been compared with the information content filtering based on the k Nearest Neighbor algorithm (kNN).

The designed experimental environment and the defined methodology allow to statistically diagnose the information filtering based on modal symbolic data has a better performance in speed and storage than the kNN one.

Keywords: E-commerce, Personalization, Recommendation Systems, Information Filtering, Symbolic Data Analysis.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

INTRODUÇÃO

Não é possível falarmos sobre Sistemas de Recomendação sem que antes alguns aspectos relevantes associados ao Comércio Eletrônico sejam contextualizados.

1.1. CONTEXTO

Segundo Peter Drucker (Drucker 2003), a Internet está para a atual e crescente revolução da informação assim como a estrada de ferro estava para a revolução industrial do século XVII. Em outras palavras, quando Peter Drucker coloca a Internet no mesmo nível da estrada de ferro, ele está querendo afirmar que passamos por mudanças sociais, culturais, políticas e econômicas tão ou mais significativas quanto aquelas da revolução industrial, e isso só é possível devido à infra-estrutura tecnológica propiciada pela Internet, assim como a estrada de ferro servia de infra-estrutura para revolução industrial.

Quando a rede ARPA, considerada o embrião da Internet, foi concebida, ela tinha um propósito muito específico: o intercâmbio de informações estratégicas militares entre pontos do Departamento de Defesa dos EUA, de forma que a informação estivesse descentralizada. Depois foi verificada a possibilidade de uso dessa mesma tecnologia em ambientes de pesquisa, com a finalidade também de troca de informações, só que agora, informações contextualizadas no meio acadêmico. Após essa segunda fase, podemos passar ao cenário atual, em que é incontestável o número de aplicações providas pela Internet em conjunto com a World Wide Web (WWW ou simplesmente Web). Na Web as empresas expõem uma vitrine dos seus produtos e serviços ao consumidor sem que o mesmo precise sequer sair de sua casa. Surge então o Comércio Eletrônico.

As possibilidades do Comércio Eletrônico são inúmeras. À medida que essas novas possibilidades mudam os hábitos de compra do consumidor, surgem não só novas aplicações, mas também novos modelos de negócios. Uma das grandes vantagens dessa realidade para o consumidor é o poder de decisão que ele possui no processo de compra, devido, principalmente, à facilidade que existe em se comparar características e preços dos produtos e serviços entre os concorrentes. Por outro lado, as empresas podem usufruir de uma plataforma que provê informações estratégicas como, por exemplo, o perfil de compra dos seus clientes.

O Comércio Eletrônico também está revolucionando as estratégias de marketing utilizadas pelas empresas. Basicamente, podemos enumerar três abordagens de marketing: marketing em massa, marketing segmentado e marketing *one-to-one*. O marketing em massa é o mais antigo e possivelmente ainda o mais aplicado. Ele se baseia na divulgação dos produtos e serviços para um público único constituído dos mais diversos tipos de pessoas. Essa estratégia normalmente não requer o uso de tecnologias computacionais para sua execução. Por outro lado, as duas outras abordagens de marketing são em geral impraticáveis sem o uso de tecnologias de informação.

O *marketing segmentado* é, do ponto de vista de customização da informação, um meio termo entre a primeira e a terceira abordagem. Nesse caso, a empresa identifica que tipos de clientes teriam uma chance maior de consumir um determinado produto ou serviço. Para isso, a empresa possui normalmente uma base de dados contendo informações dos seus clientes e dos produtos que eles consumiram no passado. Assim, é possível a utilização de tecnologias de suporte a decisão como as tecnologias de Descoberta do Conhecimento [1].

Segundo Jeff Bezos (Easton e Bezos 1998), CEO[2] da Amazon.com, para cada cliente em potencial deveria existir uma loja na Web customizada as suas necessidades. A afirmação de Bezos acompanha a estratégia de *marketing one-to-one*, cada vez mais freqüente no cenário atual. Essa estratégia de marketing possui o foco no cliente, ou seja, os produtos e serviços são apresentados ao cliente na medida que existe um grau de interesse relevante naquela informação. Isso propicia aos clientes uma maior satisfação por receberem um atendimento on-line personalizado e, conseqüentemente, maximizam as chances de venda para a empresa.

[1] Normalmente é usado o termo Mineração de Dados quando se fala em tecnologias para Descoberta do Conhecimento. No entanto, é importante salientar que o processo de Mineração de Dados é apenas uma das etapas no processo de Descoberta do Conhecimento. Consulte Witten e Frank (2000) para mais detalhes.
[2] CEO - *Chief Executive Officer*. Equivale no Brasil ao diretor executivo.

1.2. MOTIVAÇÕES

A tecnologia de suporte ao *marketing one-to-one* é conhecida como Sistema de Recomendação (consulte o Capítulo 2). Basicamente, essa tecnologia permite que os sites de Comércio Eletrônico sugiram produtos aos consumidores provendo informações relevantes que os ajudem no processo de compra.

A importância dos Sistemas de Recomendação pode ser notada a partir de uma afirmação de Ravi Kalakota e Márcia Robinson (Kalakota e Robinson 2001), especialistas em estratégias de negócios: "Os portais de maior sucesso estão acumulando grande quantidade de informações sobre produtos que estão disponíveis aos consumidores de uma forma fácil de acessar, com base em suas necessidades".

Existem dois aspectos a se considerar nessa afirmação. Primeiro, o número de produtos e serviços é cada vez maior em sites de Comércio Eletrônico, o que nos leva a um problema conhecido como *sobrecarga de informação*. O segundo aspecto é que os sites de sucesso provêm mecanismos para resolver o problema de sobrecarga de informação como, por exemplo, apresentar informações personalizadas ao perfil de cada cliente. Esse fato nos leva a constatar a importância crescente dos Sistemas de Recomendação. Além disso, podemos ainda enumerar alguns benefícios do uso dessa tecnologia para o Comércio Eletrônico (Schafer et al. 2001):

- **Converter visitantes em compradores:** usuários de um *website* freqüentemente pesquisam produtos ou serviços sem comprar qualquer coisa. Os Sistemas de Recomendação poderiam ajudar os usuários a encontrar de forma mais fácil os produtos que eles teriam maior chance de comprar.
- **Potencializar venda cruzada:** a venda cruzada é aquela em que os produtos adquiridos estão de alguma forma relacionados. Os Sistemas de Recomendação podem aumentar a chance de vendas cruzadas em um *website*, a partir da sugestão de produtos relacionados aos produtos que foram em

um momento anterior diagnosticados como interessantes para o usuário.

- **Diminuição de custos através da fidelização do cliente:** estudos mostram que uma empresa típica perde a metade dos seus clientes a cada cinco anos. Além disso, custa de cinco a dez vezes mais obter um novo cliente do que manter um existente. Diante desse fato, é imperativa uma estratégia de negócios que vislumbre a fidelização. Os Sistemas de Recomendação podem auxiliar tanto na apresentação de informações que sejam pertinentes para a necessidade do cliente, quanto na forma com que essas informações são apresentadas. Aliado a isso, há a tendência de que os consumidores freqüentem os sites que melhor preencham suas necessidades. Logo, uma boa experiência do consumidor com relação ao serviço provido aumenta a chance de que esse simples usuário torne-se um assíduo cliente e, até mesmo, venha a recomendá-lo a algum amigo.

As tecnologias núcleo dos Sistemas de Recomendação são comumente chamadas de Engenhos de Personalização. Em geral, estas tecnologias utilizam técnicas de Inteligência Artificial para filtrar as informações relevantes aos usuários dos mais diversos tipos de Sistemas de Informação. Duas técnicas de filtragem de informação que têm se popularizado nos últimos anos são a Filtragem Colaborativa e a Filtragem Baseada em Conteúdo (subseção 2.3.2). Embora tenham mostrado bom desempenho, ambas possuem problemas como aqueles associados ao mau uso dos recursos computacionais, que acabam comprometendo a velocidade de resposta e a alocação de memória. Aspectos como esses são críticos em sistemas de informações que operam em ambientes com grande número de acessos, como é o caso da Internet.

Surge daí a principal motivação deste trabalho: a possibilidade de melhorar a qualidade dos Sistemas de Recomendação e, conseqüentemente, a satisfação dos usuários e das empresas que deles se beneficiam. Para isso, é necessário o desenvolvimento de um método de filtragem que supere as barreiras e/ou problemas (subseção 2.3.3) das abordagens já existentes.

1.3. OBJETIVOS

Dois são os objetivo gerais deste trabalho. O primeiro é a concepção e o desenvolvimento de um método alternativo de Filtragem de Informação Baseado em Conteúdo. Este método deve utilizar as ferramentas disponíveis no domínio de Análise de Dados Simbólicos, que se trata de um novo domínio em Descoberta do Conhecimento. O segundo objetivo geral deste trabalho é avaliar o método proposto de forma a compará-lo estatisticamente com uma outra abordagem já consolidada.

A fim de alcançar os objetivos gerais mencionados anteriormente, são traçados alguns objetivos específicos ou secundários. Assim, o primeiro objetivo secundário é o estudo sistematizado do problema de recomendação em sistemas de informação, o qual é apresentado no próximo capítulo. A partir deste estudo, é possível o agrupamento de diversos trabalhos relacionados ao tema e a exploração dos problemas não superados pelas abordagens atuais.

O segundo objetivo secundário é o estudo de um domínio específico de recomendação. Assim, foi escolhido neste trabalho o domínio de recomendação de filmes para aplicação do método proposto. Este objetivo secundário é pré-requisito para o terceiro objetivo específico: o desenvolvimento de um ambiente de avaliação experimental. Com esse ambiente criado e com uma metodologia de avaliação definida, que compreende o quarto objetivo específico, podemos cumprir o segundo objetivo geral do trabalho que é a avaliação do método proposto.

1.4. ORGANIZAÇÃO

A organização deste trabalho é descrita como segue:

Capítulo 2: Sistemas de Recomendação	São abordados vários aspectos importantes dos Sistemas de Recomendação como, por exemplo, os tipos existentes e suas aplicações. Além disso, há um detalhamento maior do Sistema de Recomendação Personalizado, em que são descritas as etapas principais no processo de recomendação, os problemas inerentes, e que tipo de técnicas de Inteligência Artificial podem ser utilizadas para resolvê-los.

Capítulo 3: Trabalhos Relacionados	O estado da arte ou os trabalhos relacionados são discutidos neste capítulo. Para efeito de organização ele é dividido em três seções. A primeira refere-se ao problema de coletar o perfil do usuário. Após isso são apresentados alguns estudos relacionados aos algoritmos de predição e geração de listas de recomendação. Por fim, são descritos alguns critérios comumente usados na análise e comparação da qualidade dos algoritmos.
Capítulo 4: Filtragem De Informação Baseada Em Dados Simbólicos Modais	Este capítulo descreve uma das principais contribuições deste trabalho. Trata-se de um novo método de filtragem de informação baseado no domínio de Análise de Dados Simbólicos, a ser introduzido logo na primeira seção. A segunda seção do capítulo explica com exemplos como se dá a representação e a construção do perfil do usuário. Seguido a isso é apresentado o algoritmo de recomendação. A fim de exemplificar o método, a seção 4 ilustra uma aplicação real no domínio de recomendação de filmes. Por fim, são descritas as conclusões do capítulo.
Capítulo 5: Análise Experimental	Neste capítulo é apresentada a análise experimental realizada no modelo proposto no capítulo 4. A idéia básica é descrita na seção inicial. Seguido a isso é introduzido o método de filtragem k Vizinhos Mais Próximos (kNN) que servirá de referência para comparação com a abordagem desenvolvida neste trabalho. O ambiente e a metodologia são descritos nas seções subseqüentes. Por fim, são apresentados os resultados dos experimentos e realizadas algumas análises a partir desses resultados.
Capítulo 6: Conclusões	Este capítulo descreve as conclusões finais do trabalho apresentado e enumera alguns trabalhos futuros.

Capítulo 2

SISTEMAS DE RECOMENDAÇÃO

2.1. INTRODUÇÃO

O capítulo anterior descreveu brevemente algumas características dos Sistemas de Recomendação bem como os impactos que essas tecnologias propiciam em ambientes de Comércio Eletrônico. Tais impactos serviram de alguma forma como motivação para o desenvolvimento deste trabalho.

Conforme descrito na seção 1.3, nosso objetivo maior é o desenvolvimento de um método de filtragem de informação que supere alguns dos problemas das abordagens atuais. As técnicas de filtragem de informação, por sua vez, são utilizadas em Sistemas de Recomendação como o núcleo das soluções para o problema de sobrecarga de informação. Dessa forma, antes de estudarmos os problemas das técnicas atuais de filtragem de informação, é fundamental a familiarização com alguns aspectos dos Sistemas de Recomendação.

Nesse sentido, o objetivo deste capítulo é introduzir os tipos de Sistemas de Recomendação, os níveis de personalização e os problemas existentes nesse contexto.

2.2. TAREFAS DOS SISTEMAS DE RECOMENDAÇÃO

Em geral, os Sistemas de Recomendação podem ser vistos como tipos especiais de Sistemas de Informação que fornecem uma visão personalizada do *repositório digital* [3] associado a esse sistema. Essa visão personalizada da informação pode ser apresentada de diversas formas, a depender do objetivo ou tarefa que o Sistema de Recomendação se destina.

A seguir são descritas as principais tarefas dos Sistemas de Recomendação (Herlocker et al. 1999, Schafer et al. 2001, Teixeira 2002).

[3] A explosão de documentos eletrônicos nos diversos tipos de rede fez surgir o conceito de repositórios de informação digital. Basicamente, esses repositórios consistem em agrupamentos lógicos de documentos eletrônicos com precária organização estrutural da informação. No contexto da Internet pode-se dizer que grandes portais são exemplos claros de repositórios de informação digital. Da mesma forma, as redes internas de corporações de médio e grande porte constituem repositórios de informação digital, visto que nesses ambientes surge uma quantidade enorme de documentos eletrônicos.

i) Apresentar os pontos de vista de usuários do sistema a cerca de uma informação, produto ou vendedor.

Websites como Amazon, CDNow, Citeseer, ACM Journals, Arremate e Infobox permitem que os usuários expressem suas opiniões sobre um determinado produto, vendedor ou artigo (Figura 1). Isso pode ajudar um novo usuário a filtrar as informações que são mais pertinentes em um dado contexto, ou a escolher um produto ou artigo em uma quantidade de opções significativas, ou mesmo a aumentar a sua credibilidade em um vendedor.

Alguns websites, como Amazon e CDNow, permitem ainda que o usuário indique a relevância de um determinado comentário, estabelecendo até mesmo uma seqüência ordenada dos comentários. Observe que tudo isso ocorre através da própria interação da comunidade de usuários com o sistema.

Figura 1 – Tela sobre a reputação de um vendedor definida pela própria comunidade do website Arremate.com.

ii) *Apresentar opiniões de críticos conceituados.*

Isso pode influenciar pessoas a comprarem um determinado livro se indicado por um crítico com opinião convergente àquela do usuário. Além disso, pode direcionar a busca de artigos em um repositório como o da ACM, na medida que o usuário conhece o conjunto de atores pertinentes em uma determinada área do conhecimento e passa a procurar artigos indicados por estes (Figura 2).

Nesse caso, assim como no item anterior, temos um processo puramente manual, em que alguns usuários (os críticos) fornecem suas opiniões a cerca de itens contidos no repositório digital.

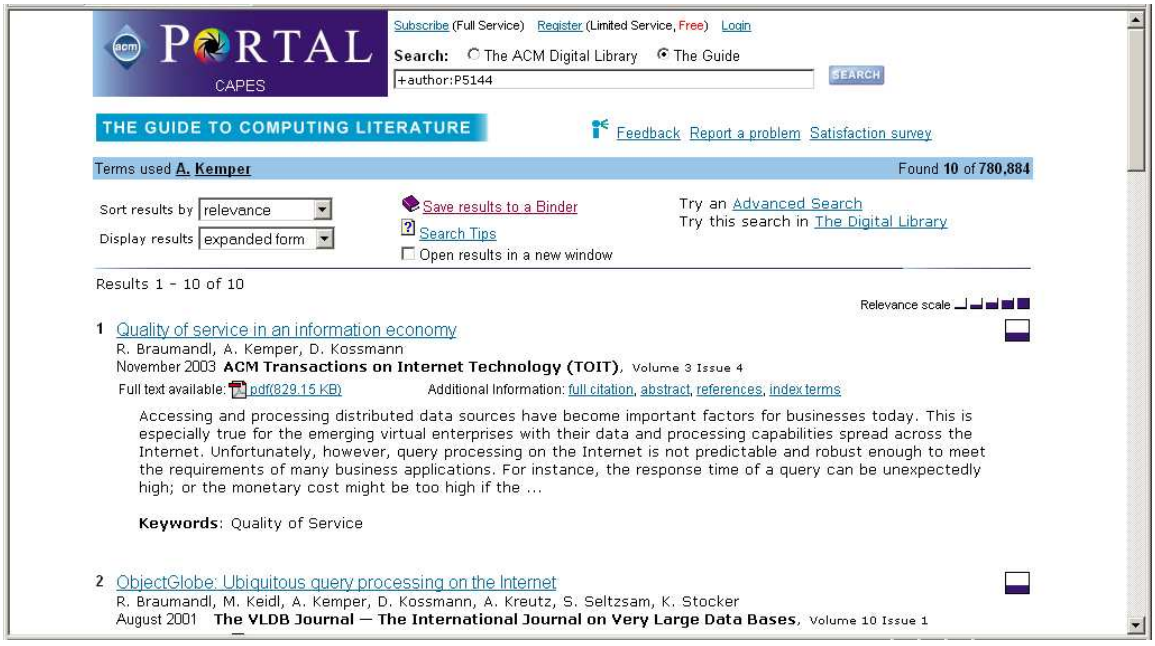


Figura 2 – Resultado da consulta por artigos relacionados a um determinado pesquisador na biblioteca digital da ACM através do portal de periódicos da CAPES.

iii) Listar os itens de informação ou os produtos mais consultados e/ou comprados.

Esse é um dos recursos mais utilizados em Sistemas de Recomendação, basicamente, por dois motivos: não requer conhecimento prévio sobre o usuário que está recebendo a informação e não exige que a comunidade de usuários crie o hábito de fornecer suas opiniões referentes a produtos e/ou outras informações do repositório digital.

Esse tipo de sugestão é gerada automaticamente pelo sistema a partir dos dados de compra dos usuários ou dos dados de acesso às páginas do site. No entanto, não requer qualquer processamento de inteligência computacional.

A Figura 3 ilustra alguns tipos de sugestões oferecidas ao usuário em uma página de DVDs no estilo MPB. Nela são apresentados os itens mais vendidos, os lançamentos e outras recomendações geradas a partir de informações de acesso da comunidade de usuários no website.

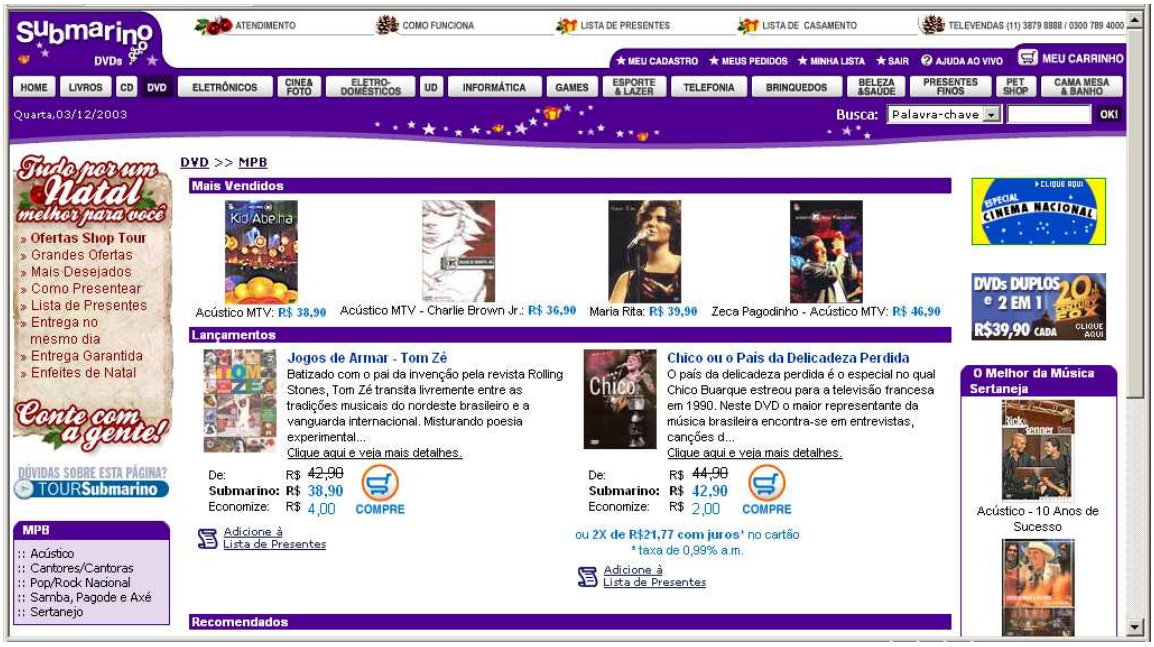


Figura 3 – Tela da loja virtual Submarino.com na categoria DVD/MPB, em que o usuário observa os itens mais vendidos, os lançamentos e outros DVDs recomendados.

iv) Listar os produtos ou itens que tenham um nível de semelhança significativo com o contexto visualizado pelo usuário.

O Citeseer (Figura 4) disponibiliza para o usuário uma lista dos artigos considerados semelhantes àquele que está sendo consultado em um dado momento. Além disso, ele provê as listas dos artigos referenciados e dos artigos que o referenciam.

Figura 4 – Tela resultante da consulta de documentos semelhantes ao texto de um artigo de referência no Citeseer.

Websites de Comércio Eletrônico como a Amazon, CDNow e Americanas fornecem também ao usuário uma lista dos produtos relacionados ao que está sendo consultado em dado momento ou aos produtos que estão inclusos no *carrinho de compras*[4] (Figura 5). Isso pode potencializar o que profissionais de *marketing* costumam chamar de *venda cruzada*.

[4] O carrinho de compras é uma abstração virtual do carrinho de supermercado. Através desse mecanismo o usuário pode indicar seu interesse por produtos em uma loja on-line mesmo que a compra não seja efetuada. Com isso, os Sistemas de Recomendação tem maiores chances de melhorar a qualidade de suas sugestões através do uso das informações que refletem os interesses dos usuários.

Figura 5 – Tela de recomendação de itens na loja virtual Americanas.com tomando como base o conteúdo do carrinho de compras.

Para esse tipo de recomendação, é necessário, em algumas situações, o uso de algoritmos mais complexos de inteligência computacional, sobretudo os de Recuperação de Informações.

v) Apresentar produtos ou itens que estejam associados com o contexto visualizado pelo usuário de acordo com o comportamento passado da comunidade.

São inúmeros os exemplos de websites que fornecem esse tipo de personalização. A Amazon (Figura 6), por exemplo, mostra uma lista dos produtos mais comprados por consumidores que também compraram o produto visualizado pelo usuário em um dado momento.

Basicamente, o sistema efetua um cruzamento das informações de compras e/ou navegação dos consumidores com o item ou o produto visualizado pelo usuário.

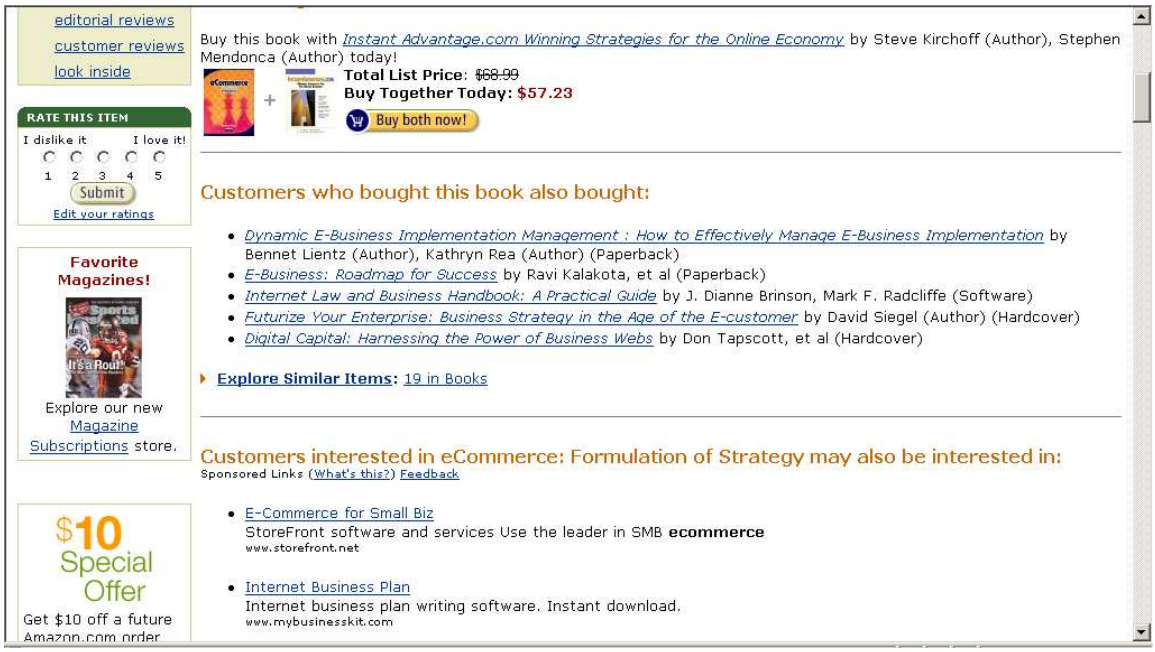


Figura 6 – Tela resultante da consulta de um livro no website Amazon.com.

vi) Apresentar itens ou produtos mais próximos aos interesses do usuário de acordo com seu perfil.

Esse é o mais alto nível de personalização que se pode ter. É também o mecanismo que viabiliza em sua totalidade a estratégia de marketing *one-to-one*, tendo em vista que através desse mecanismo o conteúdo fornecido ao usuário é totalmente customizado às suas necessidades.

A Americanas (Figura 7), por exemplo, envia mensagens de e-mail que contêm sugestões de produtos que mais se adeqüem aos interesses do usuário.

Figura 7 – Tela de um e-mail enviado pela Americanas.com.

É natural que nesse tipo de tecnologia ocorram os mais complexos problemas conhecidos em Sistemas de Recomendação, o que os faz alvo das principais pesquisas na área.

vii) Prover uma nota que reflita a relevância de um determinado item ou produto para o usuário.

A Figura 8 ilustra a página de recomendações gerada automaticamente pelo IMDB para um usuário. Para gerar esta página o IMDB não só levou em conta o perfil deste usuário como também a semelhança com o item "Godfather Trilogy: 1901-1980, The (1992) (V)". Para isso, o processo de recomendação compreende duas etapas principais. Na primeira, o sistema consulta na base de conteúdo todos os itens semelhantes a um determinado item indicado pelo usuário. No nosso exemplo, o sistema procura na base de dados todos os filmes semelhantes ao item "Godfather Trilogy: 1901-1980, The (1992) (V)". Na segunda etapa o sistema prediz uma nota para cada um dos itens recuperados na primeira etapa levando em conta o perfil do usuário. Por exemplo, verifica-

se que o filme “Cidade de Deus” possui uma nota 8,6, representando que há uma alta probabilidade de que o usuário goste deste título.

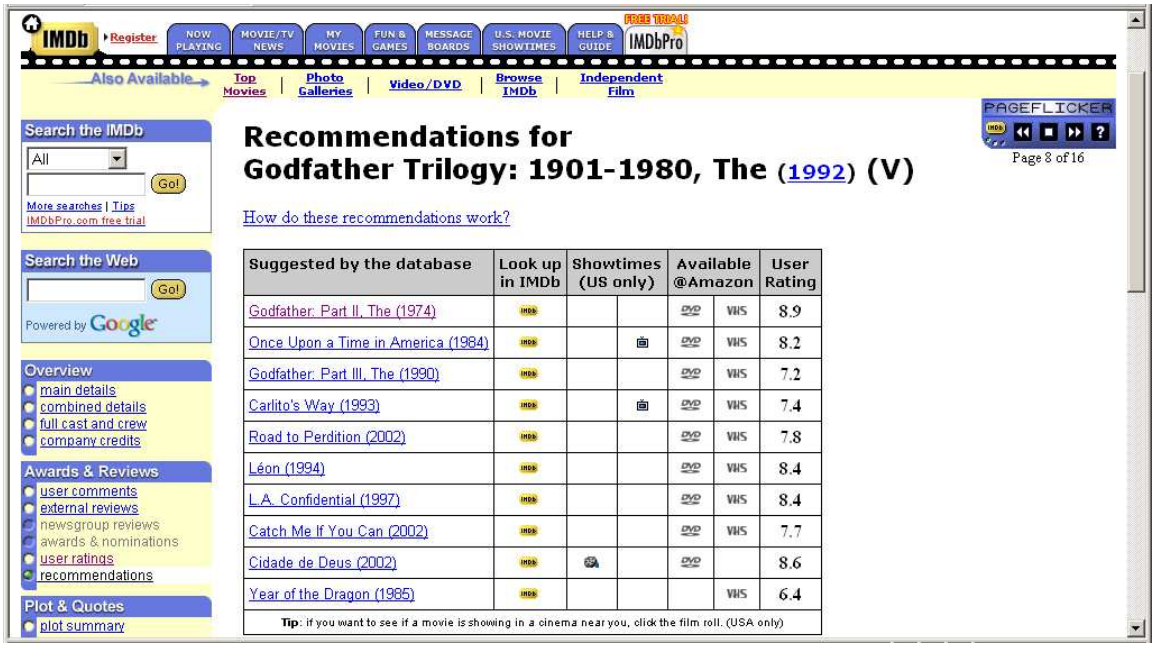


Figura 8 – Tela de recomendações do IMDB para um determinado usuário.

Portanto, essa abordagem baseia-se no perfil do usuário assim como o método anterior. O FilmConceil e o IMDB, por exemplo, possibilitam que o usuário forneça informações sobre seus interesses e posteriormente obtenham predições de filmes.

2.3. SISTEMAS DE RECOMENDAÇÃO PERSONALIZADOS

Os Sistemas de Recomendação capazes de identificar e aprender as preferências e necessidades de um usuário, gerando recomendações customizadas ao seu perfil, assim como aquelas retratadas nos itens (vi) e (vii) da seção anterior, são chamados **Sistemas de Recomendação Personalizados** (Herlocker et al. 1999, Schafer et al. 2001, Teixeira 2002).

Sistemas de Recomendação Personalizados identificam cada usuário de forma individual e coletam suas preferências que são armazenadas em uma representação interna, chamada de *perfil do*

usuário. O sistema utiliza tradicionalmente técnicas de *Filtragem de Informação* para gerar recomendações apropriadas ao interesse de cada usuário a partir da representação do seu perfil. Além disso, técnicas não inseridas no domínio de *Filtragem de Informação*, como *Regras de Associação* e *Redes Neurais*, podem ser utilizadas nesses sistemas (Agrawal et al. 1993, Good et al. 1999, Schafer et al. 1999).

A Figura 9 ilustra o funcionamento de um Sistema de Recomendação Personalizado. O sistema, representado na ilustração pelo *filtro*, separa os itens irrelevantes dos itens interessantes. Estes últimos são recomendados ao usuário através de uma das formas descritas anteriormente. O usuário, por sua vez, recebe essa informação customizada e pode eventualmente comportar-se de forma que o sistema adquira informações sobre o seu real interesse nos itens apresentados. Esse processo é comumente chamado de *feedback* do usuário. A forma com que o sistema obtém o *feedback* do usuário depende de cada tipo de domínio.

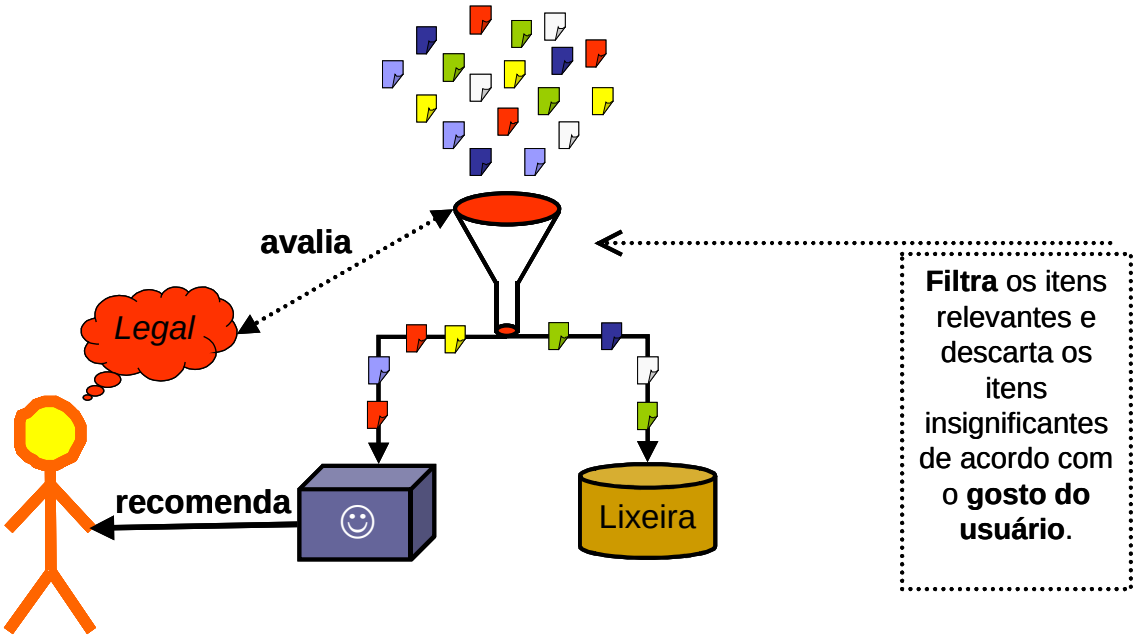


Figura 9 - Esquema de um Sistema de Recomendação Personalizado.

2.3.1 Etapas nos Sistemas de Recomendação Personalizados

Uma análise da Figura 9 nos leva a constatar que um Sistema de Recomendação Personalizado possui duas etapas bem definidas a fim de possibilitar o seu objetivo principal: recomendar ou predizer itens para um usuário de acordo com o comportamento passado do usuário. São elas:

a) Como coletar o perfil do usuário?

b) Como explorar adequadamente esse perfil?

A Coleta do Perfil do Usuário

Uma das formas de representar o comportamento do usuário é através do que chamamos de *perfil do usuário*. Este pode ser coletado algumas vezes *implicitamente* e outras *explicitamente* (seção 3.2). Uma aquisição implícita poderia ser pensada em um cenário de um website de Comércio Eletrônico. Nesse caso a seguinte política poderia ser adotada: cada vez que um usuário compre algum produto o sistema registraria essa informação no perfil do usuário de forma positiva. Essa estratégia de aquisição implícita do perfil do usuário é adotada por vários sites de Comércio Eletrônico.

No entanto, algumas vezes é mais conveniente a aquisição explícita do perfil do usuário. Considere, por exemplo, um site de recomendação de filmes. Assumindo que tal sistema não comercializa filmes, não seria adequada a estratégia de aquisição implícita citada anteriormente. Sendo assim, o sistema necessita uma resposta do usuário com relação aos filmes apresentados, dando origem a um perfil de usuário constituído dos filmes avaliados e suas respectivas notas atribuídas por esse usuário. Esse modelo explícito de aquisição de perfil é utilizado, por exemplo, nos websites PTV (http://www.ptv.ie) e Canal Plus (http://www.canalplus.fr).

Gerando Recomendações a partir do Perfil do Usuário

Como utilizar adequadamente o perfil adquirido para efetuar boas recomendações, é um problema bastante estudado nos últimos anos. Alguns trabalhos foram realizados nesse sentido utilizando técnicas de aprendizagem como *Redes Neurais*, *Regras de Associação* e *Árvores de*

Decisão (veja a seção 3.3). No entanto, a maior parte dos estudos concentra-se em técnicas de *Filtragem de Informação*. Embora as soluções propostas sejam diferentes elas podem geralmente ser classificadas em duas categorias principais com relação ao tipo de filtragem adotada. Algumas dessas soluções baseiam-se na análise da correlação entre o conteúdo dos itens com o perfil do usuário para recomendar itens relevantes e descartar os itens não pertinentes. Essa técnica é chamada de *Filtragem Baseada em Conteúdo*. Uma outra abordagem, chamada de *Filtragem Colaborativa*, é baseada na correlação entre perfis de usuários. A idéia básica é selecionar os itens preferidos pelos usuários cujas preferências mais se assemelham ao gosto do usuário alvo. Na próxima seção veremos mais detalhes dessas técnicas de filtragem.

2.3.2 Abordagens em Filtragem de Informação

Como foi exposto anteriormente, os Sistemas de Recomendação Personalizados tradicionalmente se encaixam em duas categorias de filtragem: Colaborativa e Baseada em Conteúdo (Basu et al. 1998). Diversos sistemas de filtragem de informação foram construídos utilizando uma dessas técnicas de filtragem ou mesmo uma combinação de ambas (Resnick et al. 1994, Arya 1995, Shardanand et al. 1995, Krukwich e Burkey 1996, Balanovic e Shoham 1997, Breese et al. 1998, Sarwar et al. 1998, Herlock et al. 1999, Cotter e Smyth 2000, Bezerra e De Carvalho 2004). As características dessas técnicas de filtragem bem como suas vantagens e desvantagens são discutidas com mais detalhes a seguir.

Filtragem de Informação Colaborativa

Nos sistemas de *Filtragem Colaborativos* há a formação de uma comunidade de usuários que interagem com o sistema fornecendo avaliações. Dessa forma, há uma troca de experiências entre os membros da comunidade através das avaliações de cada indivíduo na base de itens do sistema. Assim todos podem usufruir o conhecimento alheio.

Nos Sistemas Colaborativos a recomendação é baseada na similaridade entre usuários, ou seja, são recomendados itens cujos usuários com preferências semelhantes tenham gostado.

Uma das vantagens da Filtragem Colaborativa é o simples fato de não precisar representar o conteúdo dos itens. Toda a avaliação usada na recomendação é baseada no julgamento humano. Significa que se pode levar em consideração características realmente relevantes para a avaliação de qualidade, inclusive as que não são representáveis facilmente em computadores. Adicionalmente, na Filtragem Colaborativa a qualidade das recomendações não é dependente apenas das avaliações do próprio usuário, mas também dos membros da comunidade, o que trás dois benefícios: i) permite ter recomendações de qualidade mesmo com poucas avaliações e ii) torna as recomendações mais diversificadas.

Naturalmente, a Filtragem Colaborativa possui alguns problemas próprios. Um desses problemas está relacionado a novos itens dentro da base de informação. Do momento em que um novo item é adicionado ao sistema até o momento de sua primeira recomendação pode levar um bom tempo, visto que o mesmo precisa ser bem avaliado por um número significativo de membros da comunidade. Isso se dá porque na Filtragem Colaborativa o sistema depende da experiência dos membros da comunidade para determinar a relevância de um item. Em casos em que a base de informação cresce ou muda muito rapidamente ou é muito maior do que o número de usuários, a qualidade das recomendações de um sistema baseado em Filtragem Colaborativa pode ser comprometida. Outro problema está relacionado aos usuários que possuem perfil diferente dos demais membros da comunidade. Para esses usuários não haverá pessoas suficientemente semelhantes a ela no sistema capazes de servir de referência para recomendações.

O sucesso de um sistema de Filtragem Colaborativa é dependente da comunidade de usuários e da dimensão da base de informações que se deseja filtrar. Quanto maior a comunidade e a quantidade de avaliações que ela fornece melhor será a qualidade das recomendações. Adicionalmente, se a base de informações for pequena comparada à comunidade e crescer lentamente, um sistema colaborativo será capaz de gerar boas recomendações. Entretanto se ela for muito maior do que a comunidade ou crescer a uma velocidade muito grande, é provável que um sistema de recomendações puramente colaborativo gere recomendações imprecisas.

Filtragem de Informação Baseada em Conteúdo

Como vimos anteriormente, os sistemas de Filtragem Colaborativos não são aplicáveis em todas as situações. Veremos a seguir que isso serviu de motivação para os sistemas de Filtragem Baseados em Conteúdo, tendo em vista que algumas de suas maiores vantagens são aspectos associados aos problemas em sistemas de Filtragem Colaborativos.

Nos sistemas de Filtragem de Informação Baseado em Conteúdo apenas as preferências do próprio usuário são utilizadas na filtragem. A aquisição das preferências do usuário nesse tipo de filtragem depende sobretudo da descrição dos itens que ele avalia. O ideal é que a partir do perfil coletado, seja possível a recomendação de itens mais similares aos itens bem avaliados pelo usuário e ao mesmo tempo mais dissimilares dos itens mal avaliados. Para isso, a preferência do usuário freqüentemente é usada para construir um perfil contendo indicadores do interesse do usuário sobre determinados tópicos, geralmente representados através de um conjunto de palavras-chaves e pesos associados à relevância do item.

Os sistemas que utilizam apenas Filtragem Baseada em Conteúdo possuem algumas desvantagens relacionadas a esse tipo de filtragem. Uma das principais desvantagens está nas possibilidades limitadas de representação do conteúdo dos itens. Em alguns casos os tipos de informação filtrados não podem ser representados de forma satisfatória usando apenas variáveis de escala quantitativa ou qualitativa. Por exemplo, objetos multimídia como sons, vídeos e figuras são de difícil extração de características. Adicionalmente, algumas dimensões são impossíveis de se representar, como por exemplo, a trilha sonora ou a fotografia no domínio de filmes.

Outro problema relacionado a sistemas de Filtragem Baseada em Conteúdo está na super especialização do sistema em tópicos freqüentes no perfil do usuário. Dessa forma, ao aprender o perfil do usuário o sistema não inova em suas recomendações tendo em vista a pouca diversidade de itens avaliados pelo usuário.

Um último aspecto associado à qualidade das recomendações dos sistemas de filtragem baseada em conteúdo é a forte dependência em

relação à quantidade de avaliações realizadas pelo usuário. Uma vez que a única coisa que influencia a qualidade das recomendações são as avaliações do próprio usuário, será determinante o número de avaliações para aprender sobre a preferência do usuário.

No entanto, a quantidade de itens avaliados pelo usuário também pode aumentar o tempo de classificação e a memória usada pelo sistema, sobretudo em sistemas que utilizam a filtragem baseada em conteúdo. Dessa forma, é uma preocupação cada vez maior em Sistemas de Recomendação Personalizados manter um equilíbrio entre a qualidade de suas recomendações, o tempo de resposta e o espaço de memória utilizado.

2.3.3 Problemas nos Sistemas de Recomendação Personalizados

Alguns dos problemas dos Sistemas de Recomendação Personalizados provêm da própria estrutura em que eles são baseados. Assim, relatamos a seguir os problemas decorrentes da etapa de aquisição e construção do perfil do usuário. Após isso, podemos enumerar os problemas inerentes aos métodos de filtragem discutidos na seção anterior.

Adicionalmente identificamos aqueles problemas que são considerados verdadeiros desafios e que em boa parte são poucos os estudos. São eles: esparsidade, escalabilidade, desempenho em tempo real, utilização de memória, recomendações em grupo, dimensão temporal e espacial, explicação e apresentação das recomendações.

Além dos problemas intrínsecos à própria estrutura dos Sistemas de Recomendação Personalizados, existem os problemas decorrentes do domínio. Dessa forma, podemos verificar no domínio das aplicações de Comércio Eletrônico os seguintes problemas: incorporação de informações do negócio e integração com outros sistemas.

Problemas na Aquisição do Perfil do Usuário

Sabemos que nem sempre é possível adquirir informações suficientes sobre as preferências do usuário. Muitas vezes isso ocorre por que não é cômodo para os usuários fornecer notas sobre itens ou produtos

existentes no repositório digital. Assim, alguns estudos propõem métodos de *aprendizagem ativa* (Cohn et al. 1994, Hasenjäger 2000, Engelbrecht e Brits 2002, Teixeira et al. 2002, Teixeira 2002), que tentam selecionar o mínimo de itens que sejam o mais informativos possíveis a fim de que o usuário possa avaliá-los e que o sistema consiga construir um perfil adequado para o usuário.

De fato, como dito anteriormente, há situações em que é impraticável a aquisição explícita das preferências dos usuários. Em aplicações de Comércio Eletrônico isso é ainda mais visível. Alguns estudos mostram que esse problema pode ser minimizado com o mecanismo de avaliações implícitas (seção 3.2).

Problemas Inerentes aos Métodos de Filtragem

Como visto na seção 2.3.2, os principais problemas da Filtragem Colaborativa são:

- Latência na recomendação de novos itens;
- Qualidade insatisfatória na recomendação para usuários *ovelha-negra* [5];
- Qualidade dependente da participação de um número significativo de usuários.

Já os sistemas de Filtragem por Conteúdo possuem os seguintes problemas:

- Limitações na representatividade do perfil do usuário;
- Super especialização das preferências;
- Qualidade dependente do número de itens que constituem o perfil do usuário.

Problema de Esparsidade

Tradicionalmente, a esparsidade é um problema intrínseco ao processo de recomendação, tendo em vista que o propósito de um Sistema

[5] Ovelhas-negras são aqueles usuários que são considerados naturalmente diferentes. Ou seja, são pouquíssimos ou inexistentes os usuários próximos de suas preferências.

de Recomendação é minimizar o problema de sobrecarga de informação (Sarwar et al. 2000, Schafer et al. 2001).

Para exemplificar o problema de esparsidade, considere o cenário de uma loja virtual de livros. Nesse contexto, é natural que a maior parte dos usuários tenham consumido (lido, consultado ou comprado) no máximo trinta livros. Ao mesmo tempo, os usuários de livrarias on-line solicitam recomendações em uma base que pode alcançar facilmente milhares de livros. Logo, temos uma situação em que há muitos itens no repositório e poucas avaliações dos usuários, elevando a probabilidade de itens jamais avaliados no repositório.

O problema de esparsidade é bastante claro no exemplo anterior. A base de itens a serem filtrados é muito grande, impossibilitando que um usuário conheça (avalie) boa parte deles. Assim, boa parte dos itens contém pouca ou nenhuma avaliação. Dessa forma, a matriz de avaliações dos usuários torna-se esparsa. Como vimos anteriormente, esse problema pode ser minimizado pela filtragem baseada em conteúdo, visto que mesmo itens pouco avaliados, ou sequer conhecidos por qualquer usuário, podem ser recomendados. Para isso, basta que tais itens tenham conteúdo semelhante ao perfil de um usuário.

Note que a ocorrência de uma grande quantidade de usuários da comunidade não constitui um problema de esparsidade, embora constitua um problema de escalabilidade. A explicação para isso é que quanto mais usuários avaliando os itens do repositório melhores poderão ser as recomendações através da filtragem colaborativa, visto que haverá mais opções de usuários para se assemelhar a um usuário alvo. Todavia, o problema de escalabilidade surge em vista da maior dificuldade em se encontrar aqueles usuários mais próximos entre tantas opções.

O problema de esparsidade é cada vez mais comum em aplicações de Comércio Eletrônico (Demirez 2003) devido à diversidade de produtos oferecidos pelas lojas virtuais aos seus consumidores.

Problemas de Escalabilidade, Desempenho e Uso de Memória

À medida que o número de itens a serem filtrados cresce, aumentam também as preocupações com escalabilidade, desempenho e uso de memória.

A escalabilidade está associada tanto ao tamanho do conjunto de itens a serem filtrados, quanto à quantidade de usuários que acessam o sistema em busca de sugestões. Algumas técnicas como redução da dimensionalidade (Sarwar et al. 2000) e paralelismo (Olsson 2003) podem ser adotadas para minimizar esse problema.

O desempenho está associado ao tempo de resposta do sistema para gerar uma lista de sugestões ao usuário ou predizer a nota de alguns itens. Normalmente, esse tempo de resposta é diretamente proporcional a quantidade de itens avaliados pelo usuário, sobretudo na filtragem baseada em conteúdo. Adicionalmente, o desempenho também é influenciado pelo número de usuários na comunidade, por dois motivos: o sistema precisa atender em um sistema real à demanda de vários usuários ao mesmo tempo; e, em se tratando de filtragem colaborativa, o sistema requererá mais tempo para identificar os usuários com preferências similares ao usuário alvo.

Finalmente, podemos constatar que quanto maior o número de itens avaliados pelo usuário maior será a memória necessária para armazenar o seu perfil.

Alguns estudos conseguiram diminuir os problemas de desempenho e uso de memória com mudanças na representação interna do perfil do usuário ou com uma pré-seleção dos itens que constituiriam o perfil do usuário (seção 3.2).

Em grandes websites de Comércio Eletrônico, o desempenho e a memória utilizada são requisitos cruciais, tendo em vista a diversidade de produtos e a enorme quantidade de usuários que o acessam.

Recomendações em Grupo

É irreal pensar que as pessoas vivem totalmente isoladas. Se considerarmos um ambiente familiar, podemos verificar que boa parte dos

produtos adquirida é em função do bem-estar de todos os membros da família. Uma outra situação de colaboração mútua é a escolha de um programa de entretenimento entre um grupo de amigos.

A incorporação dessa característica no contexto das aplicações de Comércio Eletrônico atuais é quase que inexistente. Pouquíssimos estudos foram efetuados até o momento no que se refere a recomendações para grupos. Conseqüentemente, há uma quantidade significativa de problemas a serem pensados e solucionados (Queiroz et al. 2002, Queiroz 2003, Queiroz e De Carvalho 2003, Queiroz e De Carvalho 2004).

Ausência ou Descaso das Dimensões Temporal e Espacial

É fato que as preferências do usuário mudam com o tempo e com a sua localização geográfica. Consideremos por exemplo, a compra de roupas durante o período de um ano. Seria tolo recomendar a um cliente, que reside no sul do Brasil, a compra de camisetas nos meses de maio a setembro, tendo em vista o rigoroso inverno que ocorre nesse período do ano naquela localidade. Entretanto, caso o cliente resida em algumas cidades do nordeste do Brasil a recomendação poderia ser adequada.

Os Sistemas de Recomendação Personalizados atualmente não tratam esses atributos com a atenção que os sistemas de Comércio Eletrônico normalmente exigiriam (Schafer et al. 2001, Tang et al. 2003). A conseqüência imediata disso é o pós-processamento (implementado caso a caso) das sugestões geradas para um usuário a partir do seu perfil, a fim de considerar as dimensões temporal e espacial.

Dificuldade de Explicação e Apresentação das Recomendações

Segundo Schafer et al. (1999), o sucesso de um Sistema de Recomendação deveria ser medido através de sua eficácia em ajudar os usuários a tomar decisões que, cedo ou tarde, eles considerariam acertadas. Uma das formas de alcançar um nível de eficácia significativo seria prover argumentos que convençam o usuário das sugestões que lhe foram oferecidas em um dado momento. Em um website de Comércio Eletrônico, isso aumentaria a credibilidade, tendo em vista que o usuário deduziria que os produtos recomendados não visaram unicamente a venda e conseqüentemente o lucro, mas também a satisfação do consumidor.

Ainda são escassos os estudos que provêm métodos de explicação para os algoritmos de recomendação.

Ausência de Informações do Negócio e Integração com outros Sistemas

Freqüentemente, a equipe de vendas e/ou marketing de uma empresa deseja direcionar o consumo dos seus clientes para um determinado conjunto de produtos. Por exemplo, é comum que um supermercado faça promoções de produtos com prazo de validade próximo de vencer. Lojas de departamentos costumam fazer promoções de produtos com alto nível de estoque. Dessa forma, seria conveniente a incorporação de características do próprio negócio da empresa a fim de permitir soluções com alto valor agregado (Schafer et al. 2001).

Pela mesma razão, é interessante que os sistemas de recomendação sejam integrados com os sistemas de suporte à decisão utilizados pela equipe de marketing ou vendas. Atualmente, não se tem conhecimento de tecnologias de recomendação que gerem relatórios de marketing para apoio a tomada de decisão. Em geral tais relatórios fornecem uma visão segmentada da comunidade de consumidores. Assim, uma das formas de unir as tecnologias de recomendação com as tecnologias de suporte a decisão seria aproveitar a correlação entre os usuários, identificada a partir de algoritmos de filtragem colaborativa, para confecção de relatórios segmentados.

2.4. CONSIDERAÇÕES FINAIS

Neste capítulo foram apresentados diversos aspectos associados ao contexto de Sistemas de Recomendação. Assim, foram descritas e exemplificadas na seção 2.2 as principais tarefas realizadas por esses tipos de sistemas e na seção 2.3 foi descrito mais detalhadamente o Sistema de Recomendação Personalizado. Dessa forma, enumerou-se as principais etapas no processo de recomendação automatizado, bem como os problemas e desafios existentes. Tais aspectos são essenciais para o desenvolvimento do restante deste trabalho, visto que eles permitem a compreensão do domínio de recomendações.

Adicionalmente, foram descritas as principais abordagens em Inteligência Artificial que podem ser utilizadas como o núcleo das soluções em Sistemas de Recomendação para problemas de sobrecarga de informação. Mais especificamente, foram mencionadas as principais características, vantagens e desvantagens da abordagem de Filtragem de Informação que é baseada no conteúdo descritivo dos itens do repositório. Tal descrição a cerca das técnicas de filtragem de informação é fundamental para o restante do trabalho, tendo em vista que nosso objetivo maior é o desenvolvimento de um novo método de filtragem que supere parte dos problemas existentes nas soluções atuais.

No próximo capítulo descreveremos alguns dos principais trabalhos relacionados aos Sistemas de Recomendação, estabelecendo a relação dos mesmos com os diversos aspectos apresentados neste capítulo.

Capítulo 3

TRABALHOS RELACIONADOS

3.1. INTRODUÇÃO

O capítulo anterior descreveu aspectos relevantes a cerca das tecnologias de Sistemas de Recomendação. Isso foi importante para o restante do trabalho, visto que é necessário conhecer o máximo possível de requisitos, restrições e desafios dos Sistemas de Recomendação a fim de se propor soluções em Inteligência Computacional para esses sistemas. Além disso, é fundamental a familiarização com as possíveis soluções em Inteligência Artificial já propostas para esses sistemas. Dessa forma, o capítulo 2 introduziu as principais características das abordagens existentes em Inteligência Artificial para solução de problemas de sobrecarga de informação no contexto de Sistemas de Recomendação Personalizados.

Nesse capítulo são descritas as principais soluções propostas em Inteligência Computacional para alguns dos problemas dos Sistemas de Recomendação citados no capítulo anterior. A importância disso para o restante do trabalho é propiciar ao leitor um aprofundamento sobre os trabalhos já desenvolvidos para minimizar alguns dos problemas existentes em Sistemas de Recomendação. Com isso, estamos não só relatando as abordagens existentes, mas também diagnosticando problemas remanescentes em cada uma das abordagens que podem servir de inspiração para o desenvolvimento deste trabalho.

Para efeito de organização deste capítulo dividimos os trabalhos relacionados nos seguintes grupos:

- estudos referentes à coleta do perfil do usuário e à forma com que se representa internamente esse perfil;
- estudos relacionados aos algoritmos de predição e geração de listas de recomendação;
- critérios de análise e comparação da qualidade dos algoritmos.

3.2. AQUISIÇÃO DAS PREFERÊNCIAS DO USUÁRIO

Existem atualmente duas abordagens para o problema de aquisição das preferências do usuário: a aquisição explícita e a aquisição implícita (Claypool et al. 2001). Na primeira o usuário informa claramente o seu grau de interesse em um determinado item do repositório, através de uma nota, por exemplo. De fato, a maior parte dos Sistemas de Recomendação assume uma aquisição explícita, através de uma nota em uma escala quantitativa. Com essa abordagem, os problemas podem concentrar-se exclusivamente nos algoritmos de recomendação. Nesse sentido, daremos uma maior atenção nessa seção aos trabalhos relacionados à segunda abordagem, onde estão os mais críticos problemas de coleta do perfil do usuário.

Com relação à abordagem implícita, não é difícil enumerar algumas ações do usuário em um website que possam de alguma forma serem usadas para inferir suas preferências. O grande problema, portanto, não está na identificação dos critérios de aquisição implícita, mas sim em medir o nível de relevância de cada critério para formação do perfil do usuário.

Por exemplo, Nichols et al. (1997), Oard e Kim (1998) e Chan (1999) concordaram que as ações de examinar, imprimir, salvar informações ou registrar no *bookmark* uma determinada página de um website refletem os interesses dos usuários na informação ali contida e, conseqüentemente, são potencialmente úteis para se obter melhores recomendações. No entanto, esses trabalhos não analisam de forma direta a correlação entre esses critérios implícitos de preferências do usuário com os seus interesses explícitos.

O tempo de leitura gasto em uma determinada página foi considerado por Chan (1999), Morita e Shinoda (1994) e Konstan et al. (1997) como um importante fator de identificação do nível de interesse do usuário. Além disso, Chan (1999) propôs um tempo de leitura relativo com o tamanho da página de forma a minimizar ruídos originados por diferenças na amostra. Apesar disso, o tempo de leitura continua sendo um critério muito subjetivo tendo em vista que a complexidade de leitura entre os elementos da amostra pode variar de forma significativa, inclusive

de usuário para usuário. Uma forma de minimizar esse problema pode ser a normalização pelo tempo de leitura médio ao nível de cada usuário.

Considerando ainda o critério anterior, Konstan et al. (1997) descreve a partir do sistema *GroupLens* a correlação entre o tempo gasto na leitura de um artigo com a avaliação explícita do mesmo. Nesse trabalho, mostra-se que a qualidade das predições, tomando como base o tempo de leitura para determinados artigos, pode ser tão boa quanto a qualidade das predições baseadas em avaliações explícitas dos mesmos artigos.

Outras formas importantes de se adquirir informações sobre as preferências dos usuários são a partir de ações como: selecionar uma seção do texto, copiar no clipboard uma parte do texto, clicar com o mouse e efetuar a rolagem da tela (*scrolling*). Goecks e Shavlik (2000) modificaram um *browser* para capturar as atividades com o mouse e as ações de rolagem de tela. Apesar de ter contribuído com a definição de novos critérios, Goecks e Shavlik (2000) não os correlacionam efetivamente com o interesse do usuário.

Lieberman (1997) é um dos poucos trabalhos que propõe níveis de interesse de acordo com o tipo de ação efetuada pelo usuário. No entanto, ele não define explicitamente em seu trabalho medidas do nível de interesse para os indicadores implícitos.

Finalmente, Claypool et al. (2001) estuda de forma efetiva e experimental a correlação entre diversos tipos de avaliações implícitas citados anteriormente com os interesses explícitos do usuário, através de um Web *browser* construído especialmente para isso. Como resultado principal, é mostrado que o tempo de leitura gasto para uma página, a intensidade de *scrolling*, ou uma combinação de ambos, possuem uma alta correlação com o interesse explícito do usuário, diferentemente de ações como clique de mouse.

A seguir são descritos alguns trabalhos relacionados à representação interna do perfil do usuário e às técnicas de recomendação propostas.

3.3. METODOLOGIAS DE RECOMENDAÇÃO

Conforme discutido na seção 2.3.2, as abordagens para solução do segundo problema existente em sistemas de recomendação, ou seja, a geração das recomendações, podem ser classificadas em: métodos de filtragem baseada em conteúdo, métodos de filtragem colaborativo, métodos de filtragem híbridos e outros métodos de aprendizagem que não podem ser inseridos em nenhum grupo anterior. A seguir veremos alguns dos principais trabalhos relacionados a cada uma dessas abordagens.

3.3.1 Filtragem de Informação Colaborativa

Normalmente, nessa abordagem há uma *matriz de avaliações* que representa o histórico de preferências dos usuários, onde nas colunas estão os itens do repositório e nas linhas estão os usuários. Cada célula da matriz representa o grau de interesse de um usuário por um determinado item.

Os mais recentes Sistemas de Recomendação usam técnicas de filtragem colaborativa baseada nos *k vizinhos mais próximos* – kNN (Resnick et al. 1994, Shardanand et al. 1995). Os algoritmos baseados no kNN calculam as distâncias entre os usuários da comunidade a partir do histórico de suas preferências. Questões como "*Qual o nível de interesse de um usuário por um determinado produto ainda desconhecido pelo consumidor?*", são respondidas a partir da média ponderada das opiniões dos vizinhos mais próximos do usuário para aquele produto em questão.

É comum que as notas de tais usuários estejam em um intervalo diferente apesar de representarem o mesmo nível de interesse. Herlock et al. (1999) propõe um algoritmo para ajustar as avaliações dos usuários em uma mesma escala.

Como visto na seção 2.3.2, a filtragem colaborativa baseada no kNN possui a vantagem de rapidamente incorporar na lista de recomendações de um usuário itens totalmente inéditos (ou surpreendentes), visto que para isso é suficiente que alguns dos vizinhos desse indivíduo tenham avaliado um determinado item para que ele seja

recomendado [6]. No entanto, esses algoritmos apresentam-se inadequados em comunidades muito grandes, visto que é lenta a busca pelos usuários mais semelhantes ao usuário alvo. Assim, algumas heurísticas podem ser adotadas para fazer uma amostragem dos usuários da comunidade. Adicionalmente, Breese et al. (1998) analisou soluções para esse problema com base em uma mudança na representação interno do perfil do usuário, usando redes *bayesianas* ou métodos de *agrupamento* (*clustering*).

As *redes bayesianas* criam um modelo a partir do conjunto de treinamento. Basicamente esse modelo tem em cada nó uma estrutura semelhante a uma árvore de decisão. As arestas entre os nós são informações dos consumidores. A construção do modelo é feita *off-line* e pode levar horas ou dias. No entanto, com o modelo formado o processo de recomendação é muito rápido e tão preciso quanto os algoritmos baseados no kNN. Essa abordagem se mostrou adequada para domínios em que as preferências do usuário mudam lentamente, visto que o modelo não precisa ser atualizado freqüentemente.

Os métodos de *clustering* funcionam agrupando usuários com perfis semelhantes em grupos (*clusters*) bem definidos. É possível que um usuário coexista em vários grupos simultaneamente com um fator de pertinência diferente para cada um deles. Esse processo, como a abordagem anterior, pode ser feito *off-line*, o que provê um ganho de performance com relação aos algoritmos baseados no kNN. Uma vez que os grupos estão construídos, as predições para um usuário podem ser feitas a partir da ponderação das opiniões dos usuários do mesmo grupo. Breese et al. (1998) constatou que a precisão dos métodos de *clustering* é menor do que em abordagens baseadas no kNN, por exemplo. Mesmo assim, em alguns sistemas, como websites de Comércio Eletrônico em que o número de usuários é enorme, é conveniente o uso dessa abordagem. Pode-se, por exemplo, paralelizar o processamento em vários engenhos de personalização, cada qual com um *cluster* de usuários ou de produtos.

Gokhale e Claypool (1999) conseguiram melhorar o algoritmo de filtragem colaborativa, na medida em que obtiveram predições mais

[6] Note que isso não ocorre na filtragem baseada em conteúdo, pois nesse tipo de abordagem o conteúdo dos itens é determinante para a geração da lista de recomendação.

precisas, usando limiares para restringir sugestões a usuários com um histórico longo em comum ou com uma forte correlação.

3.3.2 Filtragem de Informação Baseada em Conteúdo

Como visto na seção 2.3.2, diferentemente dos métodos de filtragem colaborativos, a filtragem baseada em conteúdo requer a descrição dos atributos dos itens a serem filtrados. Uma possibilidade para se representar o perfil do usuário nesse tipo de abordagem é através de uma tabela contendo a descrição dos itens avaliados pelo usuário com sua respectiva classificação. A Figura 10 ilustra tal representação no domínio de recomendação de filmes.

Filme	Gênero	Ano	País	Elenco	Diretor	Sinopse	Nota
M_1	G_1	Y_1	C_1, C_4	A_1, A_2, A_5, A_7	D_3, D_4	Bla bla ...	✶ ✶
			⋮				
M_p	G_2	Y_3	C_4	A_1, A_4, A_5, A_6	D_3	Bla bla bla ...	✶ ✶ ✶ ✶ ✶

Figura 10 - Ilustra a representação do perfil do usuário em um sistema de recomendação de filmes.

O algoritmo kNN (Cover e Hart 1974), e outros algoritmos de *aprendizagem baseada em instância* (Aha et al. 1991) podem ser adaptados para Sistemas de Recomendação como soluções de filtragem de informação baseada em conteúdo (Arya 1995, Cotter e Smyth 2000, Krukwich e Burkey 1996, Balanovic e Shoham 1997).

Nesse caso, os exemplos são instâncias originais do conjunto de treinamento (itens do perfil do usuário). Durante a aprendizagem, esses algoritmos usam uma *função de distância* para determinar quão próximo um novo vetor de entrada y está a cada instância da memória, e utiliza as instâncias mais próximas para inferir a classe de saída de y.

Um dos grandes problemas da maior parte dos algoritmos de filtragem por conteúdo é a baixa eficiência em situações em que o número de itens do repositório é muito grande. Esse problema é ainda mais evidente no kNN à medida que o número de itens no perfil do usuário aumenta, visto que cada item do perfil é comparado com todos os itens do

repositório a serem filtrados. A priori esse problema pode ser minimizado de pelo menos três formas (Bezerra et al. 2002a). Uma das mais intuitivas é a redução dos itens que constituem o perfil do usuário de forma a mantê-lo o mais informativo possível (Wilson e Martinez 2000).

Outra possibilidade seria a utilização de estruturas para indexação dos exemplos que constituem o perfil do usuário. Nesse sentido, é possível utilizar uma estrutura como *k-D Trees* [7] (Bentley 1975) para efeito de organização dos itens avaliados pelo usuário não necessitando uma redução nesse conjunto. A idéia básica seria construir árvores de busca binária com instâncias do conjunto de treinamento como nós, utilizando os atributos das instâncias como chaves em uma seqüência alternada conveniente. A fim de favorecer o balanceamento das *k-D Trees*, o atributo escolhido para cada dimensão é aquele que melhor divide as instâncias da respectiva sub-árvore. Bezerra et al. (2002a) verificou que em se tratando da qualidade, no contexto de Sistemas de Recomendação Personalizados, há equivalência entre um sistema de filtragem por conteúdo baseado no método kNN sem *k-D Trees*, e um outro sistema também baseado no método kNN, contudo, acrescido de *k-D Trees*. Todavia, foi ainda constatado que com relação à velocidade as *k-D Trees* provêm melhores resultados se comparadas ao kNN padrão.

A terceira abordagem para o problema de desempenho é modificar a representação original dos exemplos armazenados no perfil do usuário criando uma nova estrutura interna. Baseado nessa idéia, foi proposto o método *RISE* por Domingos (1995). Apesar de alcançar bons resultados essa técnica não é aplicável a todos os domínios visto que não suporta atributos complexos como, por exemplo, os nominais multivalorados (o elenco de um filme, por exemplo).

No próximo capítulo é descrito um método de filtragem de informação por conteúdo que resolve o problema de desempenho através desta última abordagem, ou seja, uma mudança na representação original dos exemplos armazenados no perfil do usuário. Este método e algumas de suas variações são descritos em Bezerra et al. (2002a, 2002b), em De

[7] K-D Trees são estruturas de dados dinâmicas e adaptáveis que são muito similares a árvores binárias mas dividem o espaço geométrico de forma adequada para solucionar problemas como busca em intervalos (*range searching*).

Carvalho e Bezerra (2002), em Bezerra e De Carvalho (2003) e em Bezerra e De Carvalho (2004).

3.3.3 Abordagens de Filtragem Híbrida

A Tabela 1 mostra aspectos positivos e negativos da Filtragem de Informação Colaborativa e da Filtragem de Informação Baseada em Conteúdo. A partir dessa tabela, não é difícil observar que os problemas mais críticos das abordagens de Filtragem Colaborativa não ocorrem nas abordagens de Filtragem Baseada em Conteúdo, e vice-versa. De fato, estas duas técnicas são consideradas complementares (Balanovic e Shoham 1997). Essa foi justamente uma das grandes motivações de alguns estudos de abordagens de filtragem híbridas surgidos nos últimos anos.

Filtragem Colaborativa	Filtragem Baseada em Conteúdo
Admite itens com atributos não triviais. Por exemplo, é possível que ao avaliar um filme um determinado usuário considere quão bela é cenografia do mesmo.	Não possui problemas de latência para itens novos. É comum que os estoques de grandes lojas virtuais sejam renovados a todo momento com centenas de novos itens. Dessa forma, essa é uma característica bastante interessante, visto que mesmo os produtos mais novos do repositório seriam passíveis de recomendações.
Recomendação de itens interessantes sem conteúdo diretamente relacionado ao histórico. Na seção 3.3.1, foi mencionado o caso de que algumas recomendações poderiam surpreender ao usuário pelo fato de serem totalmente inusitadas ou ter qualquer relação com o histórico de suas avaliações.	Bons resultados mesmo para usuários incomuns. É natural que as comunidades de usuários possuam pessoas diferentes ou comumente chamadas de "ovelhas-negras". Tendo em vista que a filtragem por conteúdo é calculada apenas em função das semelhanças entre o perfil do usuário e o repositório digital, mesmo um usuário com perfil *sui generis* poderá receber boas recomendações.
Julgamento em muitas dimensões. Pelo fato da filtragem colaborativa considerar as semelhanças entre os perfis de usuários para calcular recomendações para um usuário alvo, é possível que dimensões complexas de serem representadas, como são as dimensões temporal e espacial, sejam implicitamente consideradas no processo de recomendação.	Precisão independente do número de usuários, visto que o sucesso da filtragem por conteúdo independe da comunidade de usuários.

Tabela 1 - Filtragem de Informação Colaborativa *versus* Filtragem de Informação baseada em Conteúdo.

Uma das possibilidades de combinar as duas técnicas de filtragem de informação é ponderar as predições geradas por ambas. Essa proposta foi apresentada por Claypool et al. (1999) em um sistema de jornal on-line denominado *PTango*. Nesse sistema o perfil do usuário é constituído de palavras-chave tanto fornecidas pelo usuário, quanto geradas implicitamente a partir de artigos avaliados positiva e explicitamente por ele. Uma possível extensão desse sistema é prover um mecanismo de aquisição implícita de palavras-chave.

Vimos na seção 2.3.2 que um dos problemas da filtragem colaborativa é a latência na recomendação de novos itens. Uma solução para esse problema foi proposta por Sarwar et al. (1998) no sistema *GroupLens*. A idéia básica é avaliar itens recém adicionados no repositório de forma automática com agentes denominados *filterbots*. No sistema, as avaliações do *filterbot* são tratadas como avaliações de usuários reais. O processo de avaliação do *filterbot* se dá considerando o conteúdo descritivo dos itens do repositório. Mostrou-se que essa estratégia melhora a qualidade das recomendações.

Uma outra estratégia de combinar os benefícios das duas técnicas de filtragem de informação foi proposta por Balanovic e Shoham (1997) no sistema *Fab*, um sistema de recomendação de páginas Web. No *Fab*, o perfil contendo as páginas de interesse do usuário era construído a partir de técnicas baseadas no conteúdo das páginas. Dessa forma, a determinação dos usuários mais correlatos de um usuário alvo seria efetuada a partir da semelhança existente entre os perfis dos mesmos. Determinados os usuários semelhantes o próximo passo seria recomendar novos itens usando técnicas convencionais da filtragem colaborativa.

3.3.4 Métodos de Aprendizagem de Máquina e Outras Abordagens

Classificadores são geralmente modelos computacionais para atribuir uma categoria a uma entrada (Mitchell 1997). Por exemplo, considere o problema de classificar uma pessoa como bom ou mau pagador em um sistema de análise de crédito. A entrada nesse caso poderia ser o conjunto de todas as informações pertinentes a essa pessoa, como seu endereço, sua profissão, sua renda, etc. O sistema deveria dizer a partir dessa entrada se uma pessoa é um bom ou mau cliente. No

domínio de aplicações de Comércio Eletrônico cada usuário teria seu próprio classificador. Para recomendar itens a um usuário o classificador receberia como entrada, por exemplo, um vetor das características (atributos) dos itens a serem filtrados e, poderia fornecer como saída um *score* indicando a possibilidade do item ser recomendado ou não. Classificadores podem ser implementados usando diferentes técnicas de aprendizagem de máquina como *regras de associação*, *redes neurais* e *redes bayesianas* (Mitchell 1997).

Basu et al. (1998) e Good et al. (1999) são exemplos de propostas onde há uma adaptação do problema de recomendação em um problema de classificação. Basu et al. (1998) construiu um Sistema de Recomendação híbrido usando um classificador de aprendizagem por indução, que misturava a essência da abordagem colaborativa com a abordagem baseada em conteúdo. Já, Good et al. (1999) propôs um classificador de aprendizagem indutiva através de vetores de características, o qual não provê melhores resultados que a abordagem colaborativa baseada no kNN, mas pode melhorar substancialmente a qualidade se for utilizado em conjunto.

Algumas das estratégias de marketing muito utilizadas nos últimos anos foram propiciadas pelo que se convencionou chamar de *regras de associação*. Trata-se de uma abordagem muito mais voltada para a estratégia de marketing segmentado do que a estratégia de marketing *one-to-one*. Ela permite diagnosticar com determinados níveis de *suporte* e *confiança* (Agrawal et al. 1993), por exemplo: "*50% dos consumidores que consomem cerveja na sexta-feira também compram fraudas*". Dessa forma, essas técnicas são úteis não só para analisar padrões de preferências de produtos como também para recomendar produtos a consumidores baseando-se nos produtos que tenham sido comprados pela comunidade no passado.

As regras de associação podem ser construídas *off-line*, no entanto esse processo pode ser muito custoso quando o número de itens do repositório for significativamente grande, visto que o número de regras cresce exponencialmente com o número de produtos na regra. Agrawal e Srikant (1994) propuseram uma solução para esse problema através do algoritmo APRIORI. Depois de construídas as regras de associação, o

processo de recomendação pode ser extremamente eficiente, além de ser uma representação bem mais compacta do que a *matriz de avaliações*. Finalmente, Brin et al. 1997 e Aggarwal e Yu 1998 diagnosticaram alguns problemas no *framework suporte-confiança* e propuseram uma medida alternativa para a *confiança* chamada de *Lift*.

Uma modelagem estatística do problema de recomendação, chamada de *repeat-buying theory*, foi apresentada por Ehrenberg (1988) e avaliada experimentalmente por Geyer-Schulz e Hahsler (2002). Essa teoria tem como pressuposto que em uma situação estacionária, com todas as compras de itens independentes uma da outra, o comportamento de compra segue um processo que produz uma distribuição binomial negativa (NBD) para um dado número de compras repetidas. O modelo estatístico é baseado fortemente em várias hipóteses comportamentais sobre o padrão de compra do consumidor. Constatou-se que o mesmo é adequado para vários mercados de consumo. Uma variação dessa abordagem pressupõe uma distribuição logarítmica, LSD (Geyer-Schulz e Hahsler, 2001).

Outras abordagens baseiam-se na estrutura de dados *grafos*, permitindo a exploração de relacionamentos transitivos entre os usuários, o que não ocorre com a abordagem colaborativa baseada no kNN. *Horting*, proposto por Wolf et al (1999), é um método em que os usuários são representados pelos nós de um grafo e as arestas são ponderadas para refletir o nível de semelhança entre os usuários. Constatou-se que a precisão das recomendações com *Horting* é superior ao alcançado com a abordagem baseada no kNN.

3.4. VERIFICANDO A QUALIDADE DAS RECOMENDAÇÕES

A avaliação de Sistemas de Recomendação é ainda uma questão em aberto. Isso se deve principalmente a dois fatores. Primeiro o desenvolvimento de Sistemas de Recomendação é uma inovação recente e ainda não há um consenso sobre uma métrica para avaliar esse tipo de sistema. O outro motivo está associado ao entendimento de qual é o objetivo de um Sistema de Recomendação, ou seja, que tipo de recomendação é usada e quando pode ser considerada uma boa

recomendação. Enumeramos na seção 2.2 diversas tarefas que um Sistema de Recomendação pode fazer. Nessa seção, descrevemos algumas medidas de desempenho que vem sendo utilizadas em análises estatísticas, na avaliação de algoritmos de aprendizagem de máquina e de recuperação de informação. Além disso, algumas métricas que têm se mostrado adequadas para análise de desempenho de tarefas específicas dos Sistemas de Recomendação são descritas.

3.4.1 Métricas Oriundas do Domínio de Aprendizagem de Máquina

No contexto de aprendizagem de máquina identificamos pelo menos três medidas de desempenho que podem ser adequadas ao domínio de Sistemas de Recomendação. Antes de apresentá-las, é conveniente a introdução de um conceito ilustrado na tabela abaixo.

	Predição Negativa	**Predição Positiva**	**Σ**
Saída desejada negativa	a	b	a+b
Saída desejada positiva	c	d	c+d
Σ	a+c	b+d	N

Tabela 2 - Matriz de Confusão

A matriz de confusão é uma abstração para as possíveis relações que podem ocorrer entre a saída obtida de um sistema de aprendizagem ao ser executado a partir de um conjunto de dados definido como entrada e a real saída que deveria ocorrer para esse conjunto. Nessa matriz, estão representadas duas classes possíveis, a positiva (no contexto de Sistemas de Recomendação essa classe pode significar o conjunto de itens que o usuário pode ter afinidade) e a negativa (por conseqüência poderiam ser os itens não interessantes para o usuário). Adicionalmente, essa matriz contém tanto a saída gerada pelo sistema (predição), quanto a saída esperada ou real para um determinado conjunto de dados (saída real).

Dessa forma, a matriz de confusão mostra na coluna *predição positiva* quantas das possíveis recomendações foram preditas pelo sistema e, ainda mais, quantas destas foram de fato corretas (célula *d*) e quantas

não deveriam ter sido recomendadas (célula *b*). A matriz permite identificar também, a partir da coluna *predição negativa*, quantas das possíveis recomendações foram descartadas pelo algoritmo e, adicionalmente, quantas destas foram corretamente rejeitas (célula *a*) e quantas destas deveriam na verdade ter sido recomendadas pelo sistema (célula *c*). Finalmente, temos que *N* é a quantidade de itens que constituem o conjunto de teste, ou seja, o número de itens propício a recomendações.

Baseada na matriz de confusão foram definidas várias métricas, entre elas a *acurácia* (*accuracy*), mostrada na Equação 1. No contexto de Sistemas de Recomendação (Mobasher et al. 2000, Yu et al. 2001, Lin et al. 2002), a *acurácia* é a fração correta de todas as possíveis recomendações.

$$Accuracy = \frac{a+d}{N}$$	**Equação 1**

O *erro absoluto médio* (MAE), mostrado na Equação 2, que pode ser definido a partir da matriz de confusão, também tem sido utilizado ao longo dos anos na avaliação dos Sistemas de Recomendação (Shardanand e Maes 1995, Herlock et al. 1999, Sarwar et al. 2000, Vucetic e Obradovic 2000, Yu et al. 2001, Mobasher et al. 2002, Teixeira 2002).

$$MAE = \frac{b+c}{N}$$	**Equação 2**

Uma outra medida utilizada no domínio de recomendações (Herlock et al. 1999, Mobasher et al. 2000) é a *cobertura* (Equação 3), que mede a fração de itens que o sistema está apto a recomendar. Por exemplo, pode haver alguns itens que o sistema jamais possa recomendar. Uma situação em que isso ocorre freqüentemente é quando o sistema é baseado em uma abordagem de filtragem colaborativa e existem itens que não

foram avaliados por qualquer usuário da comunidade. Assim, torna-se impossível para o sistema gerar predições para tais itens. Para isso, considere que o total de itens do repositório seja *P* e apenas *N* deles são passíveis de recomendação.

$Cobertura = \frac{N}{P}$	**Equação 3**

A seguir descrevemos algumas métricas oriundas do domínio de recuperação de informação.

3.4.2 Medidas do Domínio de Recuperação de Informação

Os Sistema de Recomendação possuem diversas tarefas (seção 2.2) com o propósito de auxiliar o usuário a encontra itens ou informações relevantes dentre uma grande diversidade de opções. De certa forma, isso pode ser visto com uma das tarefas de recuperação tratadas no domínio de *recuperação de informações*. Logo, métricas desse domínio podem ser úteis para a avaliação do desempenho dos algoritmos de recomendação. *Precisão* (Equação 4) e *recall* (Equação 5) são duas dessas métricas que são definidas a partir da matriz de confusão, assim como *MAE* e *Exatidão*.

$Precisão = \frac{itens\ corretamente\ recomendados}{total\ de\ itens\ recomendados} = \frac{d}{b+d}$	**Equação 4**

$Recall = \frac{itens\ corretamente\ recomendados}{total\ de\ recomendações\ esperadas} = \frac{d}{c+d}$	**Equação 5**

Precisão e *Recall* são medidas conflitantes quando o sistema avaliado gera predições erradas, ou seja, $b \neq 0$ ou $c \neq 0$. Assim, uma alta *precisão* (valor de *b* pequeno) significa baixo *recall* (valor de *c* grande), e

vice-versa. Para encontrar um equilíbrio entre elas uma medida chamada de *F-measure* (Equação 6) pode ser usada (Sarwar et al. 2000, Mobasher et al. 2002).

$F - measure = \frac{2 * precisão * recall}{precisão + recall} = \frac{2}{1/precisão + 1/recall}$	**Equação 6**

A seguir descrevemos duas medidas, *ROC* e *Breese*, que vem sendo cada vez mais usadas no domínio de recomendações, principalmente, pela adequação às características inerentes aos Sistemas de Recomendação.

3.4.3 Área sob a curva ROC

A curva "*relative operating characteristic*", ou simplesmente curva ROC, foi introduzida na comunidade de recuperação de informação por Swets (1979). A curva ROC é utilizada para medir o quanto um valor produzido por um sistema é capaz de distinguir os elementos relevantes dos não relevantes.

O sistema medido tem como saída uma variável de relevância associada a cada elemento. O conhecimento sobre a relevância dos elementos permite construir duas curvas, uma para os valores obtidos para os elementos relevantes e outra para os elementos não relevantes. No momento de selecionar os elementos relevantes, o sistema usa um limiar *t*. Os elementos que ultrapassem esse limiar serão considerados relevantes e portanto selecionados. Caso contrário, esse elementos serão considerados irrelevantes e então rejeitados. Para cada valor escolhido para o limiar *t* é possível calcular o índice de relevância ou *recall* (proporção dos elementos relevantes que são selecionados) e o índice de irrelevância ou *fallout* (proporção dos elementos não relevantes que são selecionados).

A utilização do *recall* e do *fallout* é bastante comum na avaliação de sistemas que tem como objetivo a seleção binária através de uma variável resultante separada por um limiar. O problema é que os valores dessas grandezas dependem do limiar escolhido. A curva *ROC* por sua vez

é capaz de avaliar a capacidade do sistema separar os elementos relevantes dos não relevantes sem depender da escolha de um limiar, permitindo comparar dois sistemas mais facilmente. A curva *ROC* é a curva obtida quando desenhado em um plano cartesiano os valores do *recall* (eixo y) versus o *fallout* (eixo x) para diferentes valores do limiar *t* (Figura 11).

As curvas *ROC* são bastante úteis para observar a capacidade de um algoritmo de filtragem de informação separar informação relevante da não relevante. Entretanto, comparar várias curvas *ROC* pode ser um processo trabalhoso e impreciso. A *área sob uma curva ROC* pode ser usada como um valor que expressa a capacidade do sistema discriminar os itens relevantes dos não relevantes. Pode-se constatar, por exemplo, que se a área sob a curva é de 0.5, então em 50% das vezes o sistema distinguiria corretamente os elementos e em 50% das vezes erroneamente, o que equivaleria a uma classificação puramente aleatória, ou seja, a um desempenho nada satisfatório. Valores menores que 0.5 indicam que o sistema está trocando os elementos relevantes pelos não relevantes, enquanto valores próximos de 1.0 indicam uma boa precisão na seleção dos elementos.

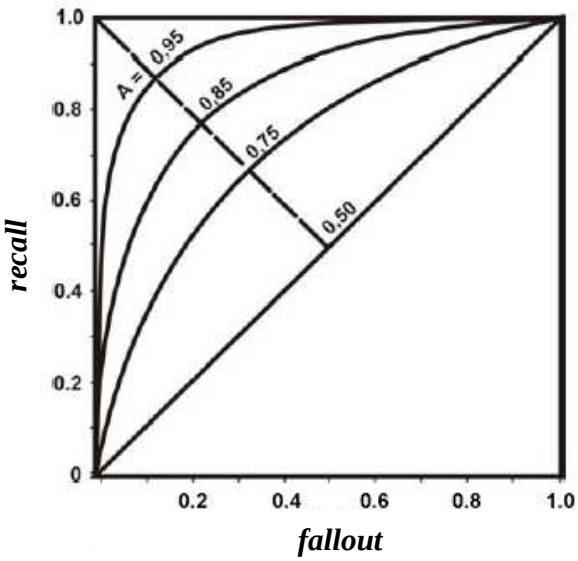


Figura 11 - Ilustra a área sob a curva ROC.

A área sobre a curva ROC é uma métrica interessante quando se está avaliando o desempenho de um Sistema de Recomendação cuja tarefa é gerar uma lista de recomendação com itens que se assemelham ao perfil do usuário para que ele possa escolher quais itens consumir (item (vi) da seção 2.2). Alguns dos trabalhos que a utilizaram são Herlock et al. 1999, Vucetic e Obradovic 2000 e Teixeira 2002. A seguir é apresentada a métrica *Breese*.

3.4.4 Métrica Breese

A métrica criada por Breese et al. (1998) é usada para calcular a utilidade para o usuário de uma lista ordenada por um sistema de filtragem. Basicamente, a utilidade esperada de uma lista para um determinado usuário é calculada em função da probabilidade deste usuário ver um item recomendado vezes a utilidade do item propriamente dito.

Em um sistema de filtragem de informação onde o resultado de uma filtragem é retornado em uma lista de itens ordenados[8] pela relevância, é fato que na maior parte dos casos o usuário irá investigar apenas os primeiros elementos da lista na sua busca por itens relevantes. A métrica *Breese* tem como principal vantagem levar em consideração essa observação no seu cálculo, o que tem justificado sua utilização cada vez mais freqüente na avaliação de Sistemas de Recomendação (Penock et al. 2000, Bezerra et al. 2002a, 2002b, De Carvalho e Bezerra 2002, Teixeira 2002, Miller et al. 2003). A métrica possui uma constante α chamada de *half-life*, que define a posição do item na ordem gerada pelo sistema em que o usuário tem 50% de chance de o observar. A utilidade de cada item será a diferença entre a avaliação dada pelo usuário e o valor médio *d* no intervalo de avaliação. A utilidade de uma lista ordenada para o usuário *a* é então:

$$R_a = \sum_j \frac{\max\left(r_{a,j} - d, 0\right)}{2^{(j-1)/(\alpha-1)}}$$	**Equação 7**

[8] Trata-se de uma variação do item (vi) da seção 2.2 devido à ordenação.

em que, *j* é a posição do item na lista ordenada pelo sistema e $r_{a,j}$ é a avaliação feita pelo usuário ao item que ocupa a posição *j*. Breese utiliza um *half-life* α igual a 5 e acrescenta que a utilização de um valor igual a 10 apresentou pouca diferença nos resultados. Para esse trabalho resolveu-se utilizar um *half-life* igual a 5. O valor final da métrica para o conjunto de usuários testados é calculado como segue:

$$R = 100 \frac{\sum_a R_a}{\sum_a R_{\max}}$$	**Equação 8**

em que, R_{max} é a utilidade máxima obtida quando todos os itens são ordenados de acordo com a avaliação do usuário. Essa última equação garante que o resultado não dependerá da seqüência nem do número das avaliações dos usuários escolhidos para um experimento.

Resumindo, a métrica proposta por *Breese* mede a capacidade do sistema ordenar os itens de acordo com sua relevância para o usuário. Além disso, ela dá mais ênfase à precisão dos primeiros itens na ordenação do sistema, os quais terão maior chance de serem apreciados pelo usuário. Devido a apresentar tais propriedades, essa métrica mostrou-se a mais adequada na avaliação experimental (Capítulo 5) do método desenvolvido nesse trabalho.

3.5. CONSIDERAÇÕES FINAIS

Nesse capítulo apresentamos vários trabalhos relacionados aos diversos aspectos dos Sistemas de Recomendação, estabelecendo na medida do possível suas relações com o Capítulo 2. Sua importância é fundamental para melhor compreensão da área e, sobretudo para o diagnóstico de problemas não tratados pelas abordagens atuais.

Nesse sentido, constatou-se a partir da seção 3.2 que apesar de haver alguns trabalhos relacionados ao problema de aquisição do perfil do usuário, muito trabalho pode ser feito no que diz respeito a aquisição implícita do perfil. De fato, o objetivo maior deste trabalho não é propor soluções para o problema de aquisição em especial. Todavia, isso não diminui a importância dessa seção, visto que um dos objetivos desse trabalho é o estudo dos Sistemas de Recomendação e sua relação com a Inteligência Artificial.

A seção 3.3 relatou diversas técnicas possíveis para geração de recomendações para o usuário baseado em um perfil previamente coletado. Uma análise dessa seção mostra que as técnicas atuais resolvem parte dos problemas relatados na seção 2.3.3, mas ainda apresentam dificuldades em se tratando do uso de recursos computacionais, como velocidade de resposta e utilização de memória. Essa observação é extremamente importante para o restante do trabalho, tendo em vista que ela serviu de motivação para o desenvolvimento de um método de filtragem que melhore a velocidade de resposta e promova um melhor uso de memória sem comprometer a qualidade das recomendações.

Finalmente, a seção 3.4 apresentou as técnicas e as métricas de avaliação no contexto de Sistemas de Recomendação. A partir dessa seção, ficou claro que há métricas mais ou menos adequadas, a depender da tarefa realizada pelo sistema. Dessa forma, pode-se adiantar que para a avaliação do método proposto neste trabalho será usada como métrica o critério definido por Breese et al. 1998, tendo em vista que a tarefa principal do sistema a ser avaliado será a qualidade do sistema na geração de listas de recomendações (item (vi) da seção 2.2).

No próximo capítulo é apresentado em detalhes o método de filtragem de informação baseado em dados simbólicos modais, que se trata de um conceito definido no contexto de Análise de Dados Simbólicos (seção 4.2).

Capítulo 4

FILTRAGEM DE INFORMAÇÃO BASEADA EM DADOS SIMBÓLICOS MODAIS

4.1. INTRODUÇÃO

Uma das motivações desse trabalho é o fato da Filtragem Colaborativa não se aplicar em todos os contextos possíveis em sistemas de recomendação, o que justifica o uso de técnicas de Filtragem Baseada em Conteúdo, inclusive em combinação com a primeira. Como mencionado anteriormente, os métodos de Filtragem por Conteúdo que possuem melhor desempenho geralmente possuem problemas de velocidade e espaço, como é o caso dos métodos baseados no kNN.

Neste capítulo é apresentado um novo método de filtragem de informação por conteúdo baseado nas teorias de Análise de Dados Simbólicos (seção 4.2). Na abordagem proposta, o perfil do usuário (seção 4.3) é modelado através de um conjunto de descrições simbólicas modais que sumarizam as informações dos itens previamente avaliados. Uma função de dissimilaridade que leva em conta as diferenças em posição e em conteúdo foi criada a fim de possibilitar a comparação entre um novo item e o perfil do usuário (seção 4.4). Para exemplificar o método proposto é apresentada na seção 4.5 uma aplicação real no domínio de recomendação de filmes.

A seguir é introduzido o domínio de Análise de Dados Simbólicos, que fornece o embasamento científico para a abordagem de Filtragem de Informação Baseada em Dados Simbólicos Modais.

4.2. ANÁLISE DE DADOS SIMBÓLICOS

Análise de Dados Simbólicos (*Symbolic Data Analysis* - SDA) é um novo domínio no que se refere à Descoberta de Conhecimento (*Knowledge Discovery*), que tem relação com: análise multivariada, reconhecimento de padrões, banco de dados e inteligência artificial. SDA provê ferramentas específicas para se trabalhar com dados complexos, agregados, relacionais e de alto nível, descritos por variáveis multivaloradas onde as entradas da tabela de dados são conjuntos de

categorias ou de números, intervalos ou distribuições de probabilidade associadas a regras e taxonomias. [9]

Os métodos SDA generalizam os métodos clássicos de análise exploratória de dados, como técnicas fatoriais, árvores de decisão, discriminação, regressão, métodos neurais, escalonamento multidimensional, classificação supervisionada, agrupamentos e reticulados conceituais.

A fim de compreender a essência de SDA é descrita a seguir a principal entrada dos algoritmos desenvolvidos na área, ou seja, a tabela de dados simbólicos.

4.2.1 Tabela de Dados Simbólicos

Por questões metodológicas, é conveniente descrever a entrada dos algoritmos de Análise de Dados clássica antes de introduzir propriamente a tabela de dados simbólicos.

Em Análise de Dados clássica, a entrada é uma tabela de dados onde as linhas são as descrições dos indivíduos, e as colunas são as variáveis. Uma célula dessa tabela ou é um valor quantitativo simples ou é uma categoria (observe a Tabela 3).

Pessoa	Idade	Altura (m)	Peso (kg)	Sexo
W_1	18	1,70	74	M
W_2	25	1,60	51	F
W_3	60	1,58	60	F
W_4	14	1,50	55	M
W_5	10	1,10	42	F

Tabela 3 - Ilustra uma tabela de dados clássica, que contém apenas variáveis categóricas ou quantitativas simples.

À medida que um conjunto considerável de informações é agregado em dados mais coesos e manipuláveis é necessária uma tabela de dados mais complexa, visto que suas células não contêm apenas dados simples como são os atributos categóricos ou quantitativos. Essa tabela é chamada de tabela de dados simbólicos.

[9] Para mais detalhes consulte o website http://www.jsda.unina2.it/

Para exemplificar estes conceitos, considere que os dados da Tabela 3 possam ser agregados para representar um determinado grupo de pessoas. A forma de agregar se dá ao nível de cada atributo. Assim, é possível que a variável idade seja representada por faixas de idade como: *criança* (de 0 a 9 anos), *pré-adolescente* (de 10 a 14 anos), *adolescente* (de 15 a 18 anos), *jovem* (de 19 a 24 anos), *adulto* (de 25 a 60 anos) e *terceira idade* (a partir de 60 anos). Adicionalmente, considere que as variáveis altura e peso sejam categorizadas em {*baixo,médio,alto*} e {*magro,normal,gordo,obeso*}, respectivamente. Dessa forma, pode-se obter a seguinte tabela de dados, que possui em sua primeira linha os dados da Tabela 3 com uma possível representação no domínio de Análise de Dados Simbólicos.

Grupo	Idade	Altura (m)	Peso (kg)	Sexo
G_1	{(Pré-adolescente, 2/5), (Adolescente, 1/5), (Adulto, 2/5)}	{(Baixo, 2/5), (Médio, 2/5), (Alto, 1/5)}	{(Magro, 1/5), (Normal, 3/5), (Gordo, 1/5)}	{(M,2/5), (F,3/5)}

Tabela 4 - Ilustra uma tabela de dados simbólicos.

É importante salientar que outros grupos de pessoas podem ser representados na mesma tabela. Para isso cada grupo ocuparia uma linha na Tabela 4. Observe também que as células dessa tabela são mais complexas que as usuais, visto que elas são estruturadas e representam variação interna. Os tipos de dados que comportam esse nível de complexidade são conhecidos como dados simbólicos e serão detalhados a seguir.

4.2.2 Dados Simbólicos

Algumas vezes no mundo real a informação registrada é muito complexa para ser descrita por dados usuais. Esse é o motivo pelo qual diferentes tipos de variáveis simbólicas e dados simbólicos têm sido introduzidos (Bock, H. H. e Diday, E. 2000). Por exemplo, para um determinado objeto, uma variável de intervalo corresponde a um intervalo do seu domínio. Ao mesmo tempo, uma variável categórica multivalorada é definida por um subconjunto do seu domínio. Por fim, uma variável modal

é representada por uma medida não-negativa na forma de uma freqüência, ou de uma distribuição de probabilidade, ou de um sistema de pesos.

Nesse trabalho são manipulados basicamente dados simbólicos modais. A fim de exemplificar esse tipo de dados considere o domínio de filmes, que possui entre outros atributos aqueles descritos na Tabela 5.

Variável	Tipo	Exemplo
Gênero	Qualitativo univalorado	*Ficção Científica*
País	Qualitativo univalorado	*EUA*
Diretor	Qualitativo univalorado	*Andy Wachowski*
Elenco	Qualitativo multivalorado	*Carrie-Anne Moss , Hugo Weaving , Joe Pantoliano , Keanu Reeves , Laurence Fishburne*
Ano	Qualitativo univalorado ordenado	*1999*
Sinopse	Textual	*Até que ponto o que vivemos é a realidade? Descobrir que o mundo é uma farsa criada por máquinas poderosas para nos controlar pode significar o fim para muita gente. Mas, em Matrix, isso é o início de uma luta surpreendente, que irá perturbar você do começo ao fim.*
Nota[10]	Quantitativo univalorado	*5*

Tabela 5 - Tipos de variáveis no domínio de filmes, exemplificado através do filme *Matrix*.

Seja O_j um conjunto finito de categorias. Por exemplo, se j representa a variável *Elenco*, então O_j poderia ser o conjunto de todos os atores e atrizes do mundo. Uma variável modal y_j com domínio O_j, definido pelo conjunto de objetos E={a, b, ...}, é uma variável multivalorada se para cada objeto $a \in E$, além de ser dado um subconjunto de O_j, é dado para cada categoria m desse subconjunto um peso $w(m)$ que reflete a importância da categoria m para o objeto a. Formalmente, $y_j(a) = (S_j(a), q_j(a))$ onde $q_j(a)$ é a distribuição dos pesos definida para $S_j(a) \subseteq O_j$ tal que para cada categoria $m \in S_j(a)$ existe um peso $w(m)$ associado. $S_j(a)$ é chamado de suporte da medida $q_j(a)$ no domínio O_j.

[10] O atributo *Nota* possui a característica de não ser um descritor, mas sim uma variável que determina uma classe ou mesmo um *score* de um filme para um usuário. Para isso, considere que o domínio deste atributo é {1,2,3,4,5}, ou seja, os filmes podem ter notas no intervalo inteiro de 1 a 5.

Suponha, por exemplo, que o filme mostrado na Tabela 5 seja "Matrix". Para esse filme o atributo *Elenco* pode ser representado pela variável simbólica modal $y_{elenco}(Matrix) = (\{Carrie\text{-}Anne\ Moss, Hugo\ Weaving, Joe\ Pantoliano, Keanu\ Reeves, Laurence\ Fishburne\}, \{0.2, 0.2, 0.2, 0.2, 0.2\})$. Dessa forma, o suporte para essa variável é $S_{elenco}(Matrix) = \{Carrie\text{-}Anne\ Moss, Hugo\ Weaving, Joe\ Pantoliano, Keanu\ Reeves, Laurence\ Fishburne\}$ enquanto que a medida que suporta tal atributo para o filme *Matrix* é $q_{elenco}(Matrix) = \{0.2, 0.2, 0.2, 0.2, 0.2\}$, que significa que todos os atores possuem a mesma importância.

Uma descrição simbólica de um item é um vetor cujos descritores (atributos) são variáveis simbólicas. Ou seja, na abordagem proposta, cada item é descrito por um vetor onde cada componente é representado por uma distribuição ponderada dada por uma variável simbólica modal.

A seguir (seção 4.3) veremos que na abordagem desenvolvida neste trabalho, o perfil do usuário é formado por dois sub-perfis, um positivo e outro negativo, em que cada qual constitui um vetor cujos descritores são variáveis simbólicas modais.

4.3. O PERFIL DO USUÁRIO

O perfil do usuário é representado por um conjunto de descrições simbólicas modais que sintetizam toda a informação contida nas descrições dos itens avaliados pelo usuário levando em conta a importância que o item possui para o usuário. Em nosso caso, essa importância é dada por uma nota.

A construção das descrições simbólicas modais do perfil do usuário envolve dois passos no processo de aprendizagem: *pré-processamento* e *generalização*. A idéia básica do método proposto é construir uma descrição simbólica modal para cada item avaliado pelo usuário (*pré-processamento*) e, após isso, agregar estas descrições simbólicas modais em um número pequeno de outras descrições simbólicas modais conforme a avaliação realizada pelo usuário (*generalização*). De fato, cada uma destas descrições simbólicas modais

obtidas a partir da fase de generalização reflete um tipo de interesse do usuário.

4.3.1 Pré-Processamento

O objetivo desse passo é associar cada item com uma descrição simbólica modal. Este passo é necessário para construir o conjunto de descrições simbólicas modais usado para representar o perfil do usuário e também para compará-lo com novos itens.

Seja $x_i = (X_i^1, ..., X_i^p)$ a descrição de um item $\boldsymbol{i}$ (i=1, ..., n), onde $X_i^j \subseteq O_j$ (j=1, ..., p) é um subconjunto de categorias do domínio O_j da variável y_j. O peso $w(m)$ de cada categoria $m \in X_i^j$ é dado por:

$$w(m) = \begin{cases} 1 \Big/ \left|X_i^j\right|, \text{ se } y_j \text{ é qualitativo simples ou multivalorado} \\ \dfrac{f(m) * IDF(m)}{\sum_{m \in X_i^j} f(m) * IDF(m)}, \text{ se } y_j \text{ é textual} \end{cases}$$

Equação 9

em que $f(m)$ e $IDF(m)$ são, respectivamente, a freqüência de m em X_i^j e a freqüência inversa de m, descrita a seguir. $|X_i^j|$ é a cardinalidade de X_i^j.

A Equação 10 representa a freqüência inversa de uma palavra m com relação a uma base de itens B ($|B|$ é a cardinalidade de B, ou seja, o total de itens da base). No domínio de recomendação de filmes, por exemplo, a base B é constituída de todos os filmes do repositório digital que permitirá a realização dos experimentos definidos no Capítulo 5. O termo Ψ_m^B define o número de itens que a palavra m aparece na base B. Dessa forma, quanto mais freqüente é o termo na base, menor sua importância ou contribuição. É importante salientar que o cálculo do IDF pode ser realizado *off-line*, de forma a não comprometer o desempenho dos métodos de filtragem.

$$IDF(m) = \log_2\left(\frac{|B|}{\Psi_m^B}\right)$$

Equação 10

A descrição simbólica modal associada ao item ***i*** é dada por $x_i=(X_i^1, ..., X_i^p)$ onde $X_i^j=(S_j(i), q_j(i))$ onde $j=1, ..., p$, e $S_j(i)$ é o suporte da distribuição ponderada $q_j(i)$. A descrição simbólica de alguns atributos representados na Tabela 5 é dada a seguir:

Variável	Exemplo
Gênero	*({Ficção Científica}, {1.0})*
País	*({EUA}, {1.0})*
Diretor	*({Andy Wachowski },{1.0})*
Elenco	*({Carrie-Anne Moss, Hugo Weaving, Joe Pantoliano, Keanu Reeves, Laurence Fishburne}, {0.2,0.2,0.2,0.2,0.2})*
Ano	*({1999},{1.0})*
⋮	
Nota	*5*

Tabela 6 - Descrição simbólica de alguns atributos obtidos a partir da Tabela 5.

Observe na Tabela 6 que o atributo *Nota* não é pré-processado pois essa variável é usada para avaliar o filme e não para descrevê-lo.

4.3.2 Generalização

O objetivo principal dessa fase é a representação do perfil do usuário através de descrições simbólicas, levando em consideração a avaliação do usuário para cada item. No domínio de recomendação de filmes, por exemplo, o perfil do usuário é formado por uma lista de filmes (tanto os filmes que o usuário gostou quanto os que ele não gostou) com suas respectivas notas. Na nossa abordagem cada perfil é formado por dois sub-perfis. O sub-perfil positivo é modelado por uma descrição simbólica modal que sintetiza toda a informação contida no conjunto de itens já avaliados positivamente pelo usuário, ou seja, os itens que receberam notas 4 ou 5. De forma semelhante, o sub-perfil negativo é modelado por uma descrição simbólica modal que sintetiza os itens que receberam notas 1 ou 2.

Formalmente, sejam u^+ e u^- os sub-perfis positivo e negativo do usuário u, respectivamente. Seja $\sigma = \{+, -\}$ e $y_u\sigma = (Y^1{}_u\sigma , ... , Y^p{}_u\sigma)$ a

descrição simbólica modal do sub-perfil u^{σ}, em que $Y^{j}{}_{u}\sigma = (S_j(u^{\sigma}), q_j(u^{\sigma}))$ com $S_j(u^{\sigma})$ representando o suporte da distribuição ponderada $q_j(u^{\sigma})$ e $j=1, ..., p$.

Se $x_i=(X_i^1, ..., X_i^p)$ onde $X_i^j=(S_j(i), q_j(i))$ $(j=1, ..., p)$ é a descrição simbólica do item i pertencente a u^{σ}, o suporte $S_j(u^{\sigma})$ associado a $q_j(u^{\sigma})$ é dado por:

$$S_j(u^{\sigma}) = \bigcup_{i \in u^{\sigma}} S_j(i)$$	**Equação 11**

Seja $m \in S_j(u^{\sigma})$ uma categoria pertencente a O_j. Então o peso $W(m) \in q_j(u^{\sigma})$ da categoria m é dado por:

$$W(m) = \frac{1}{\left|u^{\sigma}\right|} \sum_{i \in u^{\sigma}} \delta(i,m), \text{ onde}$$ $$\delta(i,m) = \begin{cases} w(m) \in q_j(i), \text{se } m \in S_j(i) \\ 0, \text{caso contrário} \end{cases}$$	**Equação 12**

em que $|u^{\sigma}|$ é a cardinalidade de u^{σ}.

Suponha que há apenas dois filmes avaliados positivamente e, portanto, fazem parte do sub-perfil u^{+}. A tabela abaixo ilustra as respectivas descrições simbólicas modais de ambos os filmes, para o atributo *Elenco*.

Variável	Filme 1	Filme 2
Elenco	*({T. Hanks, D. Morse}, {0.5, 0.5})*	*({D. Morse, B. Hunt}, {0.5, 0.5})*
	⋮	
Nota	*4*	*5*

Tabela 7 - Descrições simbólicas modais de filmes avaliados pelo usuário (apenas o atributo *Elenco*).

Observe a seguir o sub-perfil u^+ correspondente a tabela acima.

Variável	Sub-perfil u^+
Elenco	*({Tom Hanks, David Morse, Bonnie Hunt}, {0.25, 0.50, 0.25})*
	⋮

Tabela 8 - Representação do sub-perfil u^+.

4.3.3 Replicação

Uma das propriedades do domínio de sistemas de recomendação que não foi ainda prevista por nossa abordagem é a que diz respeito ao *feedback* do usuário. Lembre-se que o usuário pode efetuar, no nosso caso, avaliações no intervalo de 1 a 5. Naturalmente, um item avaliado com nota 5 é mais relevante do que um item avaliado com nota 4. Da mesma forma, um item avaliado com nota 1 é mais insignificante do que um item avaliado com nota 2.

Uma das formas de representar esse comportamento é efetuando o passo de generalização um número maior de vezes para os itens avaliados com nota 5 em relação aos itens avaliados com nota 4. De forma semelhante o passo de generalização será efetuado um número maior de vezes em um item avaliado com nota 1 em relação a um item avaliado com nota 2. Assim, estipulamos experimentalmente que os itens avaliados com nota 1 e nota 5 serão generalizados[11] 3 vezes, enquanto que os itens avaliados com nota 2 e nota 4 serão generalizados 2 vezes.

Considerando o exemplo anterior, o sub-perfil u^+ seria representado por:

Variável	Sub-perfil u^+
Elenco	*({Tom Hanks, David Morse, Bonnie Hunt}, {0.2, 0.5, 0.3})*
	⋮

[11] Dizer que um item será generalizado *n* vezes é equivalente a efetuar *n* vezes o processo descrito na seção 4.3.2 com a descrição simbólica modal deste item.

Tabela 9 - Representação do sub-perfil u^+ (usando replicação).

Uma última observação é necessária: os itens avaliados com nota 3 não foram considerados. Experimentalmente foi constatado, em nosso caso, que os itens avaliados com nota 3 refletiam dúvidas dos usuários e acabavam por confundir o Sistema de Recomendação, portanto, foram desconsiderados.

Veremos a seguir de que forma se dá o processo de recomendação através da Filtragem Baseada em Dados Simbólicos Modais. Para isso, mostraremos que a comparação entre o perfil do usuário e um item a ser recomendado decorre de uma função de similaridade apropriada, que é composta por dois componentes: o *componente de posição* e o *componente de conteúdo*.

4.4. RECOMENDANDO ITENS

Como visto anteriormente, em nosso caso, a recomendação de itens é o processo de ordená-los em uma lista segundo a relevância para o usuário. A comparação entre um item e o perfil do usuário é dada por uma função de dissimilaridade que leva em conta os sub-perfis do usuário. Mais especificamente, um item deveria ser recomendado para um usuário se esse item é similar ao sub-perfil positivo u^+ e ao mesmo tempo é dissimilar ao sub-perfil negativo u^-. A função de dissimilaridade efetua comparações ao nível de cada variável tomando as diferenças em posição e conteúdo, e feito isso, os resultados parciais são agregados em um resultado global.

Seja $x=(X^1, \ldots, X^p)$ a descrição simbólica modal de um novo item z, onde $X^j=(S_j(z), q_j(z))$, $j=1,\ldots,p$. Seja $y_u\sigma = (Y^1{}_u\sigma , \ldots , Y^p{}_u\sigma)$ a descrição simbólica modal do sub-perfil u^σ, onde $Y^j{}_u\sigma = (S_j(u^\sigma), q_j(u^\sigma))$ e $\sigma = \{+, -\}$. A comparação entre o novo item z e o perfil do usuário u é dada pela seguinte função de dissimilaridade:

$$\Phi(z,u) = \frac{(1-\phi(x, y_{u^-})) + \phi(x, y_{u^+})}{2}$$	**Equação 13**

Observe que atribuímos a mesma importância para o sub-perfil positivo e o sub-perfil negativo. Experimentalmente comprovamos que isso é conveniente.

A função $\phi(x, y_u\sigma)$ com $\sigma = \{+, -\}$ possui dois componentes: um componente livre de contexto, associado aos conjuntos $S_j(z)$ e $S_j(u^\sigma)$; e um componente dependente do contexto, associado às distribuições ponderadas $q_j(z)$ e $q_j(u^\sigma)$.

4.4.1 A Função de Dissimilaridade de Dois Componentes

A função de dissimilaridade de dois componentes ϕ é definida por:

$$\phi(x, y_{u^\sigma}) = \frac{1}{p}\sum_{j=1}^{p}[w_{cf}\phi_{cf}(S_j(z), S_j(u^\sigma)) + w_{cd}\phi_{cd}(q_j(z), q_j(u^\sigma))]$$	**Equação 14**

em que ϕ_{cf} mede a diferença em posição, quando os conjuntos $S_j(z)$ e $S_j(u^\sigma)$ são ordenados, e ϕ_{cd} mede a diferença em conteúdo entre x e $y_u\sigma$. Os pesos w_{cf} e w_{cd} expressão a importância relativa de cada componente ϕ_{cf} e ϕ_{cd} na Equação 14, e possuem as propriedades: $w_{cf} + w_{cd} = 1$, $w_{cf} > 0$ e $w_{cd} > 0$. Nesse trabalho estipulamos a mesma importância para ϕ_{cf} e ϕ_{cd} e portanto $w_{cf} = w_{cd} = 0{,}5$.

A tabela abaixo expressa os acordos (α e β) e desacordos (γ e δ) entre as distribuições ponderadas $q_j(z)$ e $q_j(u^\sigma)$.

		Item z	
Usuário u		+ (Acordos)	- (Desacordos)
	+ (Acordos)	$\alpha = \sum_{m \in S_j(z) \cap S_j(u^\sigma)} w(m)$ $\beta = \sum_{m \in S_j(z) \cap S_j(u^\sigma)} W(m)$	$\gamma = \sum_{m \in \overline{S_j(z)} \cap S_j(u^\sigma)} W(m)$
	- (Desacordos)	$\delta = \sum_{m \in S_j(z) \cap \overline{S_j(u^\sigma)}} w(m)$	

Tabela 10 - Acordos (α e β) e desacordos (γ e δ) entre as distribuições ponderadas $q_j(z)$ e $q_j(u^\sigma)$.

O componente dependente de contexto ϕ_{cd} é definido como:

$\phi_{cd}(q_j(z), q_j(u^\sigma)) = \frac{1}{2}\left(\frac{\gamma+\delta}{\alpha+\gamma+\delta} + \frac{\gamma+\delta}{\beta+\gamma+\delta}\right)$	**Equação 15**

Se o domínio O_j da variável categórica y_j for ordenado, então $m_L=min(S_j(z))$, $m_U=max(S_j(z))$, $c_L=min(S_j(u^\sigma))$ e $c_U=max(S_j(u^\sigma))$. Seja a operação de junção $S_j(z) \oplus S_j(u^\sigma)$ definida por Ichino, M. e Yaguchi, H. (1994) dada por:

$S_j(z) \oplus S_j(u^\sigma) = \begin{cases} S_j(z) \cup S_j(u^\sigma), \text{se o domínio } O_j \text{ é não ordenado.} \\ \{\min(m_L, c_L), \max(m_U, c_U)\}, \text{caso contrário.} \end{cases}$	**Equação 16**

então, o componente livre de contexto ϕ_{cf} é definido como:

$$\phi_{cf}(S_j(z), S_j(u^\sigma)) = \begin{cases} 0, \text{se } S_j(z) \cap S_j(u^\sigma) = \varnothing \\ \dfrac{\left|S_j(z) \oplus S_j(u^\sigma)\right| - \left|S_j(z)\right| - \left|S_j(u^\sigma)\right|}{\left|S_j(z) \oplus S_j(u^\sigma)\right|}, \text{caso contrário.} \end{cases}$$

Equação 17

Na próxima seção é descrito um exemplo detalhado a fim de ilustrar todo o processo de recomendação através do método proposto neste capítulo.

4.5. APLICAÇÃO DO MÉTODO: RECOMENDAÇÃO DE FILMES

Nesta seção é ilustrado um exemplo detalhado da aplicação do método de Filtragem Baseado em Dados Simbólicos Modais no domínio de recomendação de filmes. Para isso, suponha que um determinado indivíduo tenha avaliado alguns filmes conforme mostram as tabelas abaixo.

Filme	*Seven - Os Sete Crimes Capitais*	*O Exterminador Do Futuro 2*	*Pulp Fiction - Tempo De Violência*
Gênero	Suspense	Ficção Científica	Policial
País	EUA	EUA	EUA
Ano	1995	1991	1994
Diretor	David Fincher	James Cameron	Quentin Tarantino
Elenco	Gwyneth Paltrow, Morgan Freeman, Richard Roundtree, Brad Pitt, John C McGinley	Robert Patrick, Arnold Schwarzenegger, Linda Hamilton	Bruce Willis, Uma Thurman, Harvey Keitel, Christopher Walken, John Travolta, Rosanna Arquette, Eric Stoltz, Tim Roth, Amanda Plummer, Ving Rhames, Samuel L Jackson, Maria de Medeiros, Quentin Tarantino
Sinopse	Dois policiais, um jovem e impetuoso, o outro maduro e prestes a se aposentar, são encarregados de perigosa investigação: encontrar serial killer que mata pessoas seguindo a ordem dos sete pecados capitais.	Cyborg assassino é reprogramado para proteger o adolescente John Connor, futuro líder da resistência, de um humanóide perigoso enviado ao passado para liqüidá-lo. Com a ajuda do robô, John e sua mãe Sarah tentam mudar a história para impedir que a guerra nuclear aconteça.	Dois assassinos profissionais devem fazer cobrança para gângster; um deles é forçado a sair com a garota do chefe, temendo passar dos limites; enquanto isso, boxeador se mete em apuros por ganhar luta que devia perder. Homenagem à literatura pulp dos anos 40. Palma de Ouro em Cannes e o Oscar de melhor roteiro.
Nota	5	5	1

Tabela 11 - Descreve alguns filmes avaliados por um indivíduo.

Filme	*Duro de Matar 2*	*Mortal Kombat - O Filme*	*Os 12 Macacos*
Gênero	Ação	Aventura	Ficção Científica
País	EUA	EUA	EUA
Ano	1996	1995	1995
Diretor	Renny Harlin	Paul Anderson	Terry Gilliam
Elenco	Bruce Willis, Bonnie Bedelia, John Amos, Robert Constanza	Christopher Lambert, Cary-Hiroyuki Tagawa, Talisa Soto, Trevor Godard, Robin Shou, Linden Ashby, Bridgette Wilson, Kitana	Bruce Willis, Christopher Plummer, Madeleine Stowe, Brad Pitt, David Morse, Frank Gorshin, Jon Seda, Joseph Melito
Sinopse	McClane espera o vôo de sua esposa quando o aeroporto é invadido por terroristas que conseguem controlar todos os pousos e decolagens.	Três guerreiros são escolhidos a dedo e atraídos a uma ilha misteriosa onde deverão enfrentar perigosos inimigos para defender o futuro da humanidade num milenar torneio, o Mortal Kombat. Adaptação do vídeo-game homônimo.	No ano de 2035, James Cole aceita a missão de voltar ao passado para tentar decifrar mistério envolvendo vírus mortal que levou à morte da maior parte da humanidade. Tomado como louco, no passado, ele tenta provar sua sanidade a uma médica, sua única esperança de mudar o futuro.
Nota	4	1	4

Tabela 12 - Continuação da tabela anterior, contendo outros filmes.

Como descrito neste capítulo, a abordagem proposta possui dois passos para criar o perfil de um usuário. O primeiro deles é a etapa de pré-processamento, onde se obtém a descrição simbólica modal de cada item avaliado pelo usuário (seção 4.3.1). A Tabela 13 mostra o pré-processamento efetuado nos filmes *"Seven"* e *"O Exterminador do Futuro 2"*, cujas fichas técnicas são descritas na Tabela 11.

Filme	*Seven - Os Sete Crimes Capitais*	*O Exterminador Do Futuro 2*
$X_i^{gênero}$	{"Suspense"=1}	{"Ficção Científica"=1}
X_i^{pais}	{"EUA"=1}	{"EUA"=1}
X_i^{ano}	{"1995"=1}	{"1991"=1}
$X_i^{diretor}$	{"David Fincher"=1}	{"James Cameron"=1}
X_i^{elenco}	{"Brad Pitt"=0.2, "Gwyneth Paltrow"=0.2, "John C McGinley"=0.2, "Morgan Freeman"=0.2, "Richard Roundtree"=0.2}	{"Arnold Schwarzenegger"=0.3333333333, "Linda Hamilton"=0.3333333333, "Robert Patrick"=0.3333333333}
$X_i^{sinopse}$	{"Dois"=0.0000005283, "aposentar"=0.000000919, "capitais"=0.0000012981, "encarregados"=0.0000010305, "encontrar"=0.0000005577, "impetuoso"=0.0000010791, "investigação"=0.0000008277, "jovem"=0.0000004242, "killer"=0.000000881, "maduro"=0.0000012218, "mata"=0.000000697, "ordem"=0.0000008937, "pecados"=0.0000011946, "perigosa"=0.0000006567, "pessoas"=0.0000005716, "policiais"=0.0000006554, "prestes"=0.0000006802, "seguindo"=0.0000010578, "serial"=0.000000881}	{"Connor"=0.0000007827, "Cyborg"=0.0000008005, "John"=0.0000008782, "Sarah"=0.0000006159, "aconteça"=0.000000707, "adolescente"=0.0000004526, "ajuda"=0.0000002993, "assassino"=0.0000003799, "enviado"=0.0000005114, "futuro"=0.0000004095, "guerra"=0.0000003613, "história"=0.0000003194, "humanóide"=0.0000008223, "impedir"=0.0000004334, "liqüidá"=0.0000008005, "lo"=0.0000002936, "líder"=0.0000004348, "mudar"=0.000000462, "mãe"=0.0000003537, "nuclear"=0.0000004981, "passado"=0.0000003996, "perigoso"=0.0000004289, "proteger"=0.0000005064, "reprogramado"=0.0000008902, "resistência"=0.0000005792, "robô"=0.0000005659, "tentam"=0.0000004091}
Nota	5	5

Tabela 13 - Ilustra as descrições simbólicas modais dos filmes "Seven – Os Sete Crimes Capitais" e "O Exterminador do Futuro 2".

Observe que o formato em que as variáveis modais são mostradas é ligeiramente diferente daquele apresentado na seção 4.3.1. No entanto, não é difícil identificar o suporte $S_j(i)$ e a distribuição $q_j(i)$ em qualquer das variáveis. Por exemplo, $S_{elenco}(Seven)=\{Gwyneth\ Paltrow, Morgan\ Freeman, Richard\ Roundtree, Brad\ Pitt, John\ C\ McGinley\}$ e a distribuição de pesos associada a tal suporte no filme *Seven* para a variável *elenco* é $q_{elenco}(Seven)=\{0.2, 0.2, 0.2, 0.2, 0.2\}$.

Adicionalmente, uma outra observação se faz necessária. O atributo *sinopse* possui distribuições com valores relativamente baixos. Isso ocorre devido ao fator IDF, que representa a importância da palavra na base como um todo (seção 4.3.1).

Para continuar com o exemplo, suponha que todos os outros itens avaliados tenham sido pré-processados. Assim, foram geradas as descrições simbólicas modais de todos os itens avaliados pelo usuário, a partir das quais será gerado o perfil do usuário. Para isso, estas descrições devem ser generalizadas e replicadas em um dos sub-perfis do usuário. Segundo descrito nas seções 4.3.2 e 4.3.3, um item avaliado com nota 1 deve ser generalizado três vezes no sub-perfil negativo, um item avaliado com nota 2 deve ser generalizado duas vezes no sub-perfil negativo, um item avaliado com nota 4 deve ser generalizado duas vezes no sub-perfil positivo e um item avaliado com nota 5 deve ser generalizado três vezes no sub-perfil positivo. Dessa forma, após o procedimento de generalização obtém-se o perfil do usuário mostrado na tabela abaixo.

	Sub-perfil positivo – u+	*Sub-perfil negativo – u⁻*
$X^{gênero}$	{"Ação"=2, "Ficção Científica"=5, "Suspense"=3}	{"Aventura"=3, "Policial"=3}
$X^{país}$	{"EUA"=10}	{"EUA"=6}
X^{ano}	{"1991"=3, "1995"=5}	{"1994"=3, "1995"=3}
$X^{diretor}$	{"David Fincher"=3, "James Cameron"=3, "Renny Harlin"=2, "Terry Gilliam"=2}	{"Paul Anderson"=3, "Quentin Tarantino"=3}

X^{elenco}	{"Arnold Schwarzenegger"=1, "Bonnie Bedelia"=0.5, "Brad Pitt"=0.85, "Bruce Willis"=0.75, "Christopher Plummer"=0.25, "David Morse"=0.25, "Frank Gorshin"=0.25, "Gwyneth Paltrow"=0.6, "John Amos"=0.5, "John C McGinley"=0.6, "Jon Seda"=0.25, "Joseph Melito"=0.25, "Linda Hamilton"=1, "Madeleine Stowe"=0.25, "Morgan Freeman"=0.6, "Richard Roundtree"=0.6, "Robert Constanza"=0.5, "Robert Patrick"=1}	{"Amanda Plummer"=0.2307692308, "Bridgette Wilson"=0.375, "Bruce Willis"=0.2307692308, "Cary-Hiroyuki Tagawa"=0.375, "Christopher Lambert"=0.375, "Christopher Walken"=0.2307692308, "Eric Stoltz"=0.2307692308, "Harvey Keitel"=0.2307692308, "John Travolta"=0.2307692308, "Kitana"=0.375, "Linden Ashby"=0.375, "Maria de Medeiros"=0.2307692308, "Quentin Tarantino"=0.2307692308, "Robin Shou"=0.375, "Rosanna Arquette"=0.2307692308, "Samuel L Jackson"=0.2307692308, "Talisa Soto"=0.375, "Tim Roth"=0.2307692308, "Trevor Godard"=0.375, "Uma Thurman"=0.2307692308, "Ving Rhames"=0.2307692308}
$X^{sinopse}$	{"Cole"=0.000001486, "Connor"=0.000002348, "Cyborg"=0.0000024015, "Dois"=0.0000015848, "James"=0.0000009743, "John"=0.0000026346, "McClane"=0.0000044844, "Sarah"=0.0000018476, "Tomado"=0.000001916, "aceita"=0.0000008944, "aconteça"=0.0000021211, "adolescente"=0.0000013579, "aeroporto"=0.0000035057, "ajuda"=0.0000008979, "aposentar"=0.0000027569, "assassino"=0.0000011398, "capitais"=0.0000038943, "conseguem"=0.0000028898, "controlar"=0.0000032319, "decifrar"=0.0000013202, "decolagens"=0.0000050512, "encarregados"=0.0000030916, "encontrar"=0.000001673, "enviado"=0.0000015342, "envolvendo"=0.0000008782, "espera"=0.0000026078, "esperança"=0.0000010659, "esposa"=0.0000016964, "futuro"=0.0000020475, "guerra"=0.000001084, "história"=0.0000009583, "humanidade"=0.000001139, "humanóide"=0.000002467, "impedir"=0.0000013003, "impetuoso"=0.0000032374, "invadido"=0.0000039176, "investigação"=0.0000024832, "jovem"=0.0000012725, "killer"=0.000002643, "levou"=0.0000012147, "liqüidá"=0.0000024015, "lo"=0.0000008809, "louco"=0.0000010513, "líder"=0.0000013044, "maduro"=0.0000036655, "maior"=0.0000008882, "mata"=0.0000020909, "missão"=0.0000007058, "mistério"=0.0000009408, "mortal"=0.0000010443, "morte"=0.0000005995, "mudar"=0.0000023102, "mãe"=0.0000010612, "médica"=0.0000011991, "nuclear"=0.0000014942, "ordem"=0.0000026811, "parte"=0.0000007593, "passado"=0.0000027971, "pecados"=0.0000035838, "perigosa"=0.0000019701, "perigoso"=0.0000012868, "pessoas"=0.0000017148, "policiais"=0.0000019661, "pousos"=0.0000050512, "prestes"=0.0000020406, "proteger"=0.0000015193, "provar"=0.0000009278, "reprogramado"=0.0000026705, "resistência"=0.0000017376, "robô"=0.0000016977, "sanidade"=0.0000013615, "seguindo"=0.0000031733, "serial"=0.000002643, "tenta"=0.0000006414, "tentam"=0.0000012274, "tentar"=0.0000007872, "terroristas"=0.0000026754, "voltar"=0.0000008882, "vírus"=0.000001139, "vôo"=0.0000031746}	{"Adaptação"=0.0000017431, "Cannes"=0.0000013832, "Dois"=0.0000009713, "Homenagem"=0.0000020852, "Kombat"=0.0000029035, "Mortal"=0.000002647, "Oscar"=0.000001088, "Ouro"=0.0000013562, "Palma"=0.0000016494, "Três"=0.0000014504, "anos"=0.000000696, "apuros"=0.0000014268, "assassinos"=0.0000012932, "atraídos"=0.0000029035, "boxeador"=0.0000018659, "chefe"=0.000001235, "cobrança"=0.0000023079, "dedo"=0.0000028541, "defender"=0.0000018943, "devem"=0.0000014622, "deverão"=0.0000029035, "devia"=0.0000021542, "enfrentar"=0.0000014361, "escolhidos"=0.0000027715, "fazer"=0.0000010026, "forçado"=0.0000012309, "futuro"=0.0000001549, "game"=0.0000028105, "ganhar"=0.0000013541, "garota"=0.0000010945, "guerreiros"=0.0000020461, "gângster"=0.0000013541, "homônimo"=0.0000023186, "humanidade"=0.0000021543, "ilha"=0.0000016573, "inimigos"=0.0000019048, "limites"=0.0000018159, "literatura"=0.0000017835, "luta"=0.0000009899, "mete"=0.0000013904, "milenar"=0.0000025751, "misteriosa"=0.0000017869, "num"=0.0000011074, "passar"=0.0000011564, "perder"=0.0000013416, "perigosos"=0.0000019708, "profissionais"=0.0000016688, "pulp"=0.0000026885, "roteiro"=0.0000014003, "sair"=0.0000013499, "temendo"=0.0000020852, "torneio"=0.0000023649, "vídeo"=0.0000020358}

Tabela 14 - Ilustra o perfil de um indivíduo, formado a partir dos itens avaliados pelo usuário que são mostrados na Tabela 11 e na Tabela 12.

As distribuições do perfil apresentado na Tabela 14 não estão normalizadas. De fato, a normalização pode ser efetuada na etapa de

comparação, quando se calcula a similaridade em posição e em conteúdo entre o perfil do usuário e a descrição simbólica de um item (seção 4.4.1). Essa é uma otimização em nível de implementação, a fim de que novos itens sejam adicionados ao perfil na medida que o usuário efetue novas avaliações. Dessa forma, o custo para construção e manutenção do perfil é insignificante, diferentemente de alguns métodos de filtragem por conteúdo que possuem formas próprias de armazenar o perfil do usuário.

A fim de ilustrar o processo de generalização sucedido na Tabela 14, considere o termo *"Ficção Científica"* $\in O_{gênero}$. Verifique que *"Ficção Científica"* $\notin S_{gênero}(u^-)$. No entanto, *"Ficção Científica"* $\in S_{gênero}(u^+)$ e seu peso é cinco. Isto ocorre porque o usuário não avaliou negativamente (notas 1 ou 2) nenhum filme. Entretanto, os filmes *"O Exterminador Do Futuro 2"* e *"Os 12 Macacos"* foram avaliados positivamente com as notas 5 e 4, respectivamente. Significa que o termo deve ser replicado três vezes em $S_{gênero}(u^+)$ devido ao primeiro filme e duas vezes em função do filme *"Os 12 Macacos"*, resultando em um total de cinco.

Existem diversos exemplos semelhantes ao descrito anteriormente e que pode melhorar a compreensão do método de Filtragem Baseado em Dados Simbólicos Modais. Todavia, passemos a próxima etapa do processo: a recomendação de itens. Para tanto, considere os seguintes itens da tabela abaixo que deverão ser comparados com o perfil do usuário a fim de se definir uma lista de sugestões, ordenada segundo sua relevância para o usuário.

Filme	*Alien 3*	*O Silêncio Dos Inocentes*	*O Juiz*
Gênero	Ficção Científica	Suspense	Ficção Científica
País	EUA	EUA	EUA
Ano	1992	1991	1995
Diretor	David Fincher	Jonathan Demme	Danny Cannon
Elenco	Sigourney Weaver, Charles S Dutton, Charles Dance, Lance Henriksen, Pete Postlethwaite, Paul McGann, Ralph Brown, Brian Glover, Danny Webb, Christopher John Fields, Holt McCallany	Scott Glen, Anthony Hopkins, Jodie Foster, Ted Levine	Armand Assante, James Russo, Sylvester Stallone, Rob Schneider, Diane Lane, Max Von Sydow, Joanna Miles, Joan Chen, Mitchell Ryan, Scott Wilson, Jürgen Prochnow, Balthazar Getty

Sinopse	A nave da tenente Ripley cai em planeta-prisão. Ela tem a confirmação de que o monstro ainda estava a bordo quando corpos de prisioneiros começam a aparecer mutilados. E é forçada, mais uma vez, a combater a terrível criatura espacial.	Agente do FBI é destacada para encontrar assassino que tira a pele de suas vítimas. Para entender como ele pensa, ela procura um perigoso psicopata, encarcerado sob a acusação de canibalismo. Terceiro filme a receber os cinco principais Oscars: filme, direção, ator, atriz e roteiro. Do best-seller de Thomas Harris. Lançado anteriormente com o selo LK-Tel.	No ano de 2139, superjuízes decidem os litígios e executam as sentenças. Um dos maiores defensores da lei, juiz Dredd, é condenado por crime que não cometeu e banido da cidade. E descobre que tudo não passa de um plano de Rico, clone mutante condenado por Dredd que busca vingança.
Score	0.61666	0.60344	0.56897

Tabela 15 - Lista de sugestão de filmes para o usuário cujo perfil corresponde ao apresentado na Tabela 14.

Filme	*Instinto Selvagem*	*Uma Linda Mulher*	*True Lies*
Gênero	Suspense	Romance	Aventura
País	EUA	EUA	EUA
Ano	1992	1990	1994
Diretor	Paul Verhoeven	Garry Marshall	James Cameron
Elenco	Michael Douglas, Sharon Stone, George Dzundza, Dorothy Malone, Bruce A Young, Wayne Knight, Stephen Tobolowsky, Leilani Sarelle, Chelcie Ross, Benjamin Mouton, Jeanne Tripplehorn, Daniel Von Bargen, Denis Arndt	Jason Alexander, Julia Roberts, Elinor Donahue, Richard Gere, Hector Elizondo, Laura San Giacomo, Alex Hyde-White, Ralph Bellamy, Larry Miller, Jane Morris	Jamie Lee Curtis, Bill Paxton, Tom Arnold, Tia Carrere, Charlton Heston, Grant Heslov, Arnold Schwarzenegger, Art Malik, Eliza Dushku
Sinopse	Nick Curran, detetive da polícia de São Francisco, investiga assassinato de astro do rock. As pistas levam à namorada do cantor, Catherine Tramell, uma mulher rica, sexy, sedutora e aparentemente fatal. Com seu jeito insinuante, ela seduz Nick que, obcecado, começa a perder o controle da situação.	Uma bela garota de programa conhece por acaso rapaz milionário. Ele a contrata como acompanhante por algumas noites e ela se apaixona por ele e pelo luxo que oferece. Como uma cinderela moderna, ela vive um conto de fadas, com intrigas e paixões.	Superagente secreto assume para sua mulher a figura de um pacato e entediante vendedor. Cansada da monotonia, ela acaba envolvida em movimentada trama e ele é forçado a dasvendar sua verdadeira identidade. Refilmagem de ' La Totale! ' de Caude Zidi.
Score	0.55416	0.49999	0.44490

Tabela 16 - Continuação da lista de sugestão mostrada na tabela anterior.

De uma forma geral, é possível verificar através de uma observação sucinta o porquê da ordem obtida pelo algoritmo. Assim, nota-se que o filme *"Alien 3"*, que é o primeiro da lista, possui um *score* relativamente alto visto que possui algumas semelhanças com o sub-perfil positivo do perfil do usuário, além de raramente possuir intersecções com o sub-perfil negativo. Pode-se observar também que os filmes *"O Silêncio dos Inocentes"* (segundo da lista), *"O Juiz"* (terceiro da lista) e *"Instinto Selvagem"* (quarto da lista) possuem classificações menores do que *"Alien 3"* sobretudo por não haver intersecção entre estes filmes e o sub-perfil positivo na variável *diretor*.

O mesmo raciocínio pode ser usado para entender o porquê dos *scores* relativamente baixos para os filmes *"Uma Linda Mulher"* e *"True Lies"*. No caso do filme *"Uma Linda Mulher"*, os atributos *gênero* e *diretor* não possuem qualquer contribuição positiva ou negativa para definição de sua relevância para o usuário. Já para o filme *"True Lies"*, apesar da variável *diretor* contribuir positivamente no *score*, as variáveis *gênero* e *ano* contribuem negativamente, devido suas semelhanças com o sub-perfil negativo do usuário.

O exemplo descrito nesta seção consiste em um caso real de uso do método proposto neste capítulo. Portanto, o objetivo deste exemplo é tornar mais claro o processo de filtragem com base nas teorias de Análise de Dados Simbólicos. Na próxima seção são expostas algumas conclusões do capítulo.

4.6. CONCLUSÕES

O fato da abordagem proposta ter como premissa a manipulação de dados simbólicos já a torna vantajosa, tendo em vista que atributos complexos, como autores podem ser manipulados pelos algoritmos. Como visto nos trabalhos relacionados, essa é uma deficiência eminente em alguns métodos de filtragem baseada em conteúdo.

Uma das propriedades da abordagem é o fato de modelar características próprias de sistemas de recomendação na representação utilizada para o perfil do usuário. Um exemplo disso é o uso de dois sub-perfis para representar o perfil do usuário, onde um deles congrega as informações dos itens avaliados negativamente, enquanto um outro sumariza as informações dos itens avaliados positivamente. Um segundo exemplo é a técnica de replicações de itens no sub-perfil conveniente de acordo com sua importância para o usuário (um item avaliado com nota 5 é mais importante do que um item avaliado com nota 4).

Uma das principais vantagens da abordagem proposta neste trabalho é o bom desempenho no que se refere ao tempo de aprendizagem e ao tempo de classificação, como será mostrado no próximo capítulo. Adicionalmente, o uso de memória também é significativamente menor, se

comparado a sistemas baseados em outros métodos de filtragem por conteúdo. Não é difícil compreender o motivo disto. De uma forma geral, ao nível de cada atributo os valores dos itens avaliados se repetem. Por exemplo, é natural que vários dos filmes avaliados por um indivíduo sejam do mesmo gênero, possuam atores recorrentes, ou mesmo possuam em sua sinopse palavras repetidas. Nesse sentido, ao sumarizar os itens avaliados pelo usuário em estruturas como as variáveis simbólicas modais, os dados que constituem o perfil do usuário acabam sendo condensados, sem que seja perdida qualquer informação.

Como mencionado na seção 4.5, uma outra característica da abordagem baseada em dados simbólicos modais é que se trata de um método de aprendizagem incremental, ou seja, novos itens podem ser adicionados ao perfil do usuário sem refazer a construção dos sub-perfis.

No próximo capítulo, é mostrado que o método de Filtragem Baseado em Dados Simbólicos Modais é estatisticamente equivalente, no que diz respeito à predição, a um dos métodos (kNN) de filtragem por conteúdo que constadamente possui um dos melhores desempenhos nesse mesmo critério.

Capítulo 5

ANÁLISE EXPERIMENTAL

5.1. INTRODUÇÃO

No capítulo anterior foi apresentada uma abordagem simbólica para filtragem de informação baseada em conteúdo. Ou seja, teorias do domínio de Análise de Dados Simbólicos (seção 4.2) foram adaptadas dando origem a um novo método de filtragem de informação. Com isso, torna-se possível aplicar este método para solucionar problemas reais como aqueles existentes nos Sistemas de Recomendação Personalizados (seção 2.3).

Baseado nisso, mostra-se conveniente avaliar o método proposto neste trabalho através de comparações com uma outra abordagem de filtragem por conteúdo no domínio de recomendações de filmes. Nesse sentido, a filtragem por conteúdo baseada no algoritmo kNN, por ser bastante popular e ter apresentado bons resultados na literatura (Arya 1995, Cotter e Smyth 2000, Krukwich e Burkey 1996, Balanovic e Shoham 1997), é escolhido para servir de referência nos experimentos realizados neste trabalho. Assim, na seção 5.2 é descrito brevemente o processo de filtragem de informação através do kNN (para os trabalhos relacionados consulte a seção 3.3.2). Após isso, é introduzido na seção 5.3 o ambiente experimental que foi estipulado com base em trabalhos relacionados.

A forma como se darão os experimentos é definida detalhadamente na seção 5.4, quando é descrita a metodologia experimental. Finalmente, na seção 5.5 são apresentados os resultados e é realizada uma análise minuciosa dos mesmos.

5.2. FILTRAGEM COM *K VIZINHOS MAIS PRÓXIMOS*

No kNN e outros algoritmos de *aprendizagem baseada em instância* os exemplos são instâncias originais do conjunto de treinamento (itens do perfil do usuário). Durante a aprendizagem, esses algoritmos usam uma função de distância para determinar quão próximo um novo vetor de entrada y está a cada instância da memória, e utiliza as

instâncias mais próximas para inferir a classe de saída de *y*. O algoritmo kNN clássico é mostrado logo abaixo.

Fase de Treinamento **Para cada item** i **avaliado pelo usuário** u **(score** ω_{Ui}**), adicione o exemplo de treinamento** (i,ω_{Ui}) **ao perfil do usuário** P_u**.**
Fase de Classificação **Dada uma nova instância** j **no conjunto de exemplos a serem classificados (conjunto de testes** Q**):** **Seja** $i_0, \ldots, i_k$ **as k instâncias do perfil do usuário que são mais similares (próximas) a** j**, o score dessa instância com relação a** P **é dada pela Equação 18**

Tabela 17 – Descreve o algoritmo clássico dos *k vizinhos mais próximos*.

$$\omega_{U_j} = \frac{\sum_{i=0}^{k} \omega_{U_i} * s_{i,j}}{\sum_{i=0}^{k} s_{i,j}}$$	**Equação 18**

em que $s_{i,j}$ é dado pela Equação 19.

$$s_{\alpha,\beta} = \frac{\sum_{m=0}^{r} \lambda_m * \Gamma_m(\alpha_m, \beta_m)}{\sum_{m=0}^{r} \lambda_m}$$	**Equação 19**

em que λ_i é o peso do atributo *m*, e $\Gamma_m(\alpha_m,\beta_m)$ define a similaridade entre os elementos α e β com relação ao atributo *m*. Neste trabalho consideramos o peso do atributo equivalente para todos os descritores. Dessa forma, a Equação 19 é simplificada na Equação 20.

$$s_{\alpha,\beta} = \frac{\sum_{m=0}^{r} \Gamma_m(\alpha_m, \beta_m)}{r}$$	**Equação 20**

A função Γ depende do tipo de atributo considerado. A fim de sermos coerentes na avaliação dos métodos de filtragem, a forma de comparação ao nível de cada atributo se dará da mesma maneira daquela utilizada na Filtragem Baseada em Dados Simbólicos Modais.

Um problema da abordagem kNN é que o cálculo do termo $s_{i,j}$, a *função de similaridade* (Equação 19), é muito caro, especialmente se considerarmos um cenário de um sistema WEB em que podem ocorrer milhões de usuários e milhares de itens a serem classificados. Além disso, todas as instâncias no perfil do usuário devem ser comparadas com cada item de repositório para confirmar a relevância dos mesmos para esse usuário. Suponha que se queira determinar os melhores itens a partir de um conjunto de perguntas de tamanho **S** para um dado usuário com um conjunto de treinamento (perfil) de tamanho **M**. Nesse cenário, o sistema deverá processar a função *Γ*, **S*M*R** vezes, onde **R** é o número de atributos de uma instância. Considerando um sistema WEB acessado por milhões de usuários ao mesmo tempo, esse número passa a ser **N*S*M*R**, onde **N** é o número de usuários.

5.3. AMBIENTE EXPERIMENTAL

Como mencionado na introdução do capítulo, o ambiente experimental é definido em função de experimentos realizados em trabalhos afins.

O *EachMovie* consistia em um serviço de recomendação de filmes que foi desativado em setembro de 1997, funcionando durante 18 meses como parte de um projeto do antigo Centro de Sistemas e Pesquisas DEC (Digital Equipment Corporation) que fora incorporado ao Centro de Pesquisas e Sistemas Compaq. Durante os 18 meses de funcionamento, 72.916 usuário efetuaram 2.811.983 avaliações para 1.628 filmes distintos. Esta base de dados está disponível para fins não comerciais e pode ser obtida através da Internet facilmente. Dessa forma, ela se torna adequada para os experimentos definidos com o propósito de se comparar

a abordagem baseada em dados simbólicos modais com a abordagem baseada no kNN.

Todavia, o banco de dados original do *EachMovie* não contém o conteúdo descritivo de qualquer atributo considerado anteriormente (*gênero*, *diretor*, *elenco*, *país*, *ano* e *sinopse*) e por essa razão não seria possível experimentar a filtragem baseada em conteúdo. Logo, a tabela de filmes originais foi combinada com uma segunda tabela contendo as descrições dos filmes no idioma Português para os seis atributos supracitados.

Para concluir, o ambiente experimental é baseado em um subconjunto do banco de dados *EachMovie* que consiste em 22.867 usuários e 1.572.965 avaliações no intervalo de 1 a 5 para 638 filmes contendo as descrições completas em Português para os atributos *gênero, país, ano, elenco, diretor* e *sinopse* (veja a Tabela 5).

5.4. METODOLOGIA EXPERIMENTAL

A metodologia experimental deve responder a três questões principais:

- O que será avaliado?
- Como medir o que será avaliado?
- Como comparar os resultados dos sistemas avaliados?

Com relação a primeira questão foi verificado na seção 2.2 que o que se pretende avaliar em um sistema de recomendação depende efetivamente do tipo de tarefa ou objetivo desse sistema. Neste trabalho definimos que o objetivo principal de um sistema de recomendação é gerar uma lista de itens ordenados segundo sua relevância para o usuário (tarefa vi) descrita na seção 2.2. Nesse caso o usuário deseja navegar pela lista de itens para observar sua ordem de relevância. Por exemplo, quando o usuário busca o melhor item dentre as escolhas, ele busca esse tipo de recomendação.

Depois de definido o objetivo do sistema a ser avaliado é necessário escolher-se uma métrica apropriada para medir a qualidade do

sistema, ou seja, a segunda questão colocada anteriormente. De acordo com as métricas relatadas na seção 3.4, têm-se que a métrica *Breese* (seção 3.4.4) mostra-se mais adequada para o objetivo do sistema de recomendação a ser avaliado e, portanto, será utilizada para medir a qualidade das recomendações dos sistemas.

Adicionalmente, pretende-se avaliar neste trabalho o tempo de resposta de classificação e o espaço de memória utilizado. O tempo de classificação é medido em milisegundos em função do tempo gasto para geração de uma lista de recomendação. O espaço de memória é medido em função de quantos bytes se gasta para o armazenamento do perfil de um usuário.

Definidos os critérios a serem avaliados (velocidade, memória e qualidade) e as medidas utilizadas para cada um dos critérios (milisegundos, bytes e métrica *Breese*), passa-se ao processo experimental propriamente, cujo algoritmo é apresentado em pseudo-código na Tabela 18. Antes de observar o algoritmo é conveniente descrever brevemente o processo que se deseja realizar.

Foram selecionados aleatoriamente 50 usuários do conjunto inicial de forma que todos os escolhidos possuam no mínimo 300 avaliações. Do conjunto de filmes avaliados, foram selecionados aleatoriamente, ao nível de cada usuário, subconjuntos de cardinal $m \in \{20, 40, 60, 80, 100\}$ para constituição do perfil do usuário; e outro subconjunto disjunto dos anteriores e contendo 200 filmes com propósito de teste. Para cardinais fixos dos conjuntos de treinamento e de teste, este processo foi repetido 30 vezes para cada usuário dos 50 selecionados. Em cada uma das 30 vezes o sistema foi executado nas mesmas condições para geração de listas de recomendação com base nos conjuntos de testes. Neste momento, eram calculados a velocidade, o espaço de memória e a qualidade da recomendação. Ao final das 30 iterações calculava-se a média de cada uma dessas medidas. Finalmente, eram calculados as médias e os desvios padrões de cada medida para cada cardinal m do conjunto de treinamento com base nas médias obtidas das 30 iterações para os 50 usuários.

Entrada

$U=\{u_1,u_2,...,u_{50}\}$: o conjunto dos 50 usuários aleatoriamente selecionados onde o número de filmes avaliados por cada usuário é maior ou igual a 300

Algoritmo

- **Para cada valor de m em {20, 40, 60, 80, 100} faça:**
  - **Para cada usuário u_i do conjunto de usuários U faça:**
    - **Repita 30 vezes:**
      - **Faça $\rho_i \leftarrow \emptyset$, onde ρ_i é o perfil do i-ésimo usuário**
      - **Escolha m filmes aleatoriamente e insira-os em ρ_i**
      - **Faça Q o conjunto de testes contendo 200 filmes escolhidos aleatoriamente onde $Q \cap \rho_i = \emptyset$**
      - **Execute os sistemas para gerar as listas de recomendações**
        - **Calcule a qualidade $P_{MP,i}$, a velocidade $T_{MP,i}$ e o espaço de memória utilizado $S_{MP,i}$ para a abordagem proposta**
        - **Calcule a qualidade $P_{kNN,i}$, a velocidade $T_{kNN,i}$ e o espaço de memória utilizado $S_{kNN,i}$ para a abordagem kNN**
    - **$P_{MP}[m,u_i] \leftarrow \Sigma_i P_{MP,i} / 30$**
    - **$T_{MP}[m,u_i] \leftarrow \Sigma_i T_{MP,i} / 30$**
    - **$S_{MP}[m,u_i] \leftarrow \Sigma_i S_{MP,i} / 30$**
    - **$P_{kNN}[m,u_i] \leftarrow \Sigma_i P_{kNN,i} / 30$**
    - **$T_{kNN}[m,u_i] \leftarrow \Sigma_i T_{kNN,i} / 30$**
    - **$S_{kNN}[m,u_i] \leftarrow \Sigma_i S_{kNN,i} / 30$**
  - **Calcule as médias $P_{MP,media}[m]$, $T_{MP,media}[m]$ e $S_{MP,media}[m]$ e os desvios padrões $P_{MP,desvio}[m]$, $T_{MP,desvio}[m]$ e $S_{MP,desvio}[m]$ em função dos vetores $P_{MP}[m,u_i]$, $T_{MP}[m,u_i]$ e $S_{MP}[m,u_i]$, respectivamente, onde $i \in \{1,...,50\}$**
  - **De forma semelhante ao passo anterior, calcule as respectivas médias e desvios padrões para o método kNN**
  - **Calcular o teste de hipótese para m**

Tabela 18 - Descreve o algoritmo utilizado para realização dos experimentos.

A última linha do algoritmo mostrado na Tabela 18 refere-se à terceira questão colocada anteriormente, ou seja: "como se dará a comparação entre os métodos de filtragem?".

A comparação entre sistemas de filtragem de informação pode depender de cada domínio, mas em geral pode-se usar testes de hipótese para prover maior confiança nos resultados obtidos.

O teste de hipótese, basicamente, define duas hipóteses: a *hipótese de trabalho* (ou nula) que normalmente diz que não há diferenças entre dois sistemas segundo o critério de avaliação escolhido; e a *hipótese alternativa* que diz que existe diferença entre os dois sistemas também

para um mesmo critério de avaliação previamente definido. Alguns testes de hipótese admitem outras hipóteses alternativas, que poderia ser, por exemplo, um sistema é melhor do que o outro. Então o objetivo desses testes é medir a evidência que existe em favor de se rejeitar a hipótese de trabalho.

A metodologia indicada para a execução de testes de hipótese é o uso de testes pareados, ou seja, os dois sistemas são executados nas mesmas condições (mesma plataforma, mesmo conjunto de dados, mesmo conjunto para treinar, mesmo conjunto de teste para avaliar, etc). Nesse caso usa-se o teste *t-student* pareado[12].

Caso tenha sido usado um esquema de *validação cruzada k-fold* em um sistema e *j-fold* em outro sistema, o cálculo muda ligeiramente. Em casos onde não há independência dos conjuntos de treinamento sucessivos, ou seja, não se pode assumir nada sobre a distribuição dos dados, os testes mais indicados são *Wilcoxon matched-pairs signed-ranks test*. Maiores informações sobre testes de hipótese podem ser obtidas em Witten e Frank 2000.

Para resumir, neste trabalho cada usuário representa um problema em particular e o objetivo é medir o desempenho global do sistema. Para isso, repete-se o experimento para cada usuário durante 30 vezes e, no final, são tiradas as médias das métricas de interesse para cada usuário. Por fim, são obtidas as métricas de desempenho globais do sistema através das médias calculadas a partir das médias dos 50 usuários. Além disso, são calculados os desvios padrões para permitir o cálculo dos testes de hipóteses. Na próxima seção são apresentados os resultados destes experimentos bem como realizadas algumas análises.

5.5. APRESENTAÇÃO E ANÁLISE DOS RESULTADOS

Conforme descrito na seção anterior, nosso objetivo é comparar a qualidade das recomendações, a velocidade e o espaço de memória do novo método de filtragem descrito no capítulo anterior com o kNN clássico introduzido na seção 5.2. Experimentos preliminares mostraram adequado

[12] De fato, veremos na próxima seção que este é o teste de hipótese mais adequado em nosso caso.

o valor 11 para a variável k. Assim, os experimentos a seguir utilizam como referência este resultado para definição do valor de k.

Tendo em mente o ambiente e a metodologia supracitados, considere μ_i, σ_i, X_i e s_i, respectivamente, a média, o desvio padrão, a média amostral e o desvio padrão amostral das variáveis aleatórias X_1 até X_6 descritas na Tabela 19.

X_1	**Qualidade média das recomendações segundo Breese obtida a partir da abordagem simbólica**
X_2	**Qualidade média das recomendações segundo Breese obtida a partir da filtragem com o kNN padrão, sendo k=11**
X_3	**Tempo médio em milisegundos gasto na geração das listas através da abordagem simbólica**
X_4	**Tempo médio em milisegundos gasto na geração das listas através da filtragem com o kNN padrão, sendo k=11**
X_5	**Espaço médio em bytes utilizado pelo perfil do usuário na geração das listas através da abordagem simbólica**
X_6	**Espaço médio em bytes utilizado pelo perfil do usuário na geração das listas através da filtragem com o kNN padrão, sendo k=11**

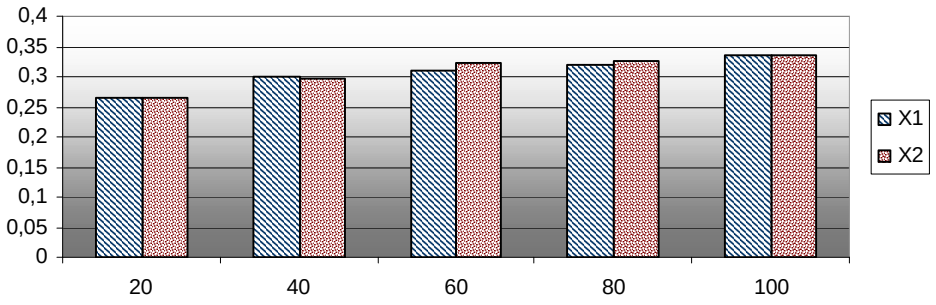
Tabela 19 - Descreve as variáveis aleatórias consideradas nos testes de hipóteses dos experimentos realizados.

Visto que os sistemas a serem comparados são executados nas mesmas condições[13], os testes de hipóteses mais adequados são os testes pareados cujas variáveis estatísticas possuem uma distribuição *t-Student* com 98 graus de liberdade. Dessa forma, nosso objetivo é testar, aos níveis de significância de 1% e 5%, as hipóteses nulas H_0: $\mu_1=\mu_2$, H_0': $\mu_3=\mu_4$ e H_0": $\mu_5=\mu_6$ contra, respectivamente, as hipóteses alternativas H_1: $\mu_1>\mu_2$, H_1': $\mu_3<\mu_4$ e H_1": $\mu_5<\mu_6$.

Os gráficos a seguir mostram os resultados destes experimentos. Neles podem ser observados os valores médios nos dois tipos de sistemas analisados (Simbólica Modal versus kNN, com k=11), variando o tamanho do conjunto de treinamento (20, 40, 60, 80 e 100), para os três critérios definidos na metodologia experimental, sendo a qualidade das recomendações ilustrada no Gráfico 1, a velocidade ilustrada no Gráfico 2

[13] Para execução dos experimentos foi utilizado um computador PC com processador AMD-Duron de 1 Ghz, 512 MB de memória, sistema operacional Linux (versão RedHat) e ambiente de execução Java 2 Runtime Environment versão 1.4.1.

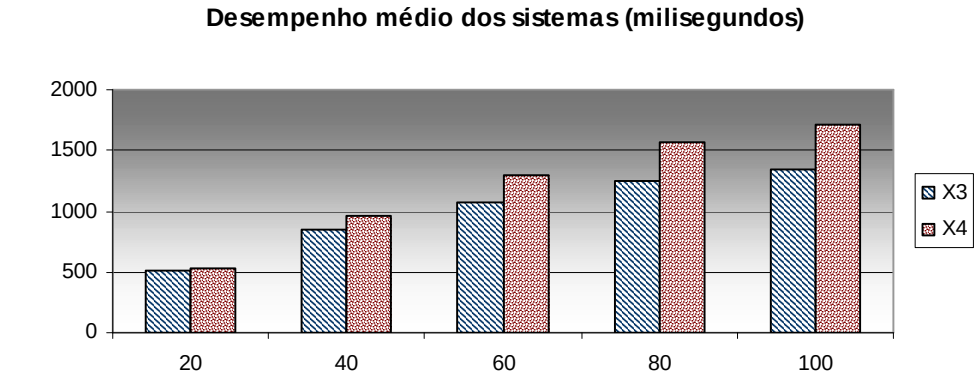
e a memória ilustrada no Gráfico 3. São também apresentados os desvios padrões (s_1, ..., s_6) associados a cada configuração *sistema* x *critério* x *tamanho do conjunto de treinamento*. Por fim, são observados os valores para as variáveis estatísticas Z1, Z2 e Z3 associadas às hipóteses nulas H_0, H_0' e H_0", respectivamente.



	20	40	60	80	100
X_1	0,2659	0,2987	0,3097	0,3192	0,3356
s_1	0,1493	0,1609	0,1590	0,1680	0,1717
X_2	0,2651	0,2952	0,3210	0,3271	0,3362
s_2	0,1553	0,1531	0,1568	0,1724	0,1691
Z_1	0,0269	0,1116	-0,3575	-0,2332	-0,0155
Resultado (1%)	Aceito	Aceito	Aceito	Aceito	Aceito
Resultado (5%)	Aceito	Aceito	Aceito	Aceito	Aceito

Gráfico 1 - Este gráfico ilustra valores que representam as qualidades das recomendações segundo a métrica Breese para o sistema de filtragem baseado em dados simbólicos (X1) em comparação com o sistema de filtragem baseado no método kNN clássico (X2).

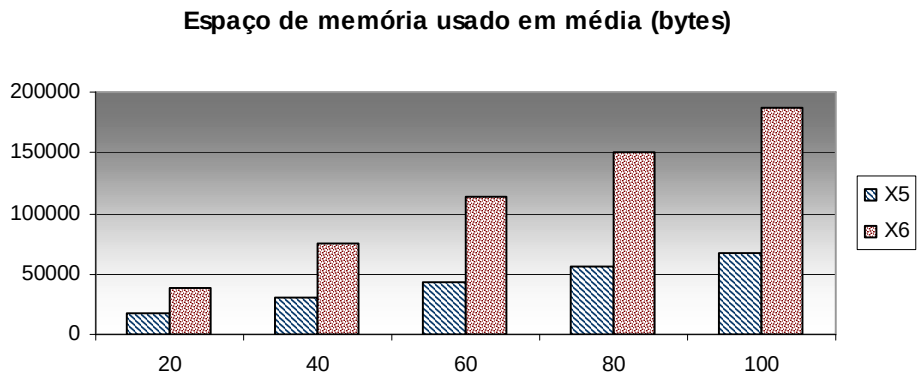
O Gráfico 1 mostra que a qualidade das recomendações segundo a métrica Breese é levemente maior no método kNN em comparação à abordagem simbólica quando o número de itens avaliados pelo usuário é maior ou igual a 60. Apesar disso, com um nível de significância de 1% (ou um grau de confiança de 99%), não se pode rejeitar a hipótese de que não há diferenças entre os dois métodos no que se refere à qualidade das recomendações segundo Breese.



	20	40	60	80	100
X_3	517,4247	841,4247	1065,8840	1243,3987	1347,6067
s_3	67,3107	117,2465	160,8172	198,1887	238,5304
X_4	535,9133	957,0413	1291,7653	1560,0247	1714,1227
s_4	79,0087	147,8273	205,3004	255,6256	299,9091
Z_2	-1,2596	-4,3329	-6,1246	-6,9218	-6,7632
Resultado (1%)	Aceito	Rejeitado	Rejeitado	Rejeitado	Rejeitado
Resultado (5%)	Aceito	Rejeitado	Rejeitado	Rejeitado	Rejeitado

Gráfico 2 - Este gráfico ilustra valores que representam o desempenho em função do tempo gasto na geração das listas de recomendação para o sistema de filtragem baseado em dados simbólicos (X3) em comparação com o sistema de filtragem baseado no método kNN clássico (X4).

Por outro lado, o Gráfico 2 mostra que quando a quantidade de itens avaliados pelo usuário aumenta, a Filtragem Baseada em Dados Simbólicos Modais supera significativamente a abordagem baseada no kNN. De fato, com um grau de confiança de 99%, pode-se rejeitar a hipótese de que não há diferenças entre as duas abordagens, no que se refere a velocidade de resposta, para perfis com 40 ou mais itens. Em contrapartida, aceita-se a hipótese alternativa de que o tempo gasto para geração das recomendações é menor na abordagem simbólica do que no método com kNN, quando o conjunto de treinamento é maior ou igual a 40.



	20	40	60	80	100
X_5	17062,5920	30838,2760	43601,7587	55730,5200	67200,0440
s_5	2631,6826	4856,1707	6743,3564	8467,4672	9968,8881
X_6	38856,0560	75802,5880	112979,8167	150053,7680	187334,8887
s_6	687,5496	1231,0202	1771,9680	2182,6430	2903,3973
Z_3	-56,6552	-63,4651	-70,3610	-76,2748	-81,8140
Resultado (1%)	Rejeitado	Rejeitado	Rejeitado	Rejeitado	Rejeitado
Resultado (5%)	Rejeitado	Rejeitado	Rejeitado	Rejeitado	Rejeitado

Gráfico 3 - Este gráfico ilustra valores que representam o espaço de memória em bytes utilizado pelo perfil do usuário durante a geração das listas de recomendação pelo sistema de filtragem baseado em dados simbólicos (X5) em comparação com o sistema de filtragem baseado no kNN clássico (X6).

Por último, observa-se a partir do Gráfico 3 uma diferença bastante expressiva entre as duas abordagens no que se refere ao espaço de memória ocupado pelo perfil do usuário. Neste caso, a abordagem simbólica modal provê melhorias de no mínimo 50%, chegando até 65% em alguns casos. De fato, pode-se aceitar, com um nível de significância de 1%, a hipótese de que o espaço de memória utilizado pela abordagem simbólica é menor do que o espaço exigido pelo método kNN.

A filtragem de informação baseada em dados simbólicos modais é mais rápida e muito mais econômica do que a filtragem com kNN, embora mantenha a qualidade das recomendações. Umas das implicações desse fato é, por exemplo, que para se alcançar uma precisão média de 0,3 na qualidade das recomendações, a abordagem simbólica leva em média 1,065 segundos para geração de uma lista de sugestões para um usuário e requer aproximadamente 43602 bytes em média para armazenar um perfil com 60 itens avaliados. Já a abordagem baseada no kNN gasta para a

mesma qualidade 1,291 segundos e requer cerca de 112980 bytes em média para um perfil também com 60 itens.

A fim de visualizar melhor essa diferença, considere um sistema de recomendação para WEB, que pode chegar facilmente ao pico de 1000 requisições de uma só vez. Nesta situação, um sistema com filtragem baseada no método kNN poderia levar até 21 minutos (1,291 segundos vezes 1000 dividido por 60) e requerer até 107 MB (112980 bytes vezes 1000 dividido por 1048576) para responder a todas estas requisições. Para lidar com isso, é comum que o processamento seja dividido em *clusters* de processadores ou mesmo servidores, o que implica em gastos.

Agora, imagine se o sistema de recomendação tivesse como método de filtragem de informação a nova abordagem apresentada neste trabalho. Com ela, seria possível responder as 1000 requisições em no máximo 18 minutos (1,065 segundos vezes 1000 dividido por 60) e ocupando apenas 42 MB (43602 bytes vezes 1000 dividido por 1048576) de memória. A conseqüência disso seria a diminuição de custos visto que seria necessário um menor número de recursos computacionais.

Não é difícil verificar porque o método de filtragem baseado em dados simbólicos alcança melhores resultados do que a abordagem com kNN no que diz respeito à velocidade de resposta na classificação e ao espaço de memória utilizado no armazenamento do perfil. A fim de analisar este fato, considere o cenário mencionado na seção 5.2, em que supunha-se um ambiente de um Sistema de Recomendação na WEB. Nesse caso, esse sistema deve poder gerar listas de recomendações para N usuários ao mesmo tempo. Para isso, a filtragem baseada no kNN possui um tempo de classificação ao nível de cada usuário proporcional a $S*M*R$, em que S é a quantidade de itens do repositório a serem filtrados, M é a quantidade de itens que constituem o perfil do usuário e R é o número de descritores de uma instância. Já a filtragem baseada em dados simbólicos possui um tempo de classificação proporcional a $S*2*R$, visto que os itens que constituem o perfil do usuário são agregados em uma das duas descrições simbólicas modais que representam o perfil do usuário. Como M é geralmente maior do que 2, a filtragem baseada em dados simbólicos modais acaba provendo melhores resultados do que a filtragem com kNN

tanto no tempo de classificação quanto no espaço de memória utilizado pelo perfil.

De fato, o resultado mostrado no exemplo anterior pode ser generalizado a outros domínios de filtragem de informação visto que a forma como o perfil do usuário é representado, ou seja, através de descrições simbólicas modais, permanece igual independentemente do domínio. Logo, ao modelar o perfil através de descrições simbólicas modais, elimina-se necessariamente uma dimensão que influi no tempo de resposta e no espaço utilizado para armazenamento do perfil. Esta dimensão é a quantidade de itens avaliados pelo usuário.

No próximo capítulo são apresentadas as conclusões deste trabalho bem como possíveis trabalhos futuros que possam aprimorar o método desenvolvido ou mesmo aplicá-lo em outros domínios em que recorrem problemas de sobrecarga de informação.

Capítulo 6

CONCLUSÕES

6.1. CONCLUSÕES

A quantidade de informações acessíveis às pessoas é cada vez maior, sobretudo, devido ao crescimento vertiginoso que a Internet vem obtendo nos últimos anos. Aliado a isso, o número de usuários de Internet bem como o de adeptos ao Comércio Eletrônico tem crescido expressivamente. Estes fatos vêm trazendo implicações nos mais diversos segmentos. Uma delas é a possibilidade de lojas virtuais oferecerem atendimentos únicos aos seus clientes a fim de fidelizá-los e maximizar suas compras. Como descrito nos capítulos 1 e 2, esse tipo de interação com o cliente é atualmente chamado de marketing *one-to-one* e a tecnologia que dá suporte a essa estratégia de relacionamento é conhecida como Sistema de Recomendação.

Conforme descrito nos capítulos 2 e 3, os Sistemas de Recomendação são tecnologias baseadas em algoritmos de filtragem de informação que, em geral, seguem uma destas estratégias de filtragem: filtragem colaborativa, filtragem baseada em conteúdo ou filtragem híbrida. Segundo mencionado também nestes capítulos, as estratégias colaborativa e por conteúdo apresentam problemas. Assim, algumas abordagens procuram minimizar os problemas de ambos através de uma filtragem híbrida. No entanto, por melhores que sejam as abordagens híbridas elas continuarão esbarrando nos problemas inerentes dos métodos de filtragem que lhes servem de base.

Este fato constitui uma das principais motivações da abordagem simbólica modal, apresentada no capítulo 4. Esse método de filtragem de informação baseada em conteúdo possui como alicerce as teorias de Análise de Dados Simbólicos (seção 4.2), que é uma extensão do domínio de Descoberta do Conhecimento (*Knowledge Discovery*).

Nesta nova abordagem o perfil do usuário é modelado através de um conjunto de descrições simbólicas modais que sumarizam as informações dos itens previamente avaliados pelo usuário (seção 4.3). Uma função de dissimilaridade que leva em conta as diferenças em posição e em

conteúdo foi criada a fim de possibilitar a comparação entre um novo item e o perfil do usuário (subseção 4.4.1). De fato, esta não é a única função criada. Outras possibilidades são descritas em Bezerra et al. 2002a e em De Carvalho e Bezerra 2002.

O desempenho deste novo método é avaliado a partir de comparações com uma outra abordagem em filtragem de informação: o kNN clássico. Para isso, foi modelado um ambiente experimental baseado no EachMovie e definida uma metodologia baseada em testes estatísticos para avaliação dos resultados.

Para concluir, podem ser enumeradas as seguintes contribuições do trabalho:

- Modelagem de estruturas complexas (filmes) via dados simbólicos modais. Pode-se citar, por exemplo, a modelagem da *sinopse* de um filme, que se trata de uma variável textual repleta de informações.
- Modelagem do perfil do usuário via generalização das descrições dos itens com dois perfis (um positivo e outro negativo) via dados simbólicos modais. Esse conceito pode ser modificado de domínio para domínio mas sua essência é representar mais fielmente os diversos interesses do usuário em estruturas agregadas e passíveis de manipular.
- No método desenvolvido neste trabalho, o perfil do usuário pode ser construído incrementalmente, de forma que novos itens sejam adicionados ao perfil na medida que o usuário faça novas avaliações. Com isso, o custo para construção e manutenção do perfil é insignificante, diferentemente de alguns métodos de filtragem por conteúdo que possuem formas próprias de armazenar o perfil do usuário.
- Comparação entre um item e um perfil via um índice de dissimilaridade que mede as diferenças em posição e conteúdo (comparação de distribuições de pesos).
- Construção de um ambiente experimental de avaliação do desempenho do modelo. O ambiente criado, baseado na

base de dados já consolidada EachMovie, permite tanto experimentos envolvendo algoritmos de filtragem colaborativa quanto algoritmos de filtragem por conteúdo. Aliado a isso, definiu-se uma metodologia baseada em testes estatísticos para avaliação do desempenho dos sistemas.

- O método criado possui a mesma qualidade nas recomendações e ainda assim é melhor tanto em velocidade quanto em memória com relação ao kNN clássico. Foi mostrado no capítulo anterior, com um nível de significância de 1%, que a Filtragem Baseada em Dados Simbólicos Modais é mais rápida e requer bem menos memória (pelo menos duas vezes menos) do que a filtragem com o kNN. Isso permite uma economia relevante para as empresas que utilizam Sistemas de Recomendação.

6.2. TRABALHOS FUTUROS

Diversas evoluções deste trabalho podem ser vislumbradas tanto no aspecto teórico quanto experimental:

- Prover mecanismos de aquisição implícita de perfil como, por exemplo, associar as ações do usuário em um website com seus interesses. Uma das formas de se fazer isso seria mapear cada tipo de ação do usuário a um subperfil específico.
- Aplicar o método desenvolvido em cenários complexos como o de lojas virtuais, em que não há aquisição explícita do perfil. Neste caso, a filtragem com dados simbólicos modais apresentada no Capítulo 4 seria adaptada para suportar diversos sub-perfis, cada qual com uma semântica em particular. Por exemplo, podemos imaginar o sub-perfil dos produtos comprados, o sub-perfil dos produtos consultados, o sub-perfil dos produtos colocados no carrinho de compras, etc. Dessa forma, o perfil do

usuário seria constituído de vários sub-perfis e o processo de recomendação seria em função do peso de cada um deles em uma aplicação real.

- Criar um método de filtragem híbrida, inspirado na idéia utilizada no sistema *Fab* (Balanovic e Shoham 1997). Nesse caso, a Filtragem por Conteúdo Baseada em Dados Simbólicos Modais (Capítulo 4) seria útil para construir o perfil do usuário e determinar a semelhança entre usuários. Por sua vez, o método de filtragem colaborativa (subseção 3.3.1) baseado, por exemplo, no kNN (Resnick et al. 1994, Shardanand e Maes 1995), seria utilizado para efetuar recomendações.
- Outra alternativa de filtragem híbrida, seria a combinação, via ponderação, por exemplo, das recomendações efetuadas pela filtragem baseada em dados simbólicos modais com uma das técnicas de filtragem colaborativa.
- Uma possibilidade de aumentar a velocidade das respostas nos Sistemas de Recomendação é a redução do conjunto de instâncias que fazem parte do perfil do usuário. Dessa forma, poderíamos aplicar um método como o proposto por Wilson e Martinez (2000) antes da execução dos algoritmos de filtragem híbridos.
- Uma outra possibilidade de trabalho futuro, visando melhores níveis de recomendações, poderia ser a aplicação de métodos de *agrupamento* no perfil do usuário. Os algoritmos de *agrupamento* em sua maioria procuram maximizar a similaridade intra-grupo e minimizar a similaridade inter-grupo. Se tratarmos os perfis positivo e negativo de um usuário como dois grupos distintos, poderíamos aplicar um algoritmo de *agrupamento* a fim de aumentar a similaridade no interior de cada perfil e diminuir a similaridade entre os perfis positivo e negativo.

Bibliografia

[1] Aggarwal, C. C. and Yu, P. S. *A new framework for itemset generation*. In PODS 98, Symposium on Principles of Database Systems, pages 18–24, Seattle, WA, USA, 1998.

[2] Agrawal, R., Imielinski, T. and Swami, A. *Mining associations between sets of items in large databases*. In Proc. of the ACM SIGMOD Int'l Conference on Management of Data, pages 207–216, Washington D.C., May 1993.

[3] Agrawal, Rakesh and Srikant, Ramakrishnan. *Fast algorithms for mining association rules*. In Jorge B. Bocca, Matthias Jarke, and Carlo Zaniolo, editors, Proc. 20th Int. Conf. Very Large Data Bases, VLDB, pages 487–499, Santiago, Chile, Sept 1994.

[4] Aha, David W., Kibler, D. and Albert, Marc K. *Instance-Based Learning Algorithms*. Machine Learning, 6, pages 37-66, 1991.

[5] Arya, S. *Nearest Neighbor Searching and Applications*. Ph.D. thesis, University of Maryland, College Park, MD, 1995.

[6] Baeza-Yates, R. e Berthier, R.N. *Modern Information Retrieval*. Addison-Wesley, Boston, 1999.

[7] Balanovic, M. e Shoham, Y. *Content-based, collaborative recommendation*. Communications of the ACM, 40, pages 88-89, Mar. 1997.

[8] Basu, C., Hirsh, H., and Cohen W. *Recommendation as classification: using social and content-based information in recommendation*. In Proceedings of the 1998 National Conference on Artificial Intelligence (AAAI-98), pages 714-720.

[9] Bentley, J. *Multidimensional binary search trees used for associative searching*. Communications of the ACM, Vol.18, pages 509-517, 1975.

[10] Bezerra, B. L. D., De Carvalho, F. A. T. , Ramalho, G. L., and Zucker, Jean-Daniel. *Speeding up Recommender Systems with Meta-*

Prototypes. Lecture Notes in the Proceedings of the XVI Brazilian Symposium on Artificial Intelligence, Porto de Galinhas/Recife, Brazil, November 11-14, 2002a, Springer, pages 227-236.

[11] Bezerra, B. L. D., De Carvalho, F. A. T. , Ramalho, G. L., e Zucker, Jean-Daniel. *Speeding up Recommender Systems*. Proceedings of the UM 2002 Workshop on Personalization in Future TV in conjunction with 2nd International Conference on Adaptive Hypermedia and Adaptive Web Based Systems, 2002b, pp.. http://www.di.unito.it/~liliana/TV02/completeProceedings.pdf.

[12] Bezerra, B. L. D. e De Carvalho, F. A. T. *Aprimorando o Desempenho de Sistemas de Recomendação Personalizados com Meta-Protótipos*. Shortpaper do IV Encontro Nacional de Inteligência Artificial em conjunto com XXIII Congresso da Sociedade Brasileira de Computação, 2003, páginas 2363-2368.

[13] Bezerra, B. L. D. e De Carvalho, F. A. T. *Information Filtering based on Modal Symbolic Objects*. Information Processing Letters, aceito em 2003, a ser publicado em 2004.

[14] Bock, H. H. and Diday, E. *Analysis of Symbolic Data*. Springer, Heidelberg, 2000.

[15] Breese, J., Heckerman, D., and Kadie, C. 1998. *Empirical analysis of predictive algorithms for collaborative filtering*. In Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence (UAI-98), pages 43-52.

[16] Brin, Sergey, Motwani, Rajeev, Ullman, Jeffrey D., and Tsur, Shalom. *Dynamic itemset counting and implication rules for market basket data*. In SIGMOD 1997, Proceedings ACM SIGMOD International Conference on Management of Data, pages 255–264, Tucson, Arizona, USA, May 1997.

[17] Chan, P. *A non-invasive learning approach to building web user profiles*. Workshop on Web usage analysis and user profiling, Fifth

International Conference on Knowledge Discovery and Data Mining, San Diego, August 1999, pages 7-12.

[18] Claypool, M., Brown, D., Phong Le, Waseda M. *Inferring User Interests*. IEEE Internet Computing, 5(6), pages 32-39, Nov/Dec 2001.

[19] Claypool, M., Gokhale, A., Miranda, T., Murnivok, P., Netes, D. e Sartin, M. *Combining Content-based and Collaborative Filters in an On-line Newspaper*. In Proceedings of ACM SIGIR Workshop on Recommender Systems, August 19 1999. http://citeseer.ist.psu.edu/article/claypool99combining.html.

[20] Cohn, D., Atlas, L., and Ladner, R. *Improving generalization with active learning*. Machine Learning, v.15 n.2, pages 201-221, May 1994.

[21] Cotter, P. and Smyth, B. *PTV: Intelligent Personalised TV Guides*. Proceedings of the 12th Innovative Applications of Artificial Intelligence (IAAI) Conference. AAAI Press, 2000, pages 957-964.

[22] Cover, T. M., e Hart, P. E. (1967). *Nearest Neighbor Classifiers*. IEEE Transactions on Computers, 23-11, November, 1974, pages 1179-1184.

[23] De Carvalho, F. A. T. e Bezerra, B. L. D. *Information Filtering based on Modal Symbolic Objects*. Proceedings of the 26th Annual Conference of the Gesellschaft für Klassifikation (GfKl), Springer, 2002, pages 395-404.

[24] Demiriz, A. *Enhancing Product Recommender Systems on Sparse Binary Data*. Accepted to be published in the Journal of Data Mining and Knowledge Discovery (2003), http://citeseer.ist.psu.edu/demiriz02enhancing.html.

[25] Domingos, Pedro. *Rule Induction and Instance-Based Learning: A Unified Approach*. in: C.S. Mellish (Ed.), Proc. IJCAI-95, Montreal, Quebec, Morgan Kaufmann, San Mateo, CA, 1995, pages 1226-1232.

[26] Drucker, Peter F. *The New Realities*. 1st Edition, Transaction Pub, 2003.

[27] Easton, Jaclyn e Bezos, Jeff. *Strikingitrich.com (Striking It Rich.com): Profiles of 23 Incredibly Successful Websites You've Probably Never Heard Of.* 1st Edition, McGraw-Hill, 1998.

[28] Engelbrecht, A. P. e Brits, R.. *Supervised Training Using an Unsupervised Approach to Active Learning.* In: Neural Processing Letters, Volume 15 , Issue 3, pages 247 – 260, 2002.

[29] Geyer-Schulz, Andreas e Hahsler, Michael. *A customer purchase incidence model applied to recommender systems.* In: Kohavi, R., Masand, B. M., Spiliopoulou, M., Srivastava, J. (eds.): WEBKDD 2001 - Mining Log Data Across All Customer Touch Points. Lecture Notes in Computer Science LNAI 2356, pages 25-47, Springer-Verlag, 2002.

[30] Geyer-Schulz, Andreas e Hahsler, Michael. *Evaluation of Recommender Algorithms for an Internet Information Broker based on Simple Association Rules and on the Repeat-Buying Theory.* In Proceedings WEBKDD'2002, Eds. B. Masand, M. Spiliopoulou, J. Srivastava, O. Zaïane, Edmonton, Canada, July 2002, pages 100–114.

[31] Goecks, J. and Shavlik, J. W. (2000). *Learning users' interests by unobtrusively observing their normal behavior.* In Proceedings of the ACM Intelligent User Interfaces Conference (IUI), Jan. 2000, pages 129-132.

[32] Good, N., Schafer, J.B., Konstan, J.A., Borchers, A., Sarwar, B., Herlocker, J., and Riedl, J. *Combining Collaborative Filtering with Personal Agents for Better Recommendations.* In Proceedings of AAAI-99, AAAI Press, pages 439-446, 1999.

[33] Hasenjäger, M. *Active Data Selection in Supervised and Unsupervised Learning.* PhD thesis, Technische Fakultät der Universität Bielefeld, Jan 2000.

[34] Herlocker, J., Konstan, J.A., Borchers, A., and Riedl, J. *An algorithmic framework for performing collaborative filtering.* Proceedings of SIGIR'99, pages 230-237, 1999.

[35] Hill, W.C. and Hollan, J.D. *Edit Wear and Read Wear*. In Proceedings of CHI, Monterey, Canada, ACM Press, pages 3-9, Apr. 1992.

[36] Kalakota, R. e Robinson, M. *e-Business 2.0 – Roadmap for Success*. 2nd Edition, Addison-Wesley Longman, 2001.

[37] Konstan, J., Miller, B., Maltz, D., Herlocker, J., Gordon, L. and Riedl, J. *GroupLens: Applying Collaborative Filtering to Usenet News*. Communications of the ACM, 40(3), pages 77-87, 1997.

[38] Krukwich, B., Burkey, C. *Learning user information interests through extraction of semantically significant phrases*. Working Notes of the 1996 AAAI Spring Symposium on Machine Learning in Information Access, March 1996.

[39] Lieberman, H. *Autonomous Interface Agents*. In Proceedings of the ACM Conference on Computers and Human Interface (CHI), 1997, pages 67-74.

[40] Lin, W., Alvarez, S. A. and Ruiz, C. *Efficient adaptive-support association rule mining for recommender systems*. Data Mining and Knowledge Discovery, 6, pages 83-105, 2002.

[41] McJones, P. *EachMovie collaborative filtering data set.* DEC Systems Research Center, http://www.research.digital.com/SRC/eachmovie/, 1997.

[42] Miller, B. N., Albert, I., Lam, S. K., Konstan, J. A. e Riedl, J. *MovieLens Unplugged: Experiences with an Occasionally Connected Recommender System.* In Proceedings of ACM 2003 International Conference on Intelligent User Interfaces (IUI'03) (Accepted Poster), January 2003.

[43] Mitchell, Tom M. *Machine Learning*. McGraw-Hill, 1997.

[44] Mobasher, B., Dai, H., Luo, T., Nakagawa, M., Sun, Y. and Wiltshire, J. *Discovery of aggregate usage profiles for web personalization*. In ACM WebKDD 2000 Web Mining for E-Commerce Workshop, 2000. http://citeseer.ist.psu.edu/mobasher00discovery.html

[45] Mobasher, B., Dai, H., Luo, T. and Nakagawa, M. *Discovery and evaluation of aggregate usage profiles for web personalization*. Data Mining and Knowledge Discovery, 6, pages 61-82, 2002.

[46] Morita, M., & Shinoda, Y. (1994). *Information filtering based on user behavior analysis and best match text retrieval*. In Proceedings of SIGIR, Dublin, Ireland, ACM Press, 1994, pages 272-281.

[47] Nichols, D. M. *Implicit Rating and Filtering*. In Proceedings of the 5th DELOS Workshop on Filtering and Collaborative Filtering, Budapest, Hungary, 10-12 November 1997, ERCIM, 31-36. ISBN: 2-912335-04-3.

[48] Oard, D. and Kim, J. *Implicit Feedback for Recommender Systems*. In AAAI Technical Report WS-98-08: Workshop on Recommender Systems, July 27, Madison, WI.

[49] Olsson, T. *Bootstrapping and Decentralizing Recommender Systems.* Licentiate Thesis 2003-006, Department of Information Technology, Uppsala University and SICS, 2003.

[50] Pennock, D. M., Horvitz, E. e Giles, C. L. *Social choice theory and recommender systems: Analysis of the axiomatic foundations of collaborative filtering.* Proceedings of the 17th National Conference on Artificial Intelligence (AAAI-2000), pages 729-734, Austin, TX, July 2000.

[51] Queiroz, S. R. M., De Carvalho, F. A. T., Ramalho, G. L., Corruble, V. *Making Recommendations for Groups using Collaborative Filtering and Fuzzy Majority.* In: 16th Brazilian Symposium on Artificial Intelligence (SBIA2002), Porto de Galinhas / Recife, 2002, pages 248-258.

[52] Queiroz, S. R. M. e De Carvalho, F. A. T. *An Item-Based Symbolic Approach for Making Group Recommendations.* In: 27th Annual Conference of the Germany Classification Society (GfKl2003), Cottubus (Germany), 2003.

[53] Queiroz, S. R. M., *Estratégias de Recomendação para Grupos baseadas em Filtragem Colaborativa*. Tese de Mestrado em Inteligência Artificial, Universidade Federal de Pernambuco, 2003.

[54] Queiroz, S. R. M. e De Carvalho, F. A. T. *A Symbolic Model-based Approach for Making Collaborative Group Recommendation* (Aceito para publicação) In: 9th Conference of the International Federation of Classification Societies, Chicago (USA), 2004.

[55] Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., and Riedl, L. 1994. *Grouplens: An Open Architecture for Collaborative Filtering of Netnews*. In Proceedings of ACM CSCW'94 Conference on Computer-Supported Cooperative Work, pages 175-186.

[56] Richardson, M., and Domingos, P. *The Intelligent Surfer: Probabilistic Combination of Link and Content Information in PageRank*, Advances in Neural Information Processing Systems 14, pages 1441-1448, 2002. Cambridge, MA: MIT Press.

[57] Sarwar, B., Karypis, G., Konstan, J. and Riedl, J. *Application of dimensionality reduction in recommender systems – a case study*. In ACM WebKDD 2000 Web Mining for E-Commerce Workshop, 2000.

[58] Sarwar, B., Konstan, J., Borchers, A., Herlocker, J., Miller, B. e Riedl, J. *Using Filtering Agents to Improve Prediction Quality in the Grouplens Research Collaborative Filtering System*. In Proceedings of the ACM Conference on Computer Supported Cooperative Work (CSCW), 1998, pages 345-354.

[59] Schafer, J.B., Konstan, J.A., and Riedl, J. *Recommender Systems in E-Commerce*. In ACM Conference on Eletronic Commerce (EC-99), pages 158-166, 1999.

[60] Schafer, J.B., Konstan, J.A. e Riedl, J. *E-Commerce Recommendation Applications*. Data Mining and Knowledge Discovery, 2001, Volume 5, Issue 1-2, pages 115-153.

[61] Shardanand, U. and Maes, P. 1995. *Social information filtering: Algorithms for automating "word of mouth"*. In Proceedings of ACM

CHI'95 Conference on Human Factors in Computing Systems, pages 210-217.

[62] Tang, T. Y., Winoto, P. e Chan, K. C. C. *On the Temporal Analysis for Improved Hybrid Recommendations.* In Proceedings of 2003 IEEE/WIC International Conference on Web Intelligence (WI 2003), October, 13-17, Halifax, Canada, pages 214-220, 2003.

[63] Teixeira, I. R., De Carvalho, F. A. T., Ramalho, G. L., Corruble, V.. *Active CP: A method for Speeding up User Preferences Acquisition in Collaborative Filtering Systems*. In: 16th Brazilian Symposium on Artificial Intelligence (SBIA2002), Porto de Galinhas / Recife, 2002, pages 237-247.

[64] Teixeira, I. R. *Um Método de Aprendizagem Ativa em Sistemas de Filtragem Colaborativa*. Tese de Mestrado em Inteligência Artificial, Universidade Federal de Pernambuco, 2002.

[65] Vucetic, S. and Obradovic, Z. *A regression-based approach for scalling-up personalized recommender systems in e-commerce*. In ACM WebKDD 2000 Web Mining for E-Commerce Workshop, 2000.

[66] Wilson, D. R. and Martinez, T. R. *Reduction techniques for exemplar-based learning algorithms*. Machine Learning, 38(3), pages 257-268, 2000.

[67] Witten, Ian H. e Frank, Eibe. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations.* San Diego, CA: Morgan Kaufmann, 2000.

[68] Wolf, J., Aggarwal, C., Wu, K-L., and Yu, P. 1999. *Horting Hatches an Egg: A New Graph-Theoretic Approach to Collaborative Filtering*. In Proceedings of the ACM KDD'99 Conference. San Diego, CA. pages 201-212.

[69] Yu, K., Xu, X., Martin, E. and Kriegel, H.P. *Selecting relevant instances for efficient accurate collaborative filtering*. In Proceedings of the 10th CIKM, pages 239-246, ACM Press, 2001.