PREVISÃO DE AVALIAÇÕES EM SISTEMAS DE RECOMENDAÇÃO PARA NICHOS DE MERCADO

Marcelo Rezende de Fazio

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Engenharia de Sistemas, COPPE, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Mestre em Engenharia de Sistemas e Computação.

Orientador: Geraldo Zimbrão da Silva

Rio de Janeiro
Abril de 2013

PREVISÃO DE AVALIAÇÕES EM SISTEMAS DE RECOMENDAÇÃO PARA NICHOS DE MERCADO

Marcelo Rezende de Fazio

DISSERTAÇÃO SUBMETIDA AO CORPO DOCENTE DO INSTITUTO ALBERTO LUIZ COIMBRA DE PÓS-GRADUAÇÃO E PESQUISA DE ENGENHARIA (COPPE) DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM CIÊNCIAS EM ENGENHARIA DE SISTEMAS E COMPUTAÇÃO.

Examinada por:

_____

Prof. Geraldo Zimbrão da Silva, D.Sc.

_____

Prof. Jano Moreira da Silva, Ph.D.

_____

Profª. Jonice de Oliveira Sampaio, D.Sc.

RIO DE JANEIRO, RJ - BRASIL

ABRIL DE 2013

Fazio, Marcelo Rezende de

Previsão de avaliações em sistemas de recomendação para nichos de mercado. /Marcelo Rezende de Fazio. - Rio de Janeiro: UFRJ/COPPE, 2013.

XIII, 97 p.: il.; 29,7 cm.

Orientador: Geraldo Zimbrão da Silva

Dissertação (mestrado) - UFRJ/COPPE/Programa de Engenharia de Sistemas e Computação, 2013.

Referências Bibliográficas: p. 97 - 101.

1. Sistemas de Recomendação. 2. Nichos de mercado. 3. Previsão de avaliações. I. Silva, Geraldo Zimbrão da. II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia de Sistemas em Computação. III. Título.

Dedicatória

À minha família.

Agradecimentos

Primeiramente, gostaria de agradecer a minha mãe que sempre demonstrou ser uma guerreira, trabalhando muito para que eu e minha irmã possamos ter uma vida melhor. Uma mulher fora de série na qual minha vida sem ela não faria sentido.

Ao meu pai que sempre me incentivou a estudar, mostrando como é importante a educação e o respeito às pessoas ao longo da vida. À minha irmã e tia Bel, que mesmo distantes, permanece um carinho enorme que tenho por elas.

Ao Prof. Geraldo Zimbrão, por me orientar e fazê-lo com imensa dedicação e apoio, mesmo com suas atividades de professor e coordenador. Há momentos que me espanta o seu conhecimento sobre tantas áreas e a facilidade para repassá-lo. Obrigado!

Aos meus chefes Giordano Almiro Machado Moraes e Fabio Marzullo por sempre me apoiarem na realização dessa dissertação. Sem vocês nada disso seria possível!

Ao Carlos Eduardo Mello, que sempre solicito às minhas dúvidas, iniciou a ideia dessa dissertação na cadeira de Mineração de Dados ministrada pelo Prof. Geraldo Zimbrão.

Ao professor Jano Moreira da Silva, pela grande experiência e conhecimento repassado ao longo de todo o meu mestrado.

À professora Jonice de Oliveira Sampaio por aceitar fazer parte desta banca.

À Universidade Federal do Rio de Janeiro - UFRJ, pelo orgulho que tenho de ser formado por uma das melhores universidades do país.

Resumo da Dissertação apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Mestre em Ciências (M.Sc.)

PREVISÃO DE AVALIAÇÕES EM SISTEMAS DE RECOMEDAÇÃO PARA NICHOS DE MERCADO

Marcelo Rezende de Fazio

Abril/2013

Orientador: Geraldo Zimbrão da Silva

Programa: Engenharia de Sistemas e Computação

Este trabalho apresenta novos algoritmos para melhorar a qualidade das recomendações realizadas pelos Sistemas de Recomendações, considerando nichos de mercado como fator para obter a melhora nos resultados. Três experimentos foram realizados para provar a eficiência dos algoritmos apresentados. Os resultados experimentais observados corroboram os resultados previstos analiticamente. Além disso, foi elaborada uma revisão bibliográfica sobre a área de Sistemas de Recomendação, realizada com base em trabalhos clássicos e incluindo os mais recentes desenvolvimentos e propostas na área.

Abstract of Dissertation presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Master of Science (M.Sc.)

PREVISION OF GRADES IN RECOMMENDATION SYSTEM FOR NICHES MARKET

Marcelo Rezende de Fazio

April/2010

Advisors: Geraldo Zimbrão da Silva

Department: Computer Science Engineering

This work show new algorithms to improve recommendations quality made by Recommendations Systems, considering niche market as a factor to obtain improvements in results. Three experiments were accomplished to prove the efficient of algorithms. Results from observed experiments corroborate with the analytically previewed results. Furthermore, a bibliographic revision was elaborated in Recommendations Systems area, based in classical works and including the most recent developments and proposals in the area.

Sumário

Lista de Quadros

Lista de Imagens

Capítulo 1 - Introdução

1.1 Motivação

A quantidade de informação disponível na web é consideravelmente grande (RASHID *et al.*, 2006), portanto há uma necessidade de filtrá-la e apresentá-la de acordo com as necessidades de visualização de conteúdo dos interessados, que, por diversas vezes, confundem-se em meio a tanta informação.

Sendo o universo de alternativas grande, uma pessoa poderá não ter conhecimentos para fazer uma escolha. Nesse caso, uma alternativa para solucionar a questão é que ela procure o que as outras estão interessadas (TERVEEN *et al.*, 2001).

Uma recomendação pode ser sugerida diretamente para um indivíduo específico, bem como para um conjunto de pessoas que estejam interessadas na sugestão. Porém, o ser humano é incapaz de recomendar, com qualidade, itens para milhões de usuários numa base de milhares de itens - apesar de ser capaz de processar uma variedade de dados heterogêneos, inclusive padrões de utilização e preferências dos usuários que o ajudam a escolher itens (KRISHNAN *et al.*, 2008).

Logo, encontrar formas de filtrar o conteúdo e oferecer relevância ao usuário tornou-se um imenso diferencial competitivo entre as organizações, principalmente para o mercado virtual.

Diante desse contexto, os sistemas de recomendação têm como objetivo conceber informações significativas para um conjunto de usuários dentro de um conjunto maior de informações, por meio de técnicas de mineração de dados. Em outras palavras, eles facilitam o encontro de itens desejáveis para as pessoas, que podem ser realizadas por meio de uma lista de itens de que o indivíduo possa gostar. Esses sistemas podem ser

utilizados, por exemplo, dentro de *sites* de comércio de vendas de filmes virtual, para sugerir filmes dos quais os usuários possam gostar e consequentemente comprá-los.

Evidências sugerem que o mercado da internet possa ajudar a mudar o balanço de vendas uma vez que a maioria das vendas realizadas era constituída de poucos produtos que possuíam um grande volume de vendas. Esse padrão está sendo substituído por outro no qual a venda de produtos é realizada para um público alvo (BRYNJOLFSSON, 2011). ANDERSON (2004) elaborou o termo “cauda longa” para descrever esse fenômeno em que a venda de produtos para um determinado público pode aumentar, tornando-os relevantes em relação ao total de vendas realizadas.

No entanto, os sistemas de recomendação baseados em filtragem colaborativa tradicional podem ser inapropriados para recomendar itens aos seus clientes (LI *et al.*, 2011). Esses são suportados por modelos estáticos nos quais suas relações são fixas ao longo do tempo. Os sistemas baseados em fatores temporais podem ser a chave no projeto de sistemas de recomendação para melhorar a qualidade tanto das previsões quanto das recomendações.

A previsão de avaliações de itens tem um papel fundamental tanto na utilização final do cliente quanto no suporte às recomendações realizadas pelo sistema de recomendação. Fundamentado na teoria da cauda longa e no papel da previsão, a proposta desta dissertação considera nichos de mercado e o tempo que ocorreram as avaliações como forma de melhorar a qualidade das previsões feitas pelo sistema de recomendação baseado na filtragem colaborativa gerando um novo sistema híbrido.

1.2 Objetivo

Os atuais sistemas de recomendação necessitam de melhorias nos métodos que fazem recomendações e previsões para torná-los mais eficientes e aplicáveis para uma escala maior de aplicações reais (ADOMAVICIUS, TUZHILIN, 2005). As melhorias

incluem métodos a fim de representar mais adequadamente tanto o comportamento do usuário quanto as informações sobre os itens que serão recomendados, incorporando-os ao contexto do processo de recomendações.

Um nicho de mercado consiste de um grupo de clientes que compartilham um conjunto similar e único de necessidades e preferências. Há poucos trabalhos que abordam nichos de mercados como fator de melhoria das previsões de avaliações nos sistemas de recomendação. O objetivo deste trabalho é melhorar a qualidade das previsões de itens destinados a nichos de mercado, por meio do comportamento dos usuários e dos itens nos filtros colaborativos baseado em usuários e itens. Além disso, será analisado o comportamento dos usuários de nichos de mercado em relação aos sistemas de recomendação.

1.3 Organização

Este trabalho está organizado em seis capítulos. Neste primeiro capítulo, de introdução, foram exibidos a motivação, o objetivo e a organização do trabalho.

No segundo capítulo, são apresentados os conceitos relativos aos sistemas de recomendação.

No terceiro capítulo, é feito um estudo mais aprofundado sobre as técnicas utilizadas pelos sistemas de recomendação baseados em filtragem colaborativa.

No quarto capítulo, são apresentados estudos que consideram o tempo e nichos de mercado como fatores que ajudam a melhorar a qualidade dos sistemas de recomendação baseada em filtragem colaborativa.

No quinto capítulo, é exibida a abordagem proposta, que consiste na construção de um sistema de recomendação baseado em nichos de mercado.

O último capítulo revela as conclusões e os resultados obtidos em cada etapa do trabalho e sugere um caminho para a realização de pesquisas no futuro.

Capítulo 2 - Sistemas de Recomendação

No contexto dos sistemas de recomendação, os usuários podem enviar suas preferências ou requisitá-las aos sistemas de recomendação, porém uma pessoa pode recebê-la sem a necessidade de uma requisição. As informações obtidas pelas pessoas são consolidadas na base de conhecimento, para que os sistemas de recomendação possam sugerir itens que, provavelmente, elas gostarão.

Sistemas de recomendação podem também manter informações pessoais dos usuários tais como sua localização, idade e informações sobre características dos itens para realizar previsões atingindo uma melhor qualidade (RASHID *et al.*, 2006). Alguns desses sistemas aplicam técnicas de descoberta de conhecimento para solucionar questões referentes à realização de recomendações personalizadas.

O modelo abaixo foi elaborado por TERVEEN *et al.* (2011) e exemplifica o processo de recomendação.

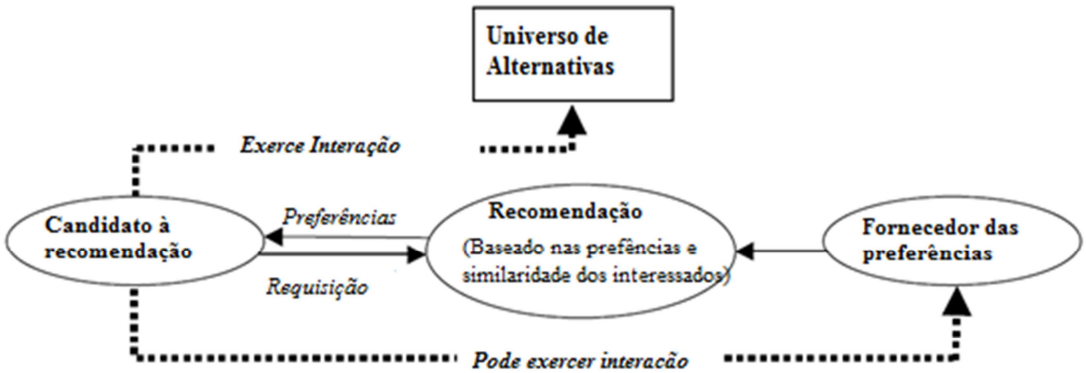

Figura 1 - Processo de Recomendação retirado de (TERVEEN, 2011)

A pessoa que recebe as recomendações é chamada de usuário ativo e, quando há uma previsão da avaliação dada por um usuário a um item, esse item é considerado um item ativo. Um item ou produto poderá ser qualquer objeto sobre o qual uma pessoa poderá expressar sua opinião (AGGARWAL *et al.*,2001).

Um item inclui:

- item varejo como por exemplo, um livro, CD de música, DVD;
- experiências, tais como um show, uma viagem;
- consumo como por exemplo, um artigo, reportagem, coluna de jornal:
- pessoas, que podem ser avaliadas pelas suas qualidades;
- emprego em uma empresa;
- propagandas.

Há várias formas de classificar os sistemas de recomendação. A mais comum classifica os sistemas em três classes: sistemas de filtragem colaborativa, os sistemas baseados em conteúdo e os sistemas híbridos. Os sistemas de filtragem colaborativa trabalham por meio de coleta das informações relativas aos usuários na forma de notas dadas aos itens e exploram as similaridades e diferenças entre perfis de vários usuários para determinar as recomendações.

Os sistemas de filtragem baseados em conteúdo fornecem recomendações comparando descrições do item com conteúdo de interesse do usuário (MELVILLE *et al.*,2002).

Há, no entanto, diferenças no modo em que esses sistemas analisam os dados históricos. Os sistemas de filtragem colaborativa trabalham somente com informações históricas. Já os sistemas de filtragem baseados em conteúdo fundamentam-se em atributos das entidades. Nesses sistemas, por exemplo, no domínio de uma locadora de

filmes, as recomendações de filmes baseados em conteúdo podem apoiar-se na localidade, na faixa etária, no sexo dos indivíduos ou no gênero, no diretor, na produtora do filme.

Já os sistemas híbridos combinam as características de ambas as classificações com o objetivo de melhorar a qualidade das previsões e recomendações.

Outra forma, menos comum, de classificar os sistemas de recomendação é por meio das avaliações explícitas ou implícitas que servem como entrada (BREESE *et al.*,1998) no processo de recomendação. Uma avaliação explícita refere-se a um usuário conscientemente avaliar um item. Já uma avaliação implícita refere-se à ação de um usuário, que pode se analisada por meio de registros históricos, padrões de utilização de um item, tempo de utilização.

Em alguns ambientes, há uma dificuldade de extrair avaliações explícitas. Por exemplo, algumas empresas provedoras de acesso à internet cobram o seu serviço baseado no tempo de conexão em que o cliente permanece conectado na internet. Para reduzir a cobrança dessas empresas ao acesso à internet, o indivíduo tenderá a diminuir o tempo conectado à internet, o que dificultará que avaliações explícitas sejam realizadas nesse ambiente. LEE *et al.* (2008) lida com essa questão por meio da construção de um efetivo sistema de recomendação para ambientes de comércio virtual sem usar avaliações explícitas.

Uma das melhores formas de quantificar a preferência dos usuários é quando um indivíduo avalia honestamente o item por intermédio de informações analisadas, no entanto, de modo geral, várias pessoas atribuem avaliações arbitrárias, sem refletir a sua verdadeira opinião sobre o item em questão (LEE *et al.* 2008). Quando se utiliza uma escala para avaliação explícita do usuário a um item, a escala par para avaliação

demonstra a neutralidade do usuário, porém valores ímpares levam o usuário a escapar-se de uma resposta direta (AGGARWAL *et al., 2*001).

Em comparação aos sistemas de recomendação, os seres humanos possuem grande facilidade para determinar a relevância, qualidade e interesse nos itens. Consequentemente, eles podem examinar dados que não são adequadamente bem investigados por computadores. Comparando-os a eles, os seres humanos são capazes de analisar diversas dimensões do item e do usuário, tais como qualidade do item e gosto pessoal do cliente, que é uma tarefa difícil de ser executada por computadores.

Nos sistemas de recomendação, um item sugerido por um sistema de recomendação pode ser útil para o usuário, no entanto pode não conter o conteúdo que ele espera. Por exemplo, um cliente que gosta de filmes de ação poderá receber uma recomendação de um filme de ação específico do qual ele não goste. Essa situação necessita ser avaliada pelos sistemas de recomendação, pois não atende plenamente os requisitos do cliente.

Um sistema de recomendação eficiente deverá, portanto, considerar todas as atividades e retorno de informações realizadas pelo usuário e analisar essas informações a fim de encontrar relacionamentos entre itens, usuários, e usuários e itens (JUNG, 2011).

2.1 Previsão e Recomendação

Previsão é um valor numérico que expressa o provável gosto de um usuário por um item o qual ele não tenha avaliado anteriormente. Já uma recomendação é uma lista de itens limitada dos quais o usuário irá gostar mais, exceto os itens que já foram avaliados ou comprados pelo usuário ativo, também conhecido como *Top-N recomendation* (SARWAR *et al.*,2001).

A previsão das avaliações realizadas para os usuários pelo sistema de recomendação é um fator, dentre outros, que aperfeiçoa a qualidade das recomendações realizadas, bem como facilita o indivíduo, enquanto ele está navegando em um *site* na internet, a prever se o usuário estará ou não interessado em um item.

2.2 Técnicas dos sistemas de recomendação

No nível mais alto, há, pelo menos, quatro técnicas de recomendação apresentadas por SCHAFER *et al.* (1999). São elas sem personalização, baseadas em atributos, correlação item-item e correlação usuário-usuário.

2.2.1 Sem personalização

Esta técnica recomenda produtos aos clientes baseada no que outros clientes relataram sobre o produto. As recomendações são independentes do cliente, logo todas as recomendações são iguais para todos eles. O esforço é baixo para gerar essas recomendações, pois a técnica não reconhece cada um dos seus clientes para sugerir itens.

2.2.2 Baseado em atributos

Os sistemas de recomendação baseados em atributos sugerem produtos para os clientes apoiados nas propriedades dos produtos. Normalmente, esses sistemas são manuais, pois os clientes devem requisitar a recomendação atribuindo propriedades desejadas aos seus produtos.

2.2.3 Correlação item-item

Esta técnica sugere os produtos por meio de um conjunto menor de produtos os quais os clientes mostraram interesse. Caso o cliente tenha escolhido alguns produtos para comprar, esse sistema recomenda produtos complementares para agregar a compra feita pelo cliente.

Os sistemas, que utilizam essa técnica, podem ser automáticos, caso eles sejam fundamentados em observações dos comportamentos padrões dos clientes e também podem ser manuais, desde que haja a necessidade de relacionar os itens. Além do mais, esses sistemas possuem como características não necessitarem de informações históricas dos clientes para sugerir as recomendações.

2.2.4 Correlação usuário-usuário

Essa técnica de recomendação sugere itens aos clientes e fundamenta-se na correlação entre o comprador e os outros compradores. Essa tecnologia é chamada frequentemente de filtragem colaborativa, pois as recomendações são produzidas por meio de técnicas de filtragem que utilizam opiniões dos usuários para recomendar itens personalizados para cada um dos clientes Os sistemas, que utilizam essa técnica, necessitam de dados históricos para produzir recomendações.

2.3 Sistemas de Recomendação em comércio virtual

Sistemas de recomendação são usados como ferramentas de negócios que remodelaram o comércio virtual no mundo (SCHAFER *et al.*,1999). Esses são usados no comércio virtual para sugerir produtos aos seus clientes dos quais gastam menos tempo procurando um item ou um serviço. Os produtos podem ser recomendados baseados em suas vendas, região onde mora o cliente, ou na análise do comportamento passado do cliente. As técnicas utilizadas para sugerir recomendações podem facilitar as personalizações das sugestões para cada indivíduo.

Ringo(SHARDANAND,1998) e *Video Recommender*(HILL,1995) foram um dos primeiros sistemas web que geraram recomendações para músicas e filmes respectivamente, por intermédio de um filtro colaborativo. Esses sistemas usaram informações históricas para que eles pudessem cumprir o seu objetivo de efetuar recomendações eficientes aos seus clientes.

Os sistemas de recomendação demonstraram capacidade de melhorar as vendas no comércio virtual através de três formas (SCHAFER *et al.*,1999):

a) Visitantes de um site Web frequentemente procuram em site de comércio virtual sem objetivo de comprar nada. Esses sistemas poderiam ajudar o cliente a encontrar produtos os quais eles gostariam de comprar.

b) Esses sistemas facilitariam a compra casada. Uma compra casada sugere que um produto adicional seja oferecido ao cliente que esteja comprando um produto.

c) Sistemas de recomendação tenderiam a melhorar a lealdade dos clientes aos *sites* das empresas de comércio virtual adicionando valor no relacionamento entre o site e o cliente. Quanto mais os clientes usam estes sistemas, ensinando o que eles gostam, mais leais eles seriam ao *site*.

Apesar de difundido em comércios virtuais, há uma gama de serviços nos quais os sistemas de recomendação ainda podem ser utilizados. Por exemplo, as empresas de televisão a cabo no Brasil oferecem pacotes com mais de 100 canais para seus clientes. Considerando que cada programa tenha duração média de uma hora, o cliente tem a opção de escolher mais de 2.400 programações para assistir por dia. Consequentemente, o cliente tenderá a ter dificuldades de escolher o programa desejado tendo um razoável universo de opções.

Dentro desse contexto, os sistemas de recomendações podem facilitar a escolha do cliente sugerindo programas baseado no tipo de programa a que ele frequentemente assiste ou baseado na avaliação dos outros clientes similares a ele.

Por intermédio da interatividade das TVs digitais, os sistemas de recomendações podem ser aplicados, melhorando a qualidade do serviço oferecido e o tempo de procura do programa desejado.

2.4 Interface de recomendações

SCHAFER *et al.*(1999) apresentou sete meios para apresentar as recomendações aos seus usuários. O método selecionado dependerá de como um *site* de comércio virtual quer que o cliente utilize as recomendações.

a) Alguns sites fornecem aos clientes recomendações baseadas diretamente nos comentários dos outros clientes. Isso ajuda ao empreendedor a ganhar mais dinheiro, pois fornece informações imparciais no serviço ou produto que está sendo vendido.

b) O relacionamento entre itens similares podem ajudar a sugerir itens que o cliente esqueceu-se de comprar, ou que está no subconsciente dele.

c) Uma característica de alguns sistemas de recomendação é a funcionalidade no qual os clientes podem avaliar a qualidade do produto, caso alguém se interesse por ele, o cliente poderá ver qual a média das notas das outras pessoas que avaliaram o produto.

d) Recomendações podem ser entregues diretamente aos clientes por intermédio de e-mails. No entanto, para implementação desse meio, é necessário analisar a política *anti-spam* e a privacidade do cliente.

e) Uma forma de aprimorar a qualidade das recomendações é aumentando a base de conhecimento desses sistemas. Por meio da navegação dos clientes nos produtos dos sites, os sistemas podem

recomendar itens que estão relacionados aos produtos que ele navegou.

f) Uma vez que o site aprendeu detalhes sobre os gostos do cliente, ele é capaz de fornecer ao cliente uma lista personalizada e de tamanho limitado, contendo itens de que ele talvez goste.

g) Uma variação menos restritiva de uma lista limitada é permitir ao cliente continuar olhando para itens altamente recomendáveis que fossem interessantes para ele.

2.5 Estrutura do Sistema de Aprendizado

Os sistemas de recomendação utilizam-se de estruturas lógicas para prever ou recomendar itens aos seus clientes. As avaliações realizadas pelos usuários podem ser representadas em um formato de tabela por meio de uma matriz n x m. Uma matriz é uma representação tabular de um conjunto de número como uma coleção de linhas e colunas. As linhas dessa matriz correspondem aos usuários e as colunas, aos itens que foram avaliados por esse conjunto de usuários. Caso a célula da matriz esteja vazia, conclui-se que o usuário não avaliou aquele item. Os sistemas de recomendação baseado em filtro colaborativo tentam prever a nota que será dada por um usuário a certo item tentando relacionar as células da matriz.

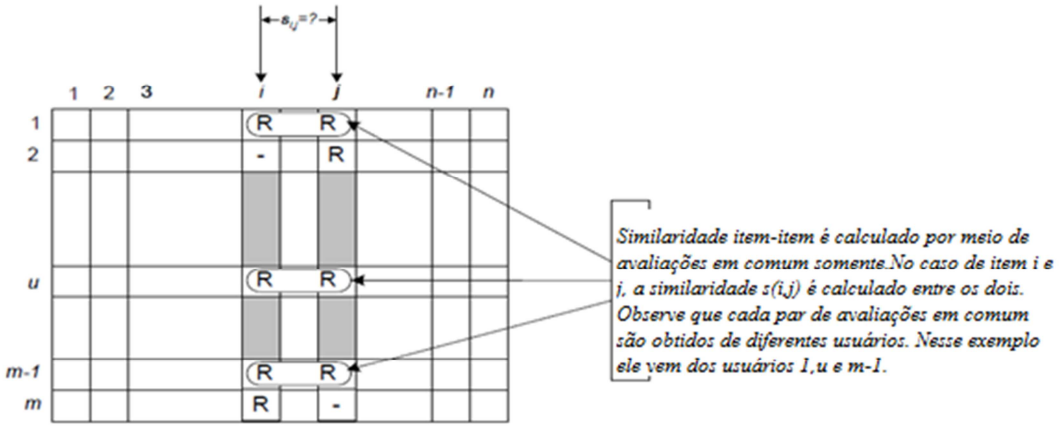

Figura 2 - Similaridade retirada de (SARWAR, 2001)

2.6 Fatores de Qualidade

Os sistemas de recomendação são suscetíveis a alguns fatores qualitativos que afetam a qualidade tanto da previsão quanto da recomendação aos usuários do sistema (RASHID *et al.*, 2005). Dentre elas, pode-se citar:

a) Número de notas: caso o usuário avalie uma quantidade considerável de itens, ele tem maiores chances de ter usuários similares a ele. Além disso, ele pode ser útil para outros usuários que estão procurando por recomendações.
b) Grau de similaridade com outros usuários: essa medida avalia o quão similar um usuário concorda com a opinião dos outro. Por meio dessa medida, é possível analisar o quanto a avaliação do usuário para um item oscila em relação aos outros.
c) Poucas avaliações de um item: essa medida é similar ao *Inverse Document Frequency*, que penaliza os itens mais frequentemente avaliados por não serem adequados para discriminar padrões de utilização e preferências dos usuários.
d) Grau de similaridade com seus vizinhos: os vizinhos mais similares são utilizados para prever uma avaliação para o usuário. O alto valor desta

medida pode ser analisado de duas formas: o usuário pode ser visto com uma pessoa que influência outras pessoas ou como uma pessoa que é influenciada por outras.

e) Muitos itens com uma considerável quantidade de avaliações: se a soma das avaliações dos itens mais populares for alta, o usuário tem grande chance de ter itens em comum com vários usuários.

f) A dispersão das avaliações de um item: a multiplicação da popularidade de um filme com sua dispersão demonstra uma medida que tenta balancear a popularidade e variância.

2.7 Sistemas baseado em conteúdo

Os sistemas de recomendação baseados em conteúdo (MAES, 1994, LIEBERMAN, 1997) usam somente a preferências de usuário ativo, por meio de recomendações baseadas nos itens de que o usuário gostou no passado. Os algoritmos desses sistemas captam a preferência do usuário, filtram as informações e selecionam novos itens que combina com as suas preferências. Esses sistemas provaram ser eficazes na busca textual em documentos relevantes por meio de técnicas como agentes inteligentes e visualização da informação (HERLOCKER, 1999).

Fundamentado no conceito que os indivíduos são capazes de formular consultas às quais expressam seus interesses ou informações necessárias para busca de itens, esses sistemas, frequentemente, indexam os dados dos documentos usando todo o texto ou somente o resumo do documento para efetuar as recomendações (SARWAR, 1998). Este tipo de sistema não sofre com o problema de um item ser necessariamente avaliado por pelo menos um vizinho para ser recomendado, logo o item pode ser recomendado mesmo que ele não tenha sido avaliado anteriormente.

O sistema é suportado na correlação entre o conteúdo do item e a preferência do usuário. Porém, a filtragem baseado em conteúdo tem algumas limitações:

a) Os atributos dos itens precisam ser atribuídos manualmente. A tecnologia atual, ainda não permite eficazmente analisar as informações relevantes dos itens automaticamente. (SHARDANAND *et al.*,1995)

b) Não há métodos dentro desta técnica que permita a recomendação de itens descobertos por acaso. O sistema, frequentemente, recomenda mais itens que já foram visualizados anteriormente pelo usuário.

c) Os métodos não são capazes de filtrar os itens baseados na qualidade, no estilo ou no ponto de vista do usuário ou item.

d) Às vezes, torna-se difícil encontrar palavras capazes de identificar o interesse do usuário (HOFMANN,2004).

e) Recomendações são geralmente realizadas pela qualidade dos itens e não pelas propriedades do objeto (SHARDANAND *et al.*,1995).

Para viabilizar recomendações baseadas em conteúdo, o uso de taxonomias facilita a identificação de classes de usuários e conteúdos que possibilitem relacioná-los. A taxonomia classifica conteúdos de forma hierárquica e exclusiva. Há vantagens nesse tipo de classificação, entre elas: a melhor organização bem como pesquisas mais eficientes do conteúdo.

Outra abordagem pode ser descrita como a utilização da folksonomia, na qual os próprios usuários definem palavras-chave para conteúdo. Nessa abordagem, o conteúdo não é classificado em grupos pré-definidos, podendo ser mais difícil de definir uma relação entre usuário e conteúdo, devido à esparsabilidade de classes, já que as palavras-

chave são definidas de modo livre pelos usuários. Pode-se entender como "folksonomia" como um sinônimo de etiquetas colaborativas (*collaborative tagging*), classificação social (*social classification*), indexação social (*social indexing*), etiquetas sociais (*social tagging*), marcação social de favorito (*social bookmarking*), entre outros nomes.

O problema da classificação e delegação aos próprios usuários contorna esse problema recorrente na taxonomia. Caso o usuário tenha poucas avaliações ou compras registradas no sistema, os algoritmos baseados em conteúdo permanecem escaláveis e com o desempenho aceitável (LINDEN *et al.*,2003). Para os usuários com uma quantidade de avaliações relevantes, é impraticável realizar uma consulta em todo o conteúdo classificado. Para solucionar essa questão, alguns algoritmos simplificam o conjunto de dados da consulta ou resumem os dados, o que leva, frequentemente, a agravarem a qualidade das recomendações (LINDEN *et al.*,2003).

2.8 Medida para utilização de sistemas de recomendação

Predictive utility (KONSTAN *et al*.,1997) refere-se ao valor de ter previsões para um item antes de investir tempo ou dinheiro consumindo o item. Um alto valor dessa medida significa que os usuários ajustarão suas decisões baseados nas previsões, já um baixo valor reduz o efeito das previsões das decisões dos usuários. Em suma, *predictive utility* é uma função que quantifica a quantidade de itens desejáveis e indesejáveis na qualidade das previsões (KONSTAN *et al*.,1997).

A análise de custo-benefício compara o valor de consumir um item desejável, o custo de perder um item desejável, o valor de rejeitar um item indesejável e o custo de consumir um item indesejável. Um falso-positivo é um item indesejável classificado pelos sistemas de recomendação como desejáveis. O custo de perder um item desejável representa o risco envolvido em fazer previsões. Já o valor de prever um item desejável

e filtrar os itens indesejáveis representa os benefícios da previsão. Logo, predictive utility é a diferença dos benefícios potenciais e o risco das previsões.

Caso haja uma grande quantidade de itens desejáveis (acima de 90% do total de itens), a filtragem de itens não agregará valor, pois há poucos itens que poderão ser rejeitados. A probabilidade de encontrar um item desejável é alta independente da utilização de uma filtragem. A filtragem pode agregar muito valor, se o total de itens desejáveis for baixo (abaixo de 1%), pois a rejeição de itens indesejáveis é alta (KONSTAN *et al*.,1997).

2.9 Problemas de Previsão

Há dois tipos mais comuns de problema relativo à previsão. A primeira chamada de previsão forçada (HOFMANN, 2004) envolve prever o valor para um item particular dado o usuário. Esse problema ocorre normalmente quando um item é apresentado a um usuário como uma recomendação, e o objetivo é antecipar qual seria o interesse do usuário.

Já na previsão livre (HOFMANN, 2004), o usuário está no controle para selecionar o item e há dois objetivos: prever o que o usuário selecionará e, opcionalmente, como ele avaliará o item.

2.10 Exemplos de aplicações que utilizam sistema de recomendação

2.10.1 1 – Last.fm

A Last.fm é um serviço de recomendações musicais. O aplicativo Scrobbler envia para a Last.fm uma mensagem para informar sobre a música que a pessoa está ouvindo no momento. Ele o ajuda a saber, quais músicas as pessoas ouvem, qual a frequência que escutam, qual a preferência das música e até quantas vezes ouviu-se um artista em um

período específico de tempo. Através dos Scrobbler, podem ser feitas diariamente recomendações personalizadas para cada ouvinte da Last.fm.

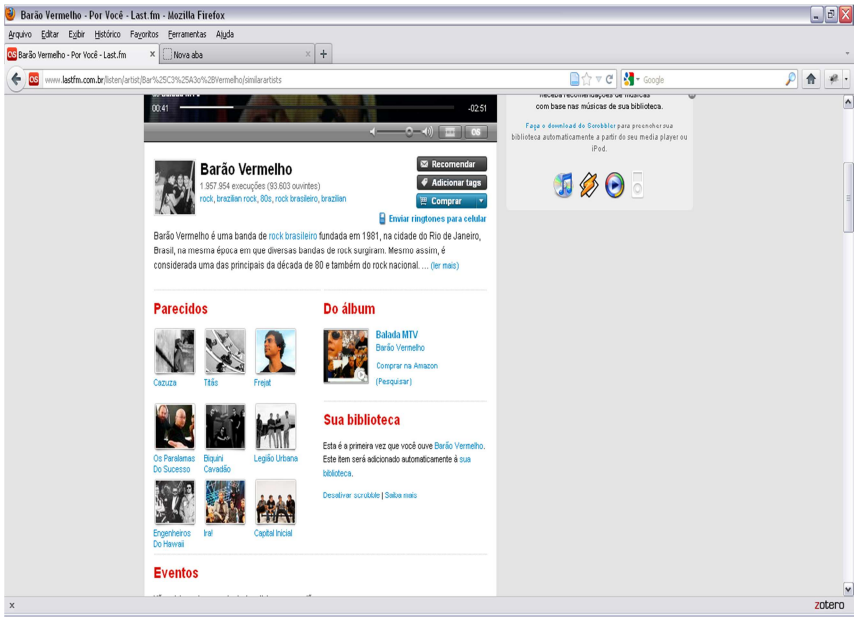

Figura 3 – Site Last.fm

2.10.2 2 – Amazon.com

A *Amazon* utiliza um algoritmo de recomendação personalizada na sua loja online para cada um dos clientes.

Figura 4 - Site Amazon.com

2.10.3 3 – Netflix

Os clientes da Netflix podem assistir, online, a filmes e a séries transmitidos pela Internet. Eles podem avaliar os filmes e séries a que assistiram na televisão ou no site da Netfix para receber recomendações de vídeos de que, talvez, eles gostem. Além da avaliação dos filmes, podem-se julgar os gêneros dos filmes que o cliente aprecie a fim de aperfeiçoar as recomendações.

Capítulo 3 - Sistemas de Recomendação baseados em filtro colaborativo

As preferências das pessoas não são aleatoriamente distribuídas, geralmente há tendências e padrões relacionados à preferência de um indivíduo assim como em grupos de pessoas (SHARDANAND *et al.*,1995). Em vez de um usuário perguntar para seus amigos sobre algum item, os sistemas baseados em filtro colaborativo (GOLDBERG *et al.*,1992) automatizam o processo de recomendação.

A ideia básica de um filtro colaborativo é manter um perfil para cada um dos seus clientes e um registro de suas preferências em certos itens. Então, compara-se o perfil do usuário ativo com o de outros usuários para obter o grau de similaridade com seus vizinhos. O sistema pode, então, considerar um conjunto de perfis mais similares ao usuário ativo e recomendar ou prever itens ao cliente em questão.

O filtro colaborativo facilita as pessoas a fazerem escolhas fundamentadas na opinião de outras (RESNICK *et al.*,1994). O sistema *Tapestry* (GOLDBERG *et al.*,1992) foi um dos primeiros sistemas de recomendação implementados. Ele foi projetado para tratar uma grande quantidade de e-mails, recebendo e filtrando os arquivos. O usuário do sistema poderia avaliar as mensagens e associar anotações dentro das mensagens. As mensagens armazenadas no banco de dados poderiam, então, ser buscadas tanto pelo conteúdo quanto pela opinião das outras pessoas.

A pretensão desse sistema era substituir os sistemas de e-mail da época. O objetivo era suportar tanto os sistemas de filtragem colaborativa quanto o baseados em conteúdo para solucionar o problema do aumento do fluxo de e-mail, filtragem de e-mails irrelevantes. GOLDBERG *et al.* (1992) definiu o filtro colaborativo como um processo em que as pessoas ajudam as outras a filtrar os documentos através de suas ações.

Os usuários colaboram no sentido de que cada ação efetuada poderá melhorar o desempenho de todo o sistema. A ideia dos sistemas de filtragem colaborativa é que, se dois usuários avaliam N itens similares, eles compartilham gostos similares (GOLDBERG *et al.*, 2001).

LEE *et al.* (2009) relata que o sistema *Tapestry* foi elaborado para trabalhar em pequenas redes, pois os usuários precisavam estar familiarizados com a preferência e opinião dos outros pertencentes à rede. As técnicas de filtragem colaborativa usadas no sistema *Tapesty* não eram automatizadas, requerendo um utilizador para construir consultas complexas. As características citadas acima fizeram que o sistema tivesse suas limitações quanto à sua aplicabilidade.

LEKAKOS (2006), então, menciona três pilares para que a abordagem baseada em filtragem colaborativa tenha sucesso:

a) Várias pessoas precisam participar para facilitar ao sistema encontrar pessoas com preferências similares.
b) Há de haver uma forma fácil das pessoas conseguirem expressar ao sistema suas preferências.
c) O algoritmo tem que ser capaz de encontrar pessoas com interesses similares.

Nos sistemas de comércio virtual, para que os sistemas possam realizar a tarefa de sugerir itens aos seus clientes, o servidor Web comunica com o sistema de recomendação para escolher o produto que será sugerido ao cliente. Este sistema de recomendação baseado em filtragem colaborativa usa a base de dados que contém as avaliações dos usuários feitas nos produtos e a similaridade entre os itens para encontrar uma vizinhança e realizar suas recomendações. As técnicas de filtragem colaborativa evitam a necessidade de conhecimento do domínio no qual será aplicado o sistema.

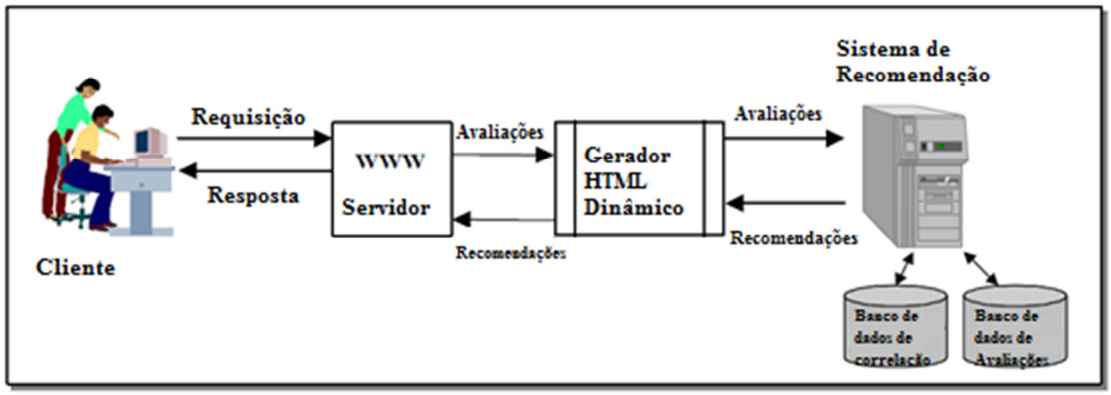

Figura 5 - Arquitetura de um Sistema de Recomendação retirado de (SARWAR, 2000)

Os sistemas de filtragem colaborativa, por capturarem a interação entre os usuários e os itens, produzem notas de recomendação devido aos usuários ou aos itens, independente da relação direta entre eles. Há, no entanto, uma tendência para alguns usuários avaliarem melhores que outros alguns itens, e esses itens de receberem notas maiores que outros itens (KOREN, 2009a).

Os sistemas de filtragem colaborativa fornecem três vantagens na filtragem de informações que não são fornecidas pelos sistemas baseado em conteúdo (HERLOCKER *et al.*,1999).

a) Suporte para filtrar conteúdo de difícil análise por processos automatizados;

b) habilidade de filtrar itens baseado na qualidade e no gosto dos usuários;

c) habilidade de fornecer recomendações ao acaso.

Porém, os algoritmos de filtragem colaborativa não são universalmente aceitos, pois os usuários não querem gastar muito tempo avaliando itens. O problema dos sistemas de filtragem colaborativa é predizer se o usuário irá gostar ou não de um item que ainda não foi avaliado por ele. Essa predição é definida por intermédio de um conjunto de preferências feitas por uma comunidade de utilizadores (HERLOCKER *et al.*,1999).

A preferência pode ser realizada por meio de um relato do usuário (exemplo: O Cliente avalia o filme com nota cinco) ou por medições implícitas que são produzidas por atividades realizadas pelo utilizador. Uma técnica para analisar dados implícitos é mapear as medidas implícitas tais como o tempo gasto lendo um documento. Pode-se definir o tempo gasto e relacioná-lo à preferência do usuário Por exemplo, caso o usuário demore lendo um documento, a preferência pelo documento pode ser avaliada como alta. Notas altas estão geralmente relacionadas ao grande interesse do usuário e notas baixas, ao pequeno interesse. No entanto, isso não quer dizer que o documento agregou valor ao usuário.

Em um filtro colaborativo, a matriz é geralmente muito esparsa, pois os usuários normalmente avaliarão uma pequena proporção do total de itens (HERLOCKER *et al.*,1999). Espera-se, então, que os algoritmos de filtragem colaborativa sejam exatos e eficientes, isto é, após as realizações de uma recomendação, o usuário deverá avaliar

bem o item e, em termos computacionais, ser eficaz para executar várias recomendações por segundo em um universo de milhões de usuários e itens (GOLDBERG *et al.*,2001).

O sistema *LikeMinds* (GREENING,1997) foi projetado para recomendar itens por meio de outros usuário do sistema, chamado Mentores, que reduzem a necessidade de uma grande quantidade de usuários similares ao usuário ativo. Mentores são usuários que possuem uma grande similaridade com o usuário ativo. AGGARWAL *et al.* (2001) demonstrou duas possíveis maneiras de implementar o algoritmo. Na primeira forma, procura-se o mentor com a maior similaridade e que tenha avaliado o item ativo, e então utiliza a nota dada pelo mentor. Na outra forma, diferentes pesos são atribuídos a cada mentor conforme a similaridade com o usuário ativo. Porém, a nota dada pelo mentor é transformada numa escala do usuário. A nota dada *r(k,j)* pelo mentor no item j é transformada em uma previsão de nota *s(k,j)*. O *min(i)* resulta na menor nota dada pelo usuário i e *max(i)*, mostra a maior nota.

$$s(k,j) = \min(i) + \frac{(r(k,j) - \min(k)) * (\max(i) - \min(i))}{\max(k) - \min(k)}$$

O resultado final será

$$L(i,j) = \frac{\sum_{AS(i,k) > 0} AS(i,k) * s(k,j)}{\sum_{AS(i,k) > 0} AS(i,k)}$$

Sendo AS(i,k) a similaridade entre os usuário i e k.

No entanto, AGGARWAL *et al*. (2001) menciona que a técnica empregada pelo sistema *LikeMinds* utiliza mais avaliações que o necessário para realizar previsões com qualidade, embora os detalhes da técnica não tenham sido revelados.

As previsões necessitam do cálculo da similaridade entre os usuários ou itens. Para isso, necessita-se de que os usuários tenham um conjunto de itens avaliados em

comum, todavia, na prática, o conjunto de avaliações feitas por cada usuário é pequeno. Geralmente, cada usuário avalia uma quantidade mínima de itens, normalmente, na ordem de 10 a 20 itens, permitindo o usuário a participar da filtragem colaborativa (GREENING,1997). Caso não seja empregada uma quantidade mínima de itens a serem avaliados, os usuários não serão capazes de aproveitar os benefícios que a filtragem colaborativa oferece.

Através da utilização de uma quantidade mínima de itens, GREENING (1997) abordou a utilização de gráficos direcionados, no qual os usuários são os nós do gráfico, e as arestas correspondem à possibilidade de previsão. Para eficácia dessa abordagem, necessita-se, então, de que mais itens sejam avaliados entre o usuário ativo e os outros usuários que terão suas avaliações utilizadas para compor o valor da previsão. Os pares de usuários precisam ser similares, exceto pelos usuários que avaliam sempre muito bem os itens. Além disso, essa abordagem leva em consideração os pares de usuário que avaliaram itens que sejam muito semelhantes ou o contrário.

A previsão da nota data ao item j pelo usuário i é realizada por meio da média dos pesos, que por intermédio de caminhos direcionados do gráfico levam a certos usuários. Cada um dos caminhos direcionados conecta o usuário i até outro usuário k que avaliou o mesmo item j. Nenhum usuário neste caminho poderá ter avaliado o item j, apenas o último que será o fim do caminho. Uma desvantagem dessa técnica, além de necessitar de uma quantidade mínima de avaliações, é que o algoritmo não é direto, pois é preciso caminhar em um gráfico para realizar a previsão, ocasionando uma demora maior na previsão de uma recomendação.

3.1 Ordem de grandeza

A filtragem colaborativa é de ordem de grandeza alta, pois considerando N o número de usuário e M o número de itens, no pior caso, a geração do filtro colaborativo

é da O(MN). A esparsidade leva a ordem de grandeza a ser *O*(M+N). Procurar cada cliente é aproximadamente *O*(M) já que são poucos itens que cada cliente avalia, porém alguns usuários avaliam uma quantidade considerável de itens na *O*(N).

3.2 Algoritmos baseados em Memória

Algoritmos baseados em memória (BREESE *et al.*, 1998) operam em todo o banco de dados do usuário para fazer previsões. Esses algoritmos são mais simples e intuitivos evitando complicações no estágio da construção do modelo comparado aos algoritmos baseados em modelo, além de poderem ser utilizados em várias situações reais. No entanto, há, pelo menos, quatro desvantagens listadas abaixo (HOFMANN, 2004):

a) A acurácia obtida por esses algoritmos nem sempre é ótima.
b) Nenhum modelo estatístico explícito é elaborado, logo nenhum conhecimento é construído por meio dos comportamentos dos usuários contidos nos registros históricos.
c) Esses algoritmos não são escaláveis, frequentemente utilizam amostras dos dados para tratar o problema.
d) Problemas de reusabilidade do algoritmo para realizar outras tarefas.

3.2.1 Filtro colaborativo baseado em usuário

Cada usuário é representado por um par item-avaliação, e pode ser resumido por uma tabela usuário-item que contém as avaliações que foram fornecidas pelos usuários nos itens.

Item	Item1	Item2	… …	Itemn
Usuário 1	R11	R12	… …	R1n
Usuário 2	R21	R22	… …	R2n
… …	… …	… …	… …	… …
Usuário m	Rm1	Rm2	… …	Rmn

Figura 6 – Tabela UsuárioXItem retirado de (GONG, 2010)

Há várias técnicas para calcular a similaridade, mas a ideia comum é calcular a similaridade entre os usuários usando alguma medida para recomendar itens baseados nessa similaridade. Os filtros colaborativos que usam o algoritmo de similaridade entre os usuários são chamados de filtros colaborativos baseados em usuário (LI *et al.*, 2005, LEKAKOS *et al.*,2006).

Para cada item que será previsto, os vizinhos que avaliaram o item com maiores similaridades com o usuário ativo são usados para o cálculo da previsão. Para cada item, um usuário poderá ter diferentes vizinhos. Portanto, todos os usuários serão examinados para o cálculo de uma vizinhança. O primeiro passo do algoritmo é calcular o peso de todos os usuários considerando a similaridade com o usuário ativo. O aumento do número de usuários e itens na base de dados aumenta a complexidade do cálculo de similaridade, oferecendo sérios problemas de escalabilidade nas técnicas de filtragem colaborativo baseadas em item (LI, 2005).

3.2.2 Soma dos pesos baseado em Usuário

Esse método calcula a previsão de um usuário i para um item j por meio da soma das notas dadas nos item j pelos usuários similares i. Cada avaliação feita pelo usuário k recebe um peso relevante equivale à similaridade entre os itens i e k.

3.2.3 Filtro colaborativo baseado em item

As técnicas de filtragem colaborativa baseada em itens (SARWAR *et al.*,2001) analisam a matriz usuário-item para identificar relações entre diferentes itens e usá-las indiretamente para sugerir recomendações para os usuários.

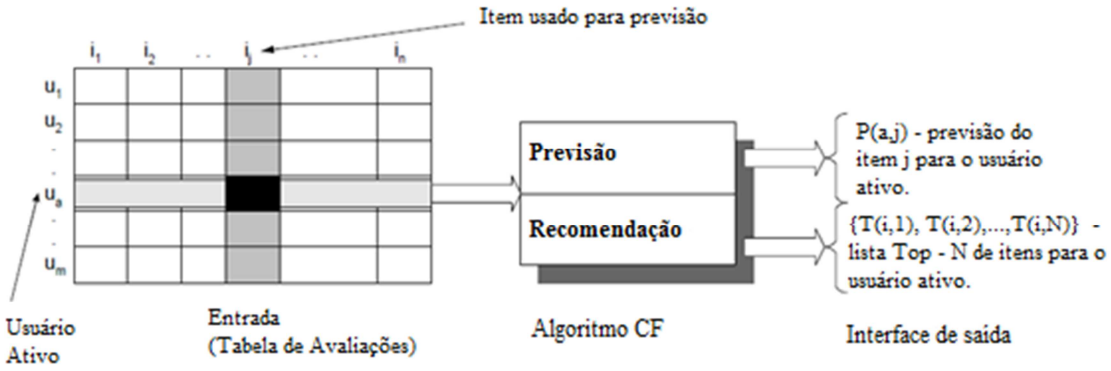

Figura 7 - Processo do Algortimo CF retirado de (SARWAR, 2001)

A abordagem baseada em itens para filtro colaborativo identifica similaridades entre dois itens comparando as notas dos usuários nesses itens. Nessa abordagem, as notas produzidas em tempos distintos possuem o mesmo peso. Logo, mudanças de interesse dos usuários não são levadas em consideração (DING *et al.*,2005).

A construção da matriz item-item é realizada por intermédio da comparação de todos os pares de itens, porém, frequentemente, encontram-se vários itens que não possuem avaliações realizadas por clientes em comum, o que torna essa abordagem ineficiente em termos de processamento de tempo e uso da memória (LINDEN *et al.*,2003).

A abordagem baseada em itens procura dentro do conjunto de itens que o usuário ativo avaliou e calcula a similaridade de cada um dos itens dentro do conjunto com o item alvo. Após essa etapa, selecionam-se os itens mais similares e suas respectivas similaridades. O cálculo da previsão é feito por meio da média das avaliações do usuário ativo nos itens similares considerando o peso da similaridade.

SARWAR *et al.* (2001) relata por meio de seus experimentos que o algoritmo baseado em item teve resultados melhores comparados ao baseado em usuário, no que diz respeito a previsões das avaliações, no entanto a melhora na acurácia das previsões não é significativa.

LINDEN *et al.* (2003) propôs um algoritmo interativo que fornece uma abordagem com o objetivo de melhorar a qualidade da similaridade de um produtos com outros produtos relacionados.

3.2.4 Soma dos pesos baseado em Item

Caso o algoritmo utilizado seja baseado em itens, o cálculo é feito por meio da soma das notas dadas pelo usuário i aos itens similares i. Cada avaliação feita pelo item k recebe um peso relevante corresponde à similaridade entre os itens j e k.

$$p_{i,j} = \frac{\sum_{todos\ os\ itens\ similares,N}(s_{i,N} * R_{u,N})}{\sum_{todos\ os\ itens\ similares,N} |s_{i,N}|}$$

Sendo $s_{i,j}$ equivale a similaridade do item i com o item j e $R_{u,j}$ a avaliação feita pelo usuário u com o item j.

3.3 Filtro Colaborativo baseado na vizinhança

Nessa técnica, um subconjunto de usuários será selecionado apoiado na similaridade com o usuário ativo, e um peso agregado será usado para predizer a preferência do usuário ativo.

Nas técnicas baseadas nessa abordagem, há uma atribuição de peso para todos os usuários considerando a similaridade com o usuário ativo. Após essa atribuição, os usuários que tiveram a maior similaridade com o usuário ativo serão selecionados para definir a predição que levará em consideração os pesos que foram atribuídos aos usuários.

Filtro Colaborativo baseado na vizinhança pode ser dividido em três etapas (HERLOCKER *et al.*,1999) :

a) colocar peso em todos os usuários considerando a similaridade com o usuários ativos;

b) definir um subconjunto de usuários para usar como um conjunto de preditores;

c) normalizar as notas e calcular a predição por meio de combinações dos pesos dos usuários selecionados.

Essas três etapas podem ser implementadas sobrepostas ou até mesmo um pouco diferente da listada acima.

GroupLens (RESNICK *et al.*,1994,KONSTAN *et al.*, 1997) foram os primeiros a adicionarem um sistema com Filtro Colaborativo baseado na vizinhança

A fórmula usada para encontrar a previsão da nota *p(i,j)* para o usuário i e item j:

$$p(i,j) = \overline{r_i} + k \sum_{k=1}^{n} w(i,p)(\overline{r_{pj}} - \overline{r_p})$$

No qual $\overline{r_i}$ é a média das avaliações feitas pelo usuário i, e o fator de normalização garante que a soma absoluta do peso não ultrapasse o valor um. O peso *w(i,p)* pode ser uma correlação ou similaridade entre os usuário i e p.

3.4 Algoritmos baseados em Modelo

Os algoritmos baseados em modelo (BREESE *et al.*,1998) usam o banco de dados do usuário para elaborar um modelo, no qual é usado para fazer previsões. Nessa categoria de algoritmos, utilizam-se métodos estatísticos para calcular o valor esperado das avaliações feitas pelos usuários.

Probabilistic Latent Semantic Analysis (pLSA) é um algoritmo fundamentado em um modelo proposto por HOFMANN (2004). O algoritmo introduz uma variável z com estados para cada par de usuário-item a fim de que o usuário e o item sejam modelados

de forma independente. O número de estados da variável Z é uma entrada para o modelo e seu estado será interpretado como um tipo de usuário. Cada indivíduo pertence a um ou mais grupos de usuários, com uma distribuição de probabilidade P(Z|u).

HOFMANN (2004) desenvolveu um método chamado *Expectation Maximization* (EM) como aprendizado dos coeficientes P(Z|u). Inicialmente, elabora-se um modelo de notas dos usuários para, então, fornecer recomendações de itens.

Dado que o valor das avaliações seja número inteiro compreendida entre 0 a m

$$P(p,m) = \sum_{z} P(p)\,P(z|p)P(m|z)$$

Sendo z uma variável oculta, a soma de todas as possibilidades da distribuição. Os parâmetros P(Z|P) correspondem ao processo estocástico de escolha da variável z tendo o usuário P.

A existência de uma variável z introduz a motivação, no caso de um sistema de recomendação de filme, de uma pessoa assistir a um filme. A pessoa p escolhe a variável z, que determina que o filme m será visto. Assumindo o conhecimento sobre z, a escolha do filme é independente do usuário. A expressão é a probabilidade do usuário ativo avaliar o item j com uma valor específico dado as avaliações anteriores feitas por ele (SCHEIN et al., 2002).

3.4.1 Modelo de Agrupamento

Técnicas de agrupamento baseados em usuários agrupam usuário que tem preferências similares. Há determinados grupos ou tipos de usuários que possuem um conjunto de preferências e gostos em comum. Esses usuários são agrupados formando conjuntos. Esses conjuntos são usados para identificar as preferências de certo grupo de

pessoas e as previsões são feitas por meio da média das opiniões dos usuários no agrupamento.

Algumas dessas técnicas associam o usuário em mais de um grupo para realizar a previsão e atribuem um peso a cada um dos grupos conforme o grau de participação do usuário. Técnicas de agrupamento geralmente produzem recomendações menos personalizadas que o algoritmo baseado em vizinhança (BREESE *et al.*,1998). Um exemplo de um modelo probabilístico para Filtro Colaborativo é o classificador Bayesiano. Classificação é a tarefa de organizar objetos em uma entre diversas categorias pré-definidas.

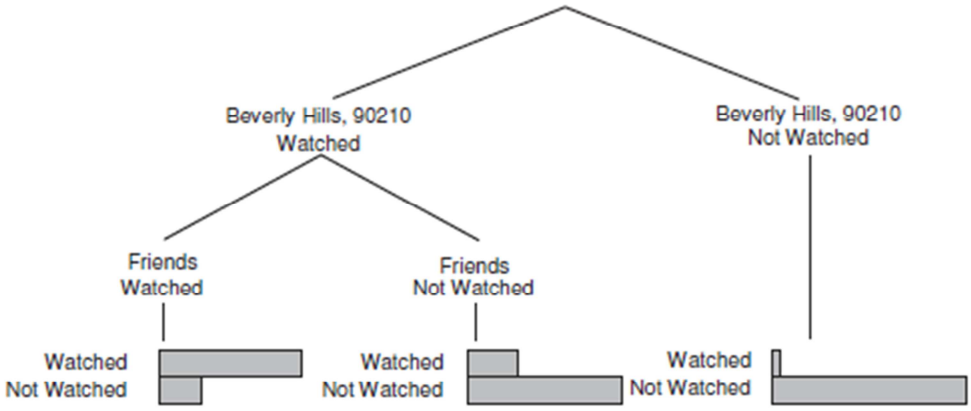

Figura 8 – Árvore de classificação retirada de (BREESE ,1998)

O algoritmo de agrupamento pode gerar partições de tamanho fixo ou fundamentado em algum limiar que pode ser requisitado por um número de partições de tamanho variado.

Uma vez criado o grupo, as previsões para o usuário ativo podem ser feitas pela média das opiniões dos outros usuários naquele grupo. Uma vez que tenha o agrupamento dos usuários realizado, o desempenho pode ser eficaz, já que o número de grupos que serão analisados será menor.

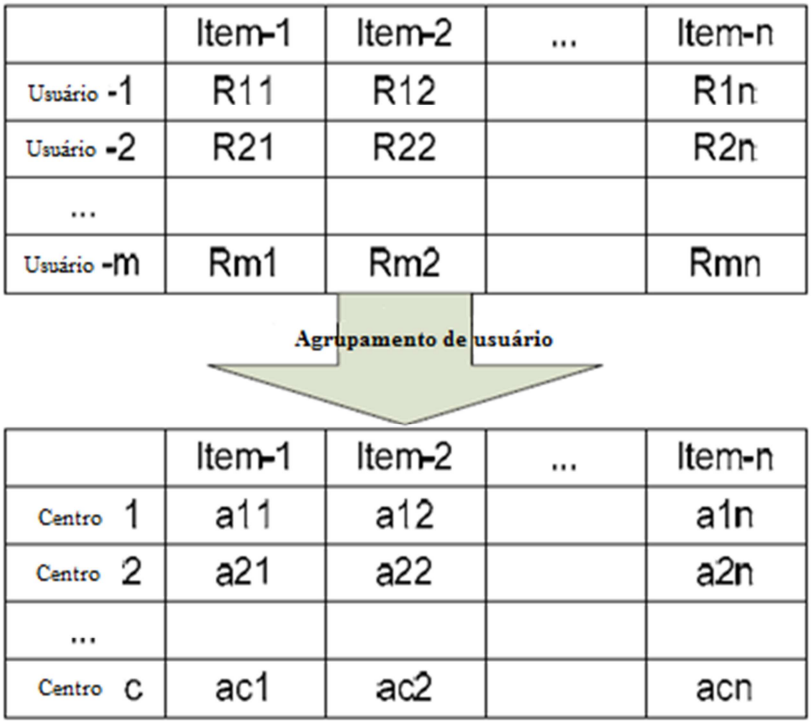

Figura 9 - Agrupamento de usuário retirado de (GONG, 2010)

Técnica de agrupamento de itens identifica itens que parecem ter avaliações similares. A previsão para um item pode ser feita pela média de opiniões de outros itens naquele agrupamento. Ela, então, é dada pela média das avaliações considerando o grau de participação de cada item no grupo.

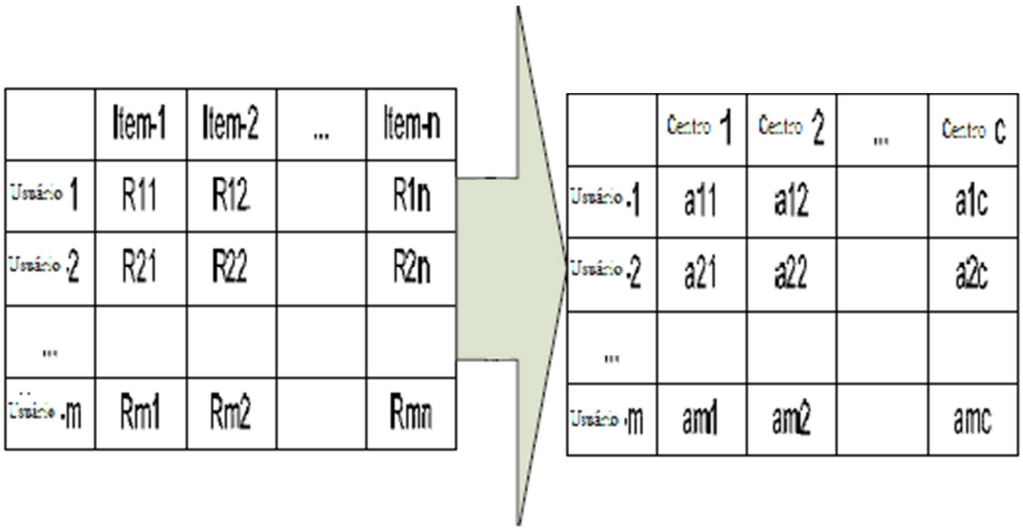

Figura 10 - Agrupamento de itens retirado de (GONG, 2010)

DING *et al.* (2005) propôs um algoritmo que utiliza agrupamento para discriminar vários tipos de itens. Para cada item agrupado, são identificadas as mudanças de

interesse do usuário e introduzido um fator de decaimento de acordo com o comportamento de compra do próprio usuário. O algoritmo assume que o mesmo usuário terá preferências similares e mudanças de interesses similares para itens semelhantes. Por meio de agrupamento, foram discriminados diferentes itens bem como uma grande quantidade de itens similares em cada grupo.

Na execução das recomendações, os sistemas de recomendação que empregam modelos de agrupamento, geralmente, possuem uma melhor escalabilidade e desempenho que os algoritmos de filtragem colaborativa, devido ao fato de que eles comparam o usuário a um número de grupo em vez de comparar como todo o banco de dados (BREESE *et al.*, 1998).

Os grupos dos usuários nem sempre agrupam os usuário mais similares, com isso as recomendações não são tão relevantes (LINDEN,2003). Pode-se melhorar a qualidade das recomendações através de grupos mais granulares, no entanto a classificação torna-se tão complexa como uma técnica de filtragem colaborativa.

A técnica de agrupamento K-means (MACQUEEN,1967) baseia-se no uso de centroide. Inicialmente, define-se K centroides iniciais. Geralmente, o centroide é calculado por meio da média dos pontos de um grupo. Ele é aplicado a objetos em um espaço n-dimensional contínuo.

Cada item ou usuário será atribuído ao centroide mais próximo sendo, então, recalculado o novo centroide considerando os itens ou usuários associados ao centroide. Por definição, o conjunto de itens ou usuários associados a um centroide é um grupo.

Algoritmos K-Means

Selecione K centroides

Repita

Atribua cada item ou usuário ao centroide mais próximo.

Recalcule o centroide de cada grupo

Até a convergência

K-means pode ter inúmeras aplicabilidades. Embora haja múltiplas execuções do algoritmo, ele é eficiente, porém, ele é restrito a dados para os quais exista uma noção do que será o centro do agrupamento. O algoritmo clássico não determina automaticamente o número de grupos, o que torna um problema ao se usar agrupamentos na detecção de elementos estranhos.

3.4.2 Modelo de Rede Bayesiana

A rede Bayesiana suporta o uso de nós. Esses correspondem a um domínio, já o estado de cada nó corresponde a uma possível avaliação de cada item. Aplica-se um algoritmo de aprendizado de redes Bayesianas para procurar sobre vários modelos considerando a dependência de cada item. Cada item terá um conjunto de itens relacionados que serão mais indicados como itens para prever a avaliação desse item.

Redes bayesianas provaram ser adaptáveis em ambientes em que as preferências dos usuários se alteram lentamente, porém esses sistemas não são apropriados nos ambientes nos quais há uma volatilidade nas preferências do usuário (SARWAR *et al.*,2001).

Supondo que os itens sejam independentes, um classificador de Bayes simples avalia a probabilidade condicional de classe.

$$P(X|Y = y) = \prod_{i=1}^{d} P(X_i|Y = y)$$

BREESE *et al.* (1998) mostrou que analisando inúmeras condições, a rede Bayesiana com árvore de decisão e os métodos de correlação(ex: correlação de Person) demonstraram um resultado melhor que o Modelo de Agrupamento e vetores de similaridade. No entanto, a escolha do algoritmo depende da natureza dos dados que estão sendo analisadas não só a natureza de sua aplicabilidade, mas também a disponibilidade das avaliações feitas pelos usuários.

3.5 Regressão

Em vez de usar as avaliações similares como o método das somas dos pesos, o método de regressão emprega avaliações com valores aproximados baseado em um modelo de regressão. Isso ocorre devido ao fato que o cálculo da similaridade, caso utilize a correlação cosseno, poderá equivocar-se no sentido que dois vetores podem estar distantes no espaço euclidiano e, mesmo assim, serem similares.

$$Rn = \alpha Ri + \beta + \varepsilon$$

Sendo N o conjunto de objetos semelhantes ao objeto i e Rn o modelo de regressão linear.

3.6 Média do Usuário

V(i,j) corresponde à avaliação realizada pelo usuário i ao item j. Considerando que *Li* seja o conjunto de itens que o usuário i avaliou. Obtém-se que a média do usuário $\overline{v_i}$ é dada por:

$$\overline{v_i} = \frac{1}{|Li|} \sum_{j \in Li} v_{i,j}$$

Já previsão da avaliação realizada pelo usuário ativo a ao item j, $p_{a,j}$ será dada pela fórmula abaixo:

$$p_{a,j} = \overline{v_a} + k \sum_{i=1}^{n} w(a,i)\,(v_{i,j} - \overline{v_i})$$

A variável n na fórmula acima refere-se ao número de usuários que serão usados para prever a avaliação para o usuário ativo. A função *w(a,i)* que é peso atribuído, refere-se a distância, correlação ou similaridade entre os usuário a e i. A variável K representa o fator de normalização a fim de que a soma dos pesos absolutos não seja maior que um. Em outras palavras, o objetivo da normalização é fazer o conjunto inteiro de valores possuir uma determinada propriedade.

3.7 Similaridade

Uma etapa crítica no processo de filtragem colaborativa é o cálculo de similaridade entre os objetos. Quanto mais dados são usados para comparar a opinião dos dois usuários, mais confiável será o cálculo da similaridade entre os usuários (GOLDBERG *et al.*,1992).

GOLDBERG *et al.* (1992) propôs uma forma de decaimento da correlação entre dois usuários. Caso o número de avaliações seja menor que cinquenta itens avaliados em comum, divide-se o decaimento por n/50. Sendo que n é o número de itens em comum avaliado pelos dois usuários. Se as avaliações forem maior que 50, não haverá um fator de decaimento.

Avaliações em certos itens demonstram-se ser ineficientes para distinguir as preferências das pessoas. Itens que são frequentemente bem avaliados ou itens que são comumente mal julgados são incapazes de refletir o compartilhamento de interesse das

pessoas. Caso todas as pessoas avaliem bem um item, não é possível diferenciar dentro desse conjunto de pessoas, os seus interesses.

Por meio da incorporação de um fator de peso na correlação de Pearson, é possível aumentar a influência dos itens com maior variação das notas e reduzir as com menor variação. Porém, esta abordagem não apresentou resultados positivos como esperado.

3.7.1 Pearson

A medida de similaridade Pearson utiliza-se do coeficiente de correlação. O valor da correlação sempre fica na faixa entre -1 e 1. Esse valor, dentro de um sistema de recomendação, indica o quanto o usuário ativo concorda com cada um dos outros usuário nas avaliações que os dois realizam em comum. Primeiramente, é necessário isolar todos os itens que foram avaliados pelos dois usuários os quais estão sendo relacionados.

A equação abaixo refere-se à equação de Pearson, que considera a covariância entres as variáveis x e y e a variância dessas variáveis.

$$\text{Pearson} = \frac{\text{cov}(\text{x}, \text{y})}{\sigma_x * \sigma_y}$$

A equação acima pode ser transformada na equação abaixo:

$$\text{Pearson} = \frac{\sum(Ux - \overline{Ux}) * (Uy - \overline{Uy})}{\sqrt{\sum(Ux - \overline{Ux})^2} * \sqrt{\sum(Uy - \overline{Uy})^2}}$$

A variável $\overline{Ux}$ é a média das notas do usuário ativo x, $\overline{Uy}$ é a média das notas do outro usuário y. *Ux* corresponde a nota dada pelo usuário x.

Quadro 1 - Matriz usuário x item

Usuário\Item	Filme1	Filme2	Filme3
Usuário X	2	4	3
Usuário Y		1	2
Usuário Z		1	3

A média do Usuário X, $\overline{Ux}$ será (2+4+3)/3 = 3

Já a média do Usuário Y, $\overline{Uy}$ será (1+2)/2 = 1,5

Supondo que o usuário X seja o usuário ativo, então, para calcular a similaridade entre o Usuário X e o Usuário Y; procede-se assim:

$$\text{Pearson} = \frac{(4-3)(1-1{,}5) + (3-3)(2-1{,}5)}{\sqrt{((4-3)^2 + (3-3)^2)} * \sqrt{((1-1{,}5)^2 + (2-1{,}5)^2)}}$$

$$\text{Pearson} = \frac{-0{,}5}{\sqrt{1} * \sqrt{0.5}} = \frac{-0{,}5}{0{,}7} = -0{,}71$$

O intervalo do coeficiente indica se caso o valor seja -1 que o relacionamento entre os dois objeto será linearmente negativa, se for 0, indica que não há relacionamento, e +1, indica um relacionamento positivo entre os usuários.

3.7.2 Pearson Restringido

O algoritmo de Person Restringido (SHARDANAND *et al.*, 1995) é uma variação do algoritmo de Pearson que leva em consideração as avaliações positivas e negativas feitas pelo indivíduo. Empregando uma escala de avaliação de 1 até 5, caso a avaliação seja maior que 3, então ela será considerada uma avaliação positiva, caso contrário será

negativa. Por esse exemplo, considera-se a equação abaixo como exemplo do Pearson Restringido:

$$sim(x,y) = \frac{\Sigma(Ux-3)*(Uy-3)}{\sqrt{\Sigma(Ux-3)*\Sigma(Uy-3)}}$$

O cálculo correlaciona o usuário ativo com todos os outros usuários, logo se utiliza o coeficiente como um delimitador para identificar todos os usuários que estão acima desse limiar. Esse algoritmo não realiza correlações negativas. O uso do coeficiente de restrição serve para garantir que somente quando ambos votarem um item positiva ou negativamente será incrementado o valor da similaridade.

3.7.3 Coeficiente de correlação de postos de *Spearman*

A correlação de postos de *Spearman* (HERLOCKER *et al.*,1999) é uma função sem parâmetros e aplicada quando não há hipótese de linearidade nos dados. Necessita-se de que os dados tenham, pelo menos, duas medidas e uma variável que identifica o dado analisado. Por exemplo, se temos dois filmes e queremos analisar a relação entre eles, tem-se como medida a avaliação feita por diversos usuário nos dois itens e a identificação é por cada usuário. Essa correlação converte duas variáveis em uma posição.

A posição é o relacionamento entre um conjunto de itens. Um item pode estar em uma posição acima, abaixo ou até na mesma posição que outro item. Reduzindo os detalhes das medidas por intermédio de uma sequência de itens, torna-se possível avaliar a complexidade da informação conforme algum critério. Por exemplo, um filme poderá ser organizado de acordo com as notas dadas pelos usuários, facilitando um usuário rapidamente a escolher o item. A correlação é calculada para as duas colunas de posicionamento.

$$w_{a,u} = \frac{\sum_{i=1}^{m}(rank_{a,i} - \overline{rank_a})(rank_{u,i} - \overline{rank_u})}{\sigma_a * \sigma_b}$$

Porém essa correlação não demonstrou ser expressivamente mais eficaz que a correlação de Peason (HERLOCKER *et al.*,1999).

3.7.4 Cosseno

Frequentemente, documentos são representados por meio de documentos. A similaridade entre dois itens pode ser medida tratando cada item como um vetor de notas e calculando o cosseno do ângulo entre os dois vetores (SALTON *et al.*,1987). Considerando a matriz usuário x item (u x i), a similaridade entre os itens i e j é definida como o cosseno de vetores de dimensão n correspondente a coluna i e j.

$$\cos(i,j) = \frac{\vec{i}.\vec{j}}{|\vec{i}| * |\vec{j}|}$$

O Denominador serve para normalizar as avaliações dos usuários para que os usuários que avaliam mais itens não tenham prioridade sobre os outros usuários. O "." significa o produto entre dois vetores. Logo, a medida do cosseno não leva em consideração a magnitude dos dois objetos.

A fundamental diferença do cálculo entre similaridade baseada em usuário e em item é que, no primeiro caso a similaridade é calculada nas linhas da matriz enquanto a segunda é calculada na coluna.

No entanto, há uma desvantagem na similaridade baseada em itens. A diferença das notas dadas por usuários diferentes não são consideradas. Para isso, SARWAR *et al.* (2001) reformulou a equação de correlação de Pearson, para calcular a similaridade entre dois itens.

$$\cos(i,j) = \frac{\sum_{u \in U}(R_{u,i} - \overline{R_i})(R_{u,j} - \overline{R_j})}{\sqrt{\sum_{u \in U}(R_{u,i} - \overline{R_i})^2} * \sqrt{\sum_{u \in U}(R_{u,j} - \overline{R_j})^2}}$$

Sendo $\overline{R_i}$ a média das avaliações feitas pelo usuário i

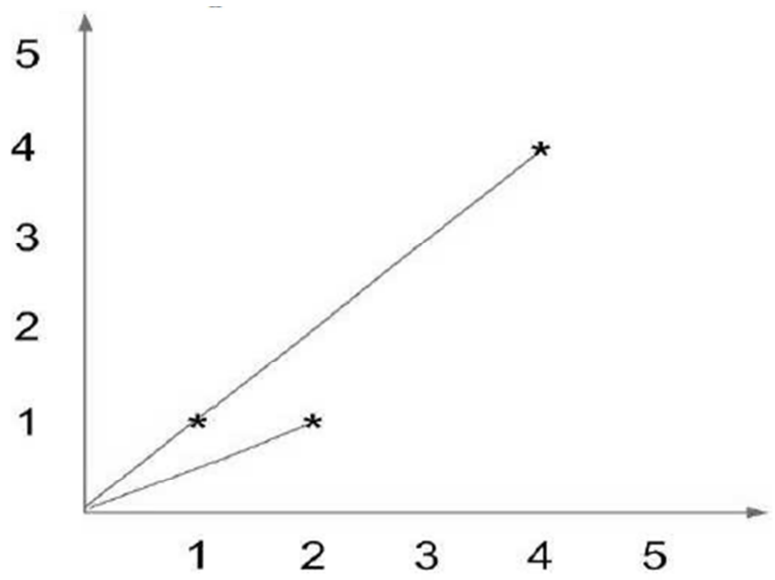

Figura 11 – Gráfico do cosseno retirado de (DING *et al.*, 2006)

Conforme mostrado na figura acima x1(1,1), x2(4,4), x3(2,1). A similaridade entre x1 e x3 é a mesma que x2 e x4, embora os vetores tenham tamanhos diferentes. Essa propriedade não é adequada para os algoritmos e filtragem colaborativa, pois o tamanho do vetor diz o quanto o usuário gostou de um item.

Sendo a correlação de Pearson é uma translação do cosseno, podendo, então, matematicamente ser exposta da seguinte forma:

$$sim^{(p)}(X,Y) = sim^{(c)}\left(\alpha(X - \overline{Z}, Y - \overline{Z})\right)$$

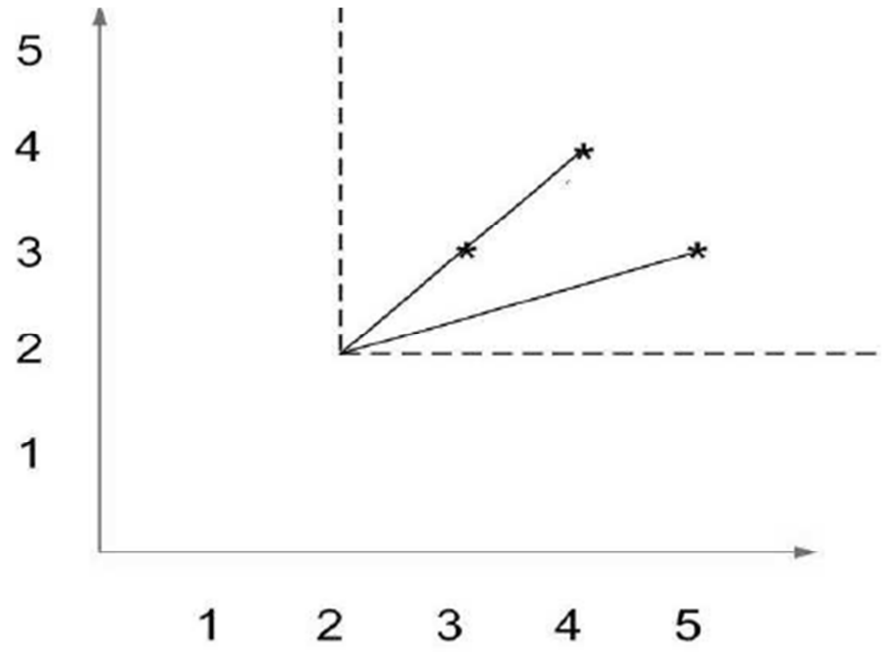

Figura 12 – Gráfico do Pearson retirada de (DING *et al.* , 2006)

Considerando os pontos x1(3;3), x2(4;4) e x3(5; 3), a similaridade de Pearson é igual entre x1 e x3 e x2 e x3.

3.7.5 Defaut Voting

Default voting (BREESE *et al.*,1998) é uma extensão do algoritmo de correlação descrita acima. A correlação entre dois usuários é calculada comparando itens que os dois usuários avaliaram. Caso os usuários tenham avaliado poucos itens em comum, a correlação será prejudicada.

O algoritmo Default Voting usa uma abordagem diferente. Esse problema pode agravar caso a matriz usuário x item seja muito esparsa. Para solucionar o problema, é adicionado um valor automaticamente para avaliações explícitas que ainda não ocorreram. Esses valores adicionados serão usados para comparar a similaridade entre os usuários.

$$\boldsymbol{w(a,i)}$$

$$= \frac{(\boldsymbol{n}+\boldsymbol{k})\sum_j(\boldsymbol{v_{a,j}v_{i,j}}+\boldsymbol{kd^2})-(\sum_j \boldsymbol{f_j}\,\boldsymbol{v_{a,j}}+\boldsymbol{kd})(\sum_j \boldsymbol{v_{i,j}}+\boldsymbol{kd})}{\sqrt{((\boldsymbol{n}+\boldsymbol{k})\sum_j \boldsymbol{v_{a,j}^2}+\boldsymbol{kd})^2)((\boldsymbol{n}+\boldsymbol{k})(\sum_j \boldsymbol{v_{i,j}^2}+\boldsymbol{kd})^2-)\,(\sum_j \boldsymbol{v_{a,j}v_{i,j}}+\boldsymbol{kd})^2\,)}}$$

Sendo n a quantidade de itens avaliados pelos usuários a e j.

Poderá ser adicionado um valor automático para avaliação de alguns itens que não foram avaliados por nenhum dos usuários. O efeito desta adição, que apesar dos itens não serem avaliados, é identificar se os usuários concordam sobre o filme.

3.7.6 Inverse User Frequency

Os itens que possuem avaliações positivas universalmente aceitas não são uteis para diferenciar itens em relação aos outros itens**.** Logo, para distinguir esses itens, a medida *Inverse User Frequency* (BREESE *et al.*,1998) usa a frequência que o item i é avaliado f'_i para esse propósito. Menor será a frequência para os itens que são bastante avaliados. A equação dessa medida é dada abaixo.

$$\boldsymbol{w(a,i)} = \frac{\sum_j \boldsymbol{f_j}\,\boldsymbol{v_{a,j}v_{i,j}} - (\sum_j \boldsymbol{f_j}\,\boldsymbol{v_{a,j}})(\sum_j \boldsymbol{f_j}\,\boldsymbol{v_{i,j}})}{\sqrt{(\sum_j \boldsymbol{f_j}\,(\sum_j \boldsymbol{f_j}\,\boldsymbol{v_{a,j}^2} - (\sum_j \boldsymbol{f_j}\,\boldsymbol{v_{a,j}})^2\,))} * \sqrt{(\sum_j \boldsymbol{f_j}\,(\sum_j \boldsymbol{f_i}\,\boldsymbol{v_{i,j}^2} - (\sum_j \boldsymbol{f_j}\,\boldsymbol{v_{i,j}})^2\,))}}$$

3.7.7 Case Amplification

A intenção desse algoritmo é amplificar o valor dos pesos. O peso é a similaridade entre dois objetos no sistema de recomendação.

$$\boldsymbol{w'_{a,i}} = \begin{cases} \boldsymbol{w_{a,i}^{\rho}}\ \boldsymbol{caso}\ \boldsymbol{w_{a,i}} \geq \boldsymbol{0} \\ -(-\boldsymbol{w_{a,i}^{\rho}}\ \boldsymbol{caso}\ \boldsymbol{w_{a,i}} < 0 \end{cases}$$

A mudança do valor do peso amplifica usuários que tenham um alto grau de similaridade e reduz os que têm um baixo grau.

3.7.8 Similaridade *LikeMinds*

A similaridade *LikeMinds* (GREENING, 1997) entre dois usuários é realizada por meio de uma função que retorna a soma da diferença entre as avaliações feitas nos itens em comum. Em outras palavras, a similaridade do usuário i com o usuário k é feita por intermédio das avaliações de todos os itens que esses usuários avaliaram em comum. Para essas avaliações, é calculada a diferença das notas.

$$CVT_{i,k} = \sum_{l \in R_i \cap R_k} C\ (|r_{i,l} - r_{k,l}|)$$

Após o cálculo da equação acima, utiliza-se o resultado para o cálculo da similaridade entre os dois usuários através da equação.

$$AST_{i,k} = \frac{CVT_{i,k*} * \log_2 card(R_i \cap R_{k)}}{card(R_i \cap R_{k)}}$$

Sendo R_i os itens avaliados pelo usuário i. Essa equação relaciona a quantidade de itens que ambos avaliaram em comum para ajustar a similaridade dos usuários.

3.8 Métricas

Várias métricas foram propostas para avaliar os sistemas de recomendação. Para avaliar a qualidade dos sistemas de recomendação utilizam-se métricas para quantificar a acurácia desses sistemas. As duas principais métricas utilizadas para avaliar a acurácia dos sistemas de recomendação são *Mean Absolute Error* e *Root Mean Square Error*. (HERLOCKER *et al.*,1999).

3.8.1 MAE (Erro Médio Absoluto)

A métrica MAE (*Mean Absolute Error*) calcula a diferença entra a média dos valores estimados e os valores reais. Portanto, o MAE quantifica o quão perto o valor

estimado está do valor real especificado pelo usuário. De acordo com a métrica, quanto menor o valor do MAE, melhor será o algoritmo.

MAE: $\sum_{i=1}^{n} \frac{|pi-qi|}{N}$

Primeiramente, calcula-se a soma do erro absoluto de N pares de avaliação e previsão, e, então, calcula-se a média.

3.8.2 NMAE

Caso as escalas numéricas usadas pelos sistemas de recomendação sejam diferentes, por exemplo, um sistema usa uma escala de nota de [0,+5], já outro utiliza uma escala de [-10,+10], a métrica MAE não será uma boa métrica de comparação destes dois sistemas, pois o intervalo de notas é diferente. Para tratar essa questão, o NMAE (GOLDBERG *et al.*,2001)normaliza essa métrica para auxiliar comparações de sistemas.

$$NMAE = \frac{MAE}{\overline{r_{max}} - \overline{r_{min}}}$$

Sendo que $\overline{r_{max}}$ e $\overline{r_{min}}$ correspondem ao maior e ao menor valor que um item poderá ser avaliado respectivamente.

3.8.3 RMSE

Uma métrica também utilizada é o RMSE (*Root Mean Square Error*) que calcula a raiz quadrada do quadrado da média da diferença entre o valor estimado e o valor real. Em outras palavras, o RMSE é a raiz quadrada da variância. Essa métrica enfatiza o erro absoluto.

$$RMSE = \sqrt{\frac{\sum_{\{u,i\}}(p_{u,i} - q_{u,i})^2}{N}}$$

Sendo $p_{u,i}$ a previsão da avaliação feita do usuário u ao item i e $q_{u,i}$ representa a real avaliação feita pelo usuário. A variável N diz a quantidade de previsões feitas pelos sistemas de recomendação.

3.8.4 Cobertura

Cobertura é uma medida de porcentagem de itens para os quais os sistemas de recomendação podem fornecer predições. São características comuns de sistemas que podem reduzir a cobertura a pequena quantidade de vizinhos e poucas avaliações feitas pelo usuário.

Considerando I*(u) o conjunto de itens I que o sistema seja capaz de prever para o usuário u e |I| o conjunto de todos os itens que estão na base de dados.

$$cobertura(u) = \frac{\mathrm{I}^*(\mathrm{u})}{|\mathrm{I}|}$$

Para calcular a cobertura de entre todos usuários.

$$cobertura = \frac{\sum_{u \in U} \mathrm{cobertura(u)}}{|\mathrm{U}|}$$

Sendo |U| o conjunto de todos os usuários que estão na base de dados.

3.8.5 *Relative Operating Characteristic* (ROC)

ROC (HERLOCKER *et al.*, 1999) é uma métrica de diagnóstico de sistemas de recomendação. A curva ROC permite medir o quanto um valor produzido por um sistema é capaz de distinguir os elementos relevantes dos não relevantes. A área abaixo de uma função plotada leva em consideração a sensibilidade e a especificação dos testes. A sensibilidade refere-se à probabilidade de selecionar um item aleatoriamente ruim que será rejeitado pelo filtro. A sensibilidade e a especificidade entre 0 e 1, obtendo, então, um conjunto de pontos que retornam um limiar para aceitação do item. A saída é uma variável de relevância associada a cada elemento permitindo construir

duas curvas de distribuição: uma para os valores obtidos para os elementos relevantes e outra para os elementos não relevantes.

Os elementos que estão acima do limiar especificado serão considerados relevantes, porém, se estiver abaixo desse limiar, serão rejeitados. A vantagem dessa métrica é que o limar não necessita ser escolhido. É possível calcular a cobertura e o ruído (que são os elementos que foram escolhidos e deveriam ter sido rejeitados).

A curva ROC está no plano cartesiano com os valores da cobertura (ordenada) versus o ruído (abscissa).

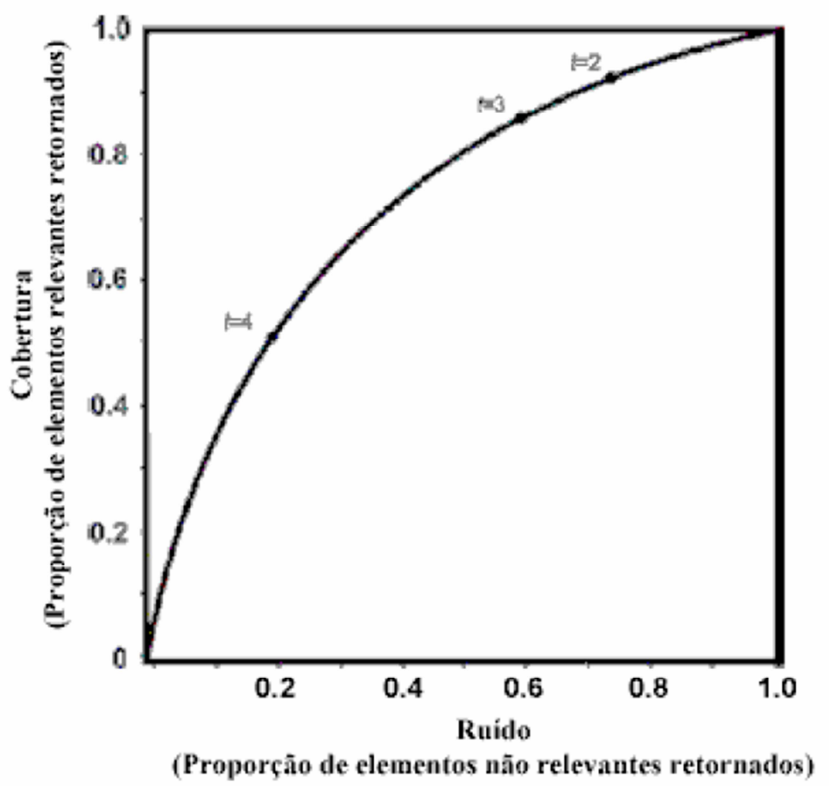

Figura 13 – Gráfico de ROC

3.8.6 Loss Function

A função perdida (*Loss function*) (HOFMANN, 2004) é uma função que quantifica o quão bom ou ruim um modelo de previsão é comparado com o valor real.

A função estipula um parâmetro no espaço do modelo H e usa um parâmetro genérico θ, no qual v é tratada no caso de avaliações implícitas. Dada uma observação (u,v,y) sendo u o usuário, y o item e v a avaliação feita do usuário u do item y, a função L será atribuída um valor para cada hipótese θ . Quanto menor a função *L(u, v, y), θ)* mais compatível será a hipótese com a observação.

$$Llg1((u, v, y), \theta) = -\log P(v|u, y; \theta), \text{ ou } Llg2((u, v, y), \theta) = -\log P(v, y|u; \theta). \quad (1)$$

3.9 Problemas

Apesar de RESNICK *et al.* (1994) relatarem que as recomendações personalizadas são mais acuradas que médias gerais, há inúmeros problemas relacionadas as recomendações personalizadas.

3.9.1 Esparsidade

A quantidade de avaliações feitas por indivíduos aos itens é pequena comparada com o máximo possível de avaliações que podem ser realizadas. Isso ocorre devido à grande quantidade de itens que podem ser avaliados por um usuário. Em alguns sistemas de recomendação, mesmo os usuários ativos avaliam menos que 1% da quantidade de itens que poderiam ser avaliados (SARWAR *et al.*,2001).

Além disso, a probabilidade de encontrar um conjunto de usuários com notas similares é geralmente baixa (MELVILLE *et al.*,2002), logo, há uma grande proporção de células da matriz usuário x item nula. Esse fenômeno dificulta a identificação de usuários que possam ser relacionados através de suas avaliações ao mesmo item. Quando há poucos itens a serem avaliados, esse fenômeno ocorre no início do uso do sistema. Isso dificulta a implementação de recomendações personalizadas.

No caso da base de dados ser consideravelmente grande, a menos que utilize uma técnica de redução da dimensionalidade, amostragem ou particionamento, a técnica de

filtragem colaborativa não é recomendada para solucionar a questão da espasidade. O modelo de agrupamento reduz essa questão, porém a qualidade da recomendação é reduzida em comparação com a filtragem colaborativa (LINDEN, 2003).

Alguns sistemas tentam aumentar o número de avaliações do usuário, analisando o comportamento dele e, então, preenchendo as supostas avaliações na matriz usuário-item. A redução de dimensionalidade, através de técnicas estatísticas, resume a dimensionalidade da matriz. O particionamento da matriz aumenta a densidade das células, já que nem todos os usuários avaliam certos tipos de itens (SARWAR, 1998).

3.9.2 Problema de Partida Fria (*Cold-Start*)

Vários sistemas de recomendação ignoram itens, que não são populares ou foram introduzidos recentemente nos sistemas de recomendação (PARK *et al*., 2008). Há uma dificuldade em encontrar itens que possam ser recomendados para novos usuários do sistema de recomendação, devido ao fato de que eles não avaliaram ainda nenhum filme.

Os sistemas de recomendação baseados em conteúdo têm dificuldades de relacionar usuários que não avaliaram, pelo menos, um mesmo item que outro usuário. Para resolver esse problema, pesquisas evidenciam que a utilização de técnicas *active-learning* reduz os problemas de partida fria. Os filmes novos e os que não tiveram nenhuma avaliação recebida também são difíceis de serem recomendados pelo mesmo problema da adição de um novo usuário.

Um classificado bayesiano com objetivo de recomendar filmes que ainda não foram avaliados por nenhuma pessoa dentro da comunidade foi proposto por SCHEIN et al. (2002). Os filmes são recomendados baseados na similaridade de seus elencos com os filmes avaliados pelo usuário.

3.9.3 Escalabilidade

Os algoritmos de filtragem colaborativa são capazes de realizar buscas de vizinhos na ordem dos milhares, mas a demanda atual está na ordem dos milhões (SARWAR *et al.*,2001). Alguns algoritmos têm dificuldades em sugerir itens quando há uma grande quantidade de usuários e itens. Caso haja uma grande quantidade de itens por usuário, isso reduz o número de vizinhos que poderão ser procurados reduzindo então a escalabilidade.

SARWAR *et al.* (2001) propôs analisar a matriz usuário-item identificando relações entre itens diferentes e usando essas relações para calcular as recomendações. Essa técnica não necessita da identificação de vizinhos similares ao usuário quando uma recomendação é requisitada, reduzindo o impacto da escalabilidade.

O aumento do número de clientes e produtos trazem desafios para os sistemas de recomendação: as recomendações produzidas necessitam de qualidade nas suas previsões de realizar milhões de recomendações por segundo e de alcançar uma alta cobertura advinda do problema da esparsidade.

O algoritmo apresentado por RASHID *et al.* (2006) apoia-se na ideia de que métodos clássicos de agrupamento são capazes de solucionar o problema da escalabilidade. A ideia é criar K grupos, e agrupar os usuários nestes grupos. Para cada centroide do grupo, calcula-se a média dada pelos usuários no grupo para cada item da base de dados. Então, para realizar uma predição, o algoritmo calcula a similaridade do usuário ativo com cada um dos grupos e usa esta similaridade e a avaliação média do item para cada grupo com o intuito de retornar o valor.

3.9.3.1 Aplicando a Decomposição de Valores Singulares (SVD) em filtros colaborativos

A redução de dimensionalidade pode eliminar características irrelevantes e reduzir o ruído. O ruído é o componente aleatório de um erro de medição. Uma das técnicas comumente usada na redução da dimensionalidade de uma matriz é a decomposição de valores singulares. O SVD é uma técnica de fatoração da matriz que faz com que a matriz seja fatorada em 3 matrizes.

$$R = U \,.S\,.\, V'$$

U e V são matrizes de tamanho m x r e n x r respectivamente; r é o grau da matriz R. S é a matriz diagonal de tamanho r x r contendo todos os valores da matriz singular na diagonal da matriz. Todas as entradas da matriz S são positivas e ordenadas pelos valores de forma decrescente. É possível reduzir a matriz S para ter somente K valores, obtendo uma matriz *Sk*, sendo k < r. Se a matriz U e V são reduzidas, a matriz reconstruída será Rk = Uk.Sk.V´k será uma matriz de grau k de R. O "´" indica que a matriz é transposta. Os novos atributos que são produzidos são combinações lineares das variáveis originais.

A implementação da matriz poderá ser realizada desta forma:

A[i][j]

$= \Sigma_{i<k}$ U[i][k] * S[k][k] * V[j][k]

$= \Sigma_{i<k}$ U[i][k] * sqrt(S[k][k]) * sqrt(S[k][k]) * V[j][k]

$= \Sigma_{i<k}$ (U[i][k] * sqrt(S[k][k])) * (sqrt(S[k][k]) * V[j][k])

$= (U[i] * sqrt(S_{diag})^T) * (V[j] * sqrt(S_{diag})^T)^T$

Pode-se capturar potenciais relacionamentos entre cliente e produto ao reduzir a dimensionalidade da matriz usuário x item para realizar as recomendações por meio do SVD.

A escolha do valor de dimensionalidade k é crítica para a qualidade das previsões. O valor de k tem que ter um tamanho que seja capaz de capturar todas as estruturas importantes e não seja pequena o bastante para evitar o *overfitting*.

3.9.4 Sinônimo

Diferentes itens podem referir-se a objetos similares. Os sistemas de recomendação baseados em correlação não conseguem encontrar a associação entre esses itens e tratá-los de forma diferente. Por exemplo, bergamota, tangerina ou mexerica referem-se à mesma fruta. Dependendo da região, o produto é chamado de outra maneira.

Capítulo 4 – Tempo e Nichos de Mercado

No cenário atual, as preferências dos usuários podem mudar com o decorrer do tempo, afetando seu humor, contexto e até tendências culturais (LI *et al.*, 2011, KOREN, 2009a). Portanto, em um determinado período de tempo, o usuário poderá ter uma ou mais preferências levando em consideração os aspectos acima.

Os sistemas de recomendação baseados em filtragem colaborativo, que adotam todo o histórico do usuário, podem ser inapropriados para recomendar itens aos seus clientes (LI *et al.*, 2011). As avaliações realizadas pelos usuários em tempos diferentes possuem o mesmo peso para esses sistemas de recomendação. Os modelos temporais podem ser a chave no projeto de sistemas de recomendação, todavia mudanças nos dados ocorrem ao longo do tempo, dificultando a atualização desses modelos para refletir o cenário atual.

O conceito de tendência das preferências (DING *et al.*, 2006) demonstra, por meio de análise de dados, que há padrões nas preferências dos usuários que alteram ao longo do tempo. Por exemplo, um indivíduo que gostava de filmes de ação, com o decorrer do tempo, por ter começado a namorar começou a gostar de comédia romântica. Logo, para melhorar a qualidade das suas recomendações, os sistemas de recomendação podem considerar esse conceito.

Por intermédio da interseção de vários itens e pessoas, as características de ambos sofrem alterações ao longo do tempo (KOREN, 2009a). Por exemplo, as televisões ao longo do tempo evoluíram sua tecnologia, antes por meio de tubo de raio catódico até as atuais LEDS. As pessoas que se interessavam por televisão de tubo de

raio catódico mudaram suas preferências por televisões mais modernas. Consequentemente, ao longo do tempo, tanto as características das preferências dos consumidores quanto das características das televisões mudaram.

4.1 Winner-take-all x Cauda Longa

Algumas pessoas acreditam que as recomendações ajudam consumidores a descobrir novos produtos e a aumentar as vendas, enquanto que outros creem que ajudam a vender produtos populares (FLEDER *et al.* 2008).

Há duas teorias opostas na literatura. A teoria *winner-take-al*l (ROSENM, 1981) prevê que o mercado será homogêneo devido à redução do custo da localização dos melhores itens, que depois estes se tornarão "superestrelas", isto é, muito bem avaliados e com uma grande quantidade de avaliações. Essa teoria assume que as pessoas terão gostos similares e considera que os ruídos na comunicação entre o consumidor e vendedor irão reduzir, e que consequentemente os produtos mais bem avaliados serão mais vendidos. Haverá, portanto, uma convergência nos interesses dos consumidores.

A teoria da cauda longa foi descrito primeiramente por Anderson (2004). Cauda longa é um fenômeno relacionado principalmente ao mercado de vendas, no qual a venda de nicho de produtos possui uma margem de crescimento, tornando-se uma considerável quantidade de vendas em relação ao total. A teoria da cauda longa (ANDERSON, 2004) argumenta que a redução do custo da localização de itens levará a fragmentação do mercado, pois os sistemas de recomendação tornam mais fáceis para os consumidores o encontro de nichos de produtos que satisfaçam suas necessidades. De acordo com a teoria da cauda longa, o mercado demandará mais produtos de nichos que produtos muito avaliados assim que o custo de localização dos produtos reduz (ANDERSON, 2006).

Embora essas duas teorias enfatizem a função dos sistemas de recomendação, ainda não se confirmou o que acontecerá com o mercado quando diferentes tipos de sistemas de recomendação forem aplicados para vários tipos de consumidores. As duas teorias não consideram a preferência individual de cada pessoa antes e depois de se tornarem clientes dos sistemas de recomendação. Consumidores com preferências únicas talvez considerem mais os nichos de produtos, enquanto que a maioria dos consumidores estarão interessados nos produtos mais populares antes da implementação desses sistemas (WU, 2011).

ELBERSE, OBERHOLZER-GEE (2009) relataram uma pequena mudança nas vendas de produtos dos mais populares para nichos, entretanto, há uma quantidade considerável de itens que não são vendidos. Conclui-se também que grande parte das vendas é de itens que possuem poucas avaliações.

4.2 Padrões comportamentais dos usuários e itens baseado no tempo

KOREN (2009a) utilizou a mudança do tempo como um fator para recomendar um filme. Na sua primeira análise, ele relatou que os filmes podem, em um determinado tempo, tornarem-se mais ou menos populares devido a algum evento externo. Por exemplo, um filme que teve um grande pico de avaliações durante o seu lançamento e, após dois anos do seu lançamento, teve outro aumento significativo de avaliações após a continuação do filme.

Na segunda análise, KOREN (2009a) avaliou a mudança de base utilizada pelo usuário para avaliação dos filmes. Um usuário que em média costumava dar nota dois, por algum motivo ele começa a avaliar, na maioria das vezes, os filmes com nota quatro. Além dos exemplos citados acima, um padrão interessante relatado por KOREN (2009a) indicou que os filmes avaliados por um mesmo usuário no mesmo dia tendem a ter notas semelhantes. Alguns usuários são mais conservadores e tendem a atribuir

avaliações aos itens que estão dentro de um intervalo de categorias enquanto que outros atribuem a um grande conjunto de itens que estão em variadas categorias (DING *et al.*, 2006).

Mudanças nas características de compra dos consumidores alteram ao longo do tempo através de mudanças sazonais, feriados específicos, dias festivos (ex.: dia dos namorados). Ademais, filmes mais antigos tendem a receber avaliações melhores que os atuais (KOREN, 2009b).

LEE *et al.* (2009) relatam que os clientes de uma empresa de música virtual, que têm uma quantidade significativa de músicas nas suas listas de reprodução, tendem a escutar expressivamente um subconjunto pequeno da sua lista e, raramente, o resto da lista.

4.3 Técnicas clássicas empregando o fator tempo

Nos sistemas de recomendações atuais, há uma necessidade de distinguir melhores os efeitos transientes de padrões que perduram por um determinado tempo. Considerando o tempo um fator relevante para melhorar a qualidade dos sistemas de recomendação, um modo para empregar esse fator é por meio das avaliações mais recentes que refletem melhor as preferências atuais das pessoas.

A maneira mais comum de modelar essa hipótese é penalizar as avaliações mais antigas (DING *et al.*, 2006). Tang *et al.* (2003) consideraram que os dados das avaliações aos itens mais antigos poderiam ser truncados, deixando somente os itens mais recentes como entrada para os sistemas de recomendação. Já Ding *et al.* (2006) propuseram um algoritmo que usa pesos nos itens baseado na expectativa da qualidade das suas futuras preferências. Em vez de penalizar os dados passados, o algoritmo analisa a distribuição dos dados ao longo do tempo.

Ainda que uma pessoa possa ter suas preferências alteradas ao longo do tempo, a distribuição dos itens deveria permanecer a mesma, pois os interesses das pessoas seriam contrabalanceados. O interesse do indivíduo poderá alterar de A para B, enquanto de outro poderá mudar de B para A. Por meio dessa hipótese, LI *et al.* (2011) definiram uma janela de tempo com uma escala de 1 até T, tendo T o mesmo tamanho da escala adotada para avaliação dos itens feitas pelos usuários. Por meio disso, relaciona-se essa janela a cada usuário e item da base de dados. Cada janela do usuário é uma combinação de K protótipos do usuário e uma distribuição dele dentro de K grupos de usuários $p_i(t)$. A janela do item será similar ao usuário, contento L grupos $q_j(t)$. Logo a nota dada em um intervalo de tempo t realizada pelo usuário i ao item j $X_{i,j}(t)$ pode ser prevista por:

$$X_{i,j}(t) = [p_i(t)]^T \, B q_j(t)$$

A restrição do tamanho da janela limita a máxima diferença de tempo permitida entre a primeira e a última ocorrência de eventos em qualquer elemento de um padrão sequencial.

Por intermédio de pesos que auxiliam a relevar as avaliações considerando o tempo como o fator importante, KOREN (2009b) analisou diferentes taxas de decaimento e concluiu que a qualidade das previsões das avaliações alcança melhores resultados quando reduz o decaimento do tempo, atingindo o melhor resultado quando não há nenhum decaimento. Apesar do fato das pessoas mudarem suas preferências e a escala de avaliação ao longo dos anos, muitas das preferências das pessoas permanecem as mesmas, possibilitando a elaboração de um padrão tanto para as pessoas quanto aos itens.

As notas mais recentes do usuário ativo refletem mais suas preferências que as antigas (DING *et al.*,2005). Portanto, um item que foi avaliado recentemente por um usuário deveria ter um impacto maior que aquele avaliado há muito tempo atrás. Uma forma de tratar esse problema seria por meio de janelas deslizantes. Nesta abordagem, define-se um tamanho para a janela, e somente as avaliações mais recentes serão utilizadas, limitadas pelo tamanho da janela. Porém, a inutilização de dados antigos agrava o problema da esparsidade, podendo, então, tornar os resultados piores (DING *et al.*,2005). Outro problema relatado por KOREN (2009b), é que a mesma importância é dada a todas as avaliações realizadas dentro da janela. Quando a mudança no padrão das avaliações é abrupta ao longo do tempo, torna-se razoável a utilização de janela deslizante, porém é inadequada quando há uma mudança gradual nas avaliações.

Uma forma tradicional de lidar com esse tipo de problema é usar séries temporais ou modelos de regressão estatística. Modelos de séries temporais, *autoregressive moving average* (ARMA) e *exponential smoothing* (BOX *et al.*,1994) usam dados do passado para fazer predições futuras. Entretanto, com a adição de novos itens ao longo da utilização dos sistemas de recomendação, esses modelos não são eficazes. Modelos regressivos também não são apropriados, pois esses novos itens não são recomendados devido a pouca quantidade de atributos associados a esses novos itens. (XIONG *et al.*,2011)

Métodos estáticos demonstraram não serem capazes de tratar a mudança de um item ou a preferência de um usuário. A preferência do mercado pode mudar ao longo do tempo, podendo ser sazonal ou não. Caso se utilize somente dados recentes, uma grande quantidade de dados importantes será perdida, tornando a matriz usuário-item esparsa (XIONG *et al.*, 2011) .

XIONG *et al.* (2011) propuseram um método baseada em fatoração que é capaz de modelar dados relacionados envolvendo o tempo. A inclusão de fatores adicionais representa características potenciais do nível de preferência da população em um tempo particular. Um tratamento especial é feito no fator tempo para garantir que a evolução dos fatores é suave. Esse modelo usa todos os dados disponíveis e adapta esses fatores a período de tempo diferente.

PARK *et al.* (2008) propuseram um algoritmo que divide os itens em cauda e cabeça os quais melhoram o desempenho comparado a alguns algoritmos alternativos. Agrupam-se os itens por meio de um algoritmo de agrupamento. Para cada grupo de itens, um modelo de previsão de nota é elaborado baseado nas avaliações feitas nos itens daquele grupo.

O agrupamento foi idealizado para aumentar a quantidade de dados utilizados na construção do modelo de previsão, pois o erro na avaliação da acurácia de uma previsão tende a aumentar quando há poucos itens. O agrupamento do item é feito somente na cauda e modelos individuais de previsão são realizados na cabeça. Para particionar os itens em cabeça e cauda, define-se um ponto de corte baseado na frequência que o item foi avaliado. Considerando um item que seja avaliado n vezes e o ponto de corte x, caso n > x o item é considerado na categoria da cabeça, caso o contrário será considerado cauda.

A escolha da quantidade de grupos e limiar para definição da categoria do item é sensível para o desempenho desse algoritmo. Essa escolha não é trivial, pois depende da qualidade do dado que está sendo tratado.

4.4 Cauda Longa e os Sistemas de Recomendação

Os consumidores procuram e seguem as recomendações por causa do aumento na probabilidade de encontrar os itens de que eles necessitam. Os consumidores dos nichos

de mercado atentam-se mais às recomendações com qualidade do que o preço dos itens. Portanto, a preferência por itens tem um impacto maior que seu preço em relação à concentração de vendas (HERVAS-DRANE, 2007).

Anderson (2004) menciona que os sistemas de recomendação podem ser fundamentados em nichos de mercado em vez de apoiados nas avaliações dos usuários. Isso ocorre devido aos sistemas de recomendação baseados em filtragem colaborativa necessitar das avaliações dos usuários para recomendar itens para os usuários.

Logo, explorar informações dos consumidores pode produzir recomendações dos produtos com maior qualidade, possibilitando o mercado a explorar novos meios de atrair consumidores (HERVAS-DRANE, 2007). Analistas e observadores propuseram que as vendas online aumentarão a oferta de produtos, possibilitando considerar os nichos de mercado (FLEDER *et al.*, 2008).

Determinar quais são os atributos adequados para discriminar a preferência do usuário na cauda-longa e explorar os especialistas que estão nela para melhorar as recomendações para os demais é um dilema (JUNG, PHAM, 2001).

A modelagem de usuários baseada em atributos pode representar as preferências do usuário por meio das atividades e *feedbacks* e é relativamente simples desde que as preferências dos usuários sejam limitadas. No entanto, caso os atributos pré-definidos não cubram todas as preferências do usuário, não é possível comparar dois usuários. Para solucionar a questão, informações adicionais são necessárias para explorar a similaridade entre eles. JUNG, PHAM (2011) focaram na exploração de informações externas para identificar as preferências do grupo na cauda-longa.

O custo de participação de um consumidor no mercado pode ser dividido em dois componentes: o custo de localizar o item desejável e o preço a ser pago pelo item (HERVAS-DRANE, 2007).

A acurácia dos sistemas de recomendação tende a declinar quando se recomendam itens que estão na cauda-longa. Uma forma de reduzir a concentração de vendas é por meio de sistemas de recomendação.

Os sistemas de recomendações baseados em conteúdo comparados aos baseados em filtragem colaborativa tendem a ter uma menor concentração de vendas (WU, 2011).

4.4.1 Problemas

O aumento da disponibilidade de produtos para os clientes é um aspecto relevante no mercado online, dado que mais nichos de consumidores poderão acessar seus itens preferidos por meio de lojas virtuais. Alguns itens são excluídos nos mercados tradicionais, pois há problemas de logística dos produtos, bem como dificuldades de armazenamento e exibição dos produtos na prateleira, o que reduz a disponibilidade dos produtos para nichos de mercado (HERVAS-DRANE, 2007).

Sistemas de Recomendação que empregam filtro colaborativo, análise baseada em conteúdo ou métodos híbridos, têm como objetivo geral recomendar tanto itens relevantes quanto novos. No entanto, devido à natureza dos algoritmos usados para implementação desses sistemas, eles influenciam a recomendação dada, tornando difícil a produção de novas recomendações (LEE *et al.*,2009).

No âmbito de recomendações de produtos que o cliente não espera obter, os sistemas de recomendação baseados em conteúdo fornecem resultados mais relevantes comparados aos sistemas citado acima porque o conteúdo que está sendo analisado não sofre influências da popularidade do item ou qualquer influência externa (LEE *et al.*,2009)

Selecionar itens que estão na cauda-longa possui uma grande chance de serem novos para quase todos os usuários, contudo não é cabível nas técnicas de recomendação eliminar os itens com maior popularidade, já que o que é novo para um

usuário pode não ser para outro (LEE *et al.*,2009). Além disso, o erro na avaliação da acurácia de uma previsão tende a aumentar quando há poucos itens.

Para obter itens que os clientes não esperam de um sistema de recomendação de músicas, LEE *et al.* (2009) dividiram os usuários em dois grupos distintos. O cliente será categorizado como "especialista" caso o item que está na sua lista de reprodução estiver na cabeça da lista. Ao mesmo tempo, o mesmo usuário será um "novato" caso a música pertencer à cauda da lista de reprodução. As recomendações são feitas somente nos itens pertencentes à cauda. Após encontrar o item na cauda, serão encontrados os usuários especialistas que avaliaram o item alvo para, então, na lista dos usuários especialistas, encontrar similaridades com o usuário alvo analisando somente os itens que estão na cabeça de sua lista de reprodução.

No entanto, há uma dificuldade em analisar a eficácia dessa abordagem. A subjetividade do que um item é novo para um usuário reduz significativamente métodos que possam avaliar o algoritmo. Outra dificuldade encontrada é separar os itens que estão na cabeça (populares) dos que não são (cauda).

4.4.2 Dificuldades

Há uma dificuldade de encontrar o ponto de corte para separar os itens que estarão na cauda e na cabeça. Um modo que não é interessante para distingui-los é utilização de porcentagem.

Quadro 2 – Tabela de DVD´s

Top títulos	Percentual do total de aluguéis	
DVDs	Blockbuster	Netflix
1-10	31.2%	13.1%

11-100	36.4%	24.8%
100-1000	19.8%	25.9%
1000 ou mais	12.6%	36.2 %

A primeira coluna desta tabela refere-se aos DVDs mais alugados. Logo, 1-10, refere-se ao top-10 títulos de DVDs mais alugados. Analisando essa linha, obtemos que os dez títulos mais alugados na Blockbuster correspondem a 31,2% do total de aluguéis e 13,1% do total de aluguéis da Netflix, Na linha abaixo, 11-100, refere-se ao décimo primeiro ao centésimo filmes mais alugados. O mesmo entendimento pode ser feito nas outras linhas e colunas.

Caso a análise seja feita em cima da porcentagem obtém-se que:

- Blockbuster (total de filmes no estoque: 3,000 DVDs): aproximadamente 80% dos aluguéis dos filmes vêm de 10% dos filmes.
- Netflix: (total de filmes no estoque: 60,000 DVDs): aproximadamente 80% dos aluguéis vêm de 5% dos filmes.

O efeito da porcentagem induz a concentração de vendas. Analisando os resultados, parece que o negócio do NetFlix é duas vezes mais concentrado que o Blockbuster. No entanto, é necessário considerar que a Netflix possui 20 vezes mais filmes que o Blockbuster.

A análise em cima do total obteve que:

- Blockbuster: Os 100 filmes mais alugados representam 67.6% do total de aluguéis de filmes.
- Netflix: Os 100 filmes mais alugados representam 37.9% do total de aluguéis de filmes.

Por meio dessa análise, é possível verificar que os 100 filmes na Blockbuster são mais concentrados que a Netflix. Isso ocorre devido à análise por meio da porcentagem a qual pune a loja que possui uma maior quantidade de itens disponíveis para serem alugados.

4.5 Coeficiente Gini

O princípio de Pareto pode ser usado para definir o fenômeno da concentração de vendas. Algumas vezes chamado de 80/20, esse princípio sugere que uma pequena porção dos produtos no mercado pode gerar uma grande porção de vendas.

Uma forma de avaliar a concentração de vendas de uma empresa é por meio do coeficiente Gini. O Coeficiente de Gini (G) é uma medida estatística baseada na Curva de Lorenz. No eixo X, dispõem-se os percentuais acumulados dos itens, sempre em ordem crescente conforme a quantidade de avaliações, e no eixo Y, os percentuais acumulados das avaliações feitas pelas pessoas nos itens.

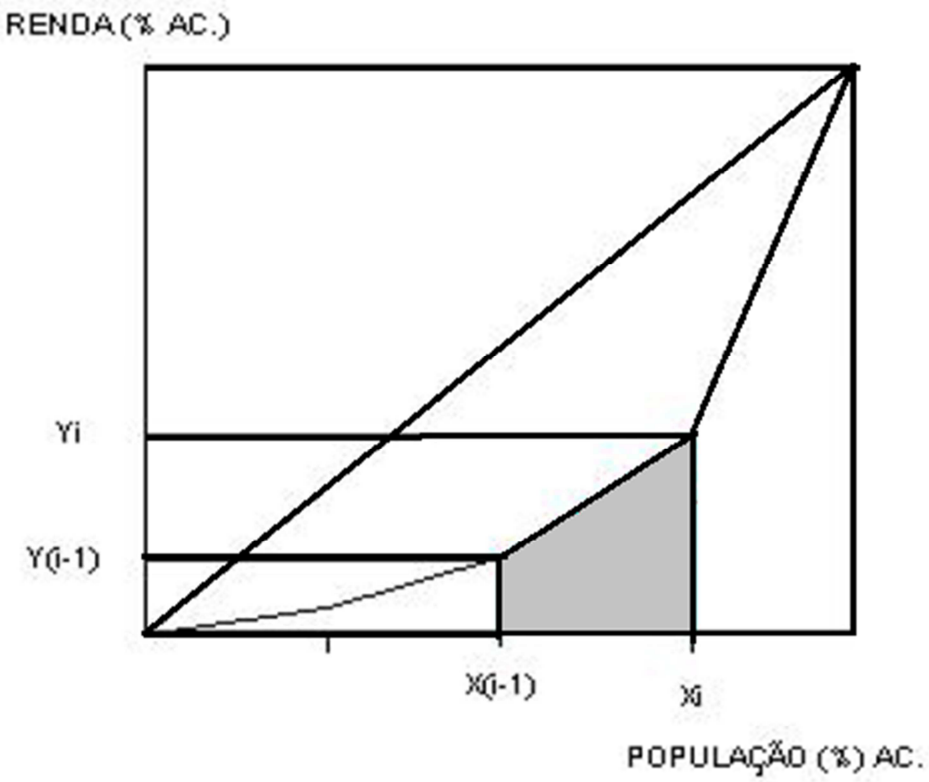

Figura 14 – Gráfico de GINI

A área entre as duas retas é chamada área de concentração. Quanto maior a concentração, maior é esta área. Dois casos extremos podem ajudar a entender a Curva de Lorenz. Em primeiro lugar, se não houvesse concentração, estaríamos sobre a reta que está a 45 graus, e a área de concentração seria zero. Por outro lado, a área de concentração seria igual ao triângulo situado abaixo da linha da reta de 45 graus caso a concentração seja máxima.

O cálculo do coeficiente de Gini é simples: divide-se a área de concentração pela área de concentração máxima, ou seja, pela área do triângulo situado abaixo da linha de perfeita igualdade:

G = Área de Concentração / Área de Concentração máxima.

Se não há concentração, o numerador é zero, e o coeficiente de Gini resulta também em zero. Se a concentração é máxima, teremos o numerador igual ao denominador, e o coeficiente assume valor um, sendo então: $0 \leq G \leq 1$.

Sabendo que a área do triangulo é ½, considerando os lados iguais a um, necessita-se definir o tamanho da área de concentração. Essa tarefa poderá ser realizada por meio de aproximação de trapézios. Calcula-se a área do trapézio conforme a equação abaixo.

$$\text{GINI: } \frac{\frac{1}{2} - \sum_{i=1}^{n}(Yi+Y_{i-1})(Xi-X_{i-1})/2}{\frac{1}{2}}$$

Logo, por meio de manipulações algébricas, o coeficiente de Gini será:

$$1 - \sum_{i=1}^{n}(Yi + Y_{i-1})(Xi - X_{i-1})/2$$

Para analisar a concentração de vendas, utiliza-se o coeficiente de Gini em uma empresa de vendas sem a utilização de um sistema de recomendação G0 (FLEDER *et al.*,2008). *Gi* será a concentração de vendas após a utilização do sistema de recomendação i.

FLEDER *et al.* (2008) definiram como tendência de concentração dependendo das condições abaixo.

Tendência	Condição
Tendência de concentração	Gi>Go
Tendência de diversidade	Gi< G0

Sem tendência	Gi=G0

Capítulo 5 - Experimento

A proposta desta dissertação apresenta novas técnicas como forma de melhorar a qualidade das previsões feitas pelo sistema de recomendação baseado em filtragem colaborativa, aplicando o tempo como fator para identificar tanto usuários que estão no mesmo nicho de mercado quanto itens de nichos de produto.

Frequentemente, encontram-se itens que inicialmente possuem uma grande quantidade de avaliações no seu lançamento, mas, conforme o decorrer do tempo, a quantidade de avaliações é reduzida comparada ao seu lançamento.

Esses filmes são itens que se tornaram destinados a pessoas que estão à procura de nicho de produtos, e que essas pessoas são nichos de mercados que deveriam ser tratadas de forma adequada, considerando a sua relevância dentro do mercado em que está inserida.

O objetivo do experimento é obter a qualidade da previsão de várias técnicas desenvolvidas para filtros colaborativos baseados em item e usuário, a fim de demonstrar qual a melhor técnica entre as listadas para filmes com esse padrão de avaliação. É avaliado também o comportamento dos usuários em determinados intervalos desses itens.

Além disso, os experimentos também servem para demonstrar o impacto do tempo na predição dos algoritmos de sistemas de recomendação descritos acima nos produtos que tenham um padrão de avaliação ABC.

Para facilitar o entendimento do experimento, definem-se os usuários através das variáveis *u* e *v* e os itens por *i* e *j*. Considerando:

$r_{u,j}(t)$: a avaliação real do usuário *u* ao item *j* no tempo *t*.

$p_{uj}(t)$: a previsão da avaliação do usuário *u* ao item *j* no tempo *t*,

$t_{u,j}$: a data em dias a partir de uma data padrão que o usuário u avaliou o item *j*,

$t_{i,min}$: a primeira data em dias a partir de uma data padrão em que um usuário avaliou item *i*,

$t_{i,max}$: a última data em dias a partir de uma data padrão em que um usuário avaliou item *i*.

L_i: a data em dias a partir de uma data padrão no qual o item *i* foi lançado.

M_j: média das avaliações do filme j

$y_{u,j}$: data em anos a partir de uma data padrão que o usuário u avaliou o item j.

Y_j : média do tempo que as avaliações no filme *j* foram realizadas.

Inicialmente, classificam-se os filmes que possuem um padrão de avaliações que decaem ao longo do tempo. Nessa classificação para cada filme, definem-se três intervalos iguais do período de avaliação.

Define-se o deslocamento $\beta = \left(\frac{t_{i,max} - t_{i,min}}{3}\right)$

O primeiro intervalo A considera as avaliações que ocorreram a partir de $t_{i,min}$ até $t_{i,min} + \beta$.

O segundo intervalo B considera as avaliações que ocorreram depois $t_{i,min} + \beta$ até $t_{i,min} + 2 * \beta$.

O terceiro intervalo C considera as avaliações que ocorreram depois $t_{i,min} + 2 * \beta$ até $t_{i,max}$.

Caso a quantidade de avaliações no intervalo A seja maior que B, B seja maior que C, a quantidade de avaliações no período A seja maior que o somatório das

avaliações realizadas no período B e C e no intervalo C obtiver mais de 10 avaliações, o filme será classificado como ABC.

Algoritmo de Classificação

Para cada filme i

A = 0, B=0, C=0

Para cada avaliação do filme i

Se ($t_{i,min} \leq t_{u,j} \leq t_{i,min} + \beta$)

A = A + 1

Senão($t_{i,min} + \beta < t_{u,j} \leq t_{i,min} + 2 * \beta$)

B = B + 1

Senão($t_{i,min} + 2(\beta < t_{u,j} \leq t_{i,max}$)

C = C + 1

Fim do para cada avaliação

Se (A > B > C E A > (B+C) E C > 10)

O filme i possui um padrão ABC

Fim do para cada filme

O gráfico abaixo representa a ideia do padrão de avaliações de um filme. Conforme o decorrer do tempo, a quantidade de avaliações por intervalo de tempo tende a reduzir. Isso mostra que, ao longo do tempo, menos pessoas se interessam pelo filme.

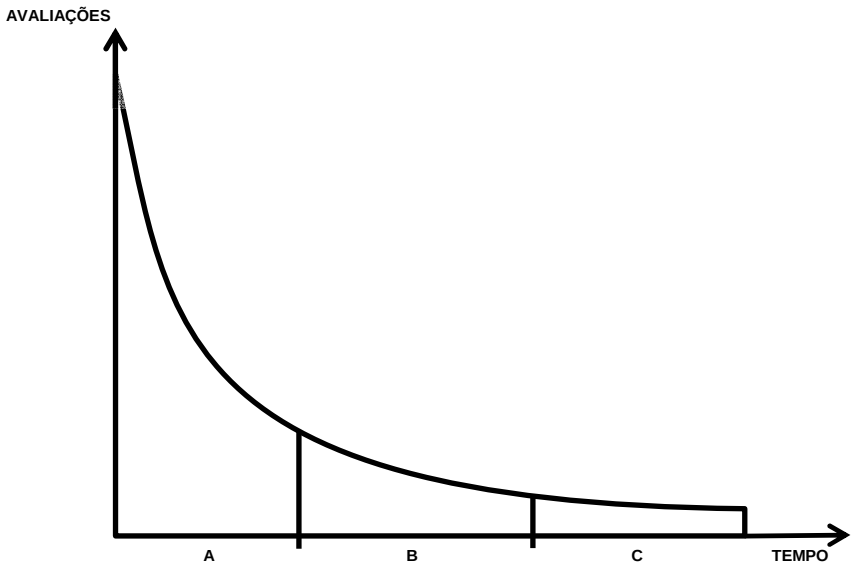

Figura 15 - Gráfico ABC

O primeiro experimento realizado identificou todos os filmes que possuem o padrão de avaliações definido como ABC, conforme descrito acima, obteve-se 858 filmes no total de 17700(aproximadamente 5% dos filmes).

Em cada intervalo A, B, C, calculou-se a média e o desvio padrão das avaliações realizadas em cada um dos intervalos para os 858 filmes classificados como ABC. O desvio padrão indica o afastamento dos valores observados em relação à média aritmética no intervalo estudado. Segue abaixo a fórmula do desvio padrão:

$$S = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n}}$$

Logo, é possível avaliar para cada filme, qual intervalo possui o menor afastamento dos seus valores em relação à média por meio do desvio padrão. O gráfico abaixo contabiliza para cada um dos filmes em qual área obteve-se o menor desvio padrão. Por exemplo, o filme X classificado como ABC, possui desvio padrão no

intervalo C menor que nos outros dois intervalos, portanto, contabiliza-se no intervalo C do gráfico abaixo.

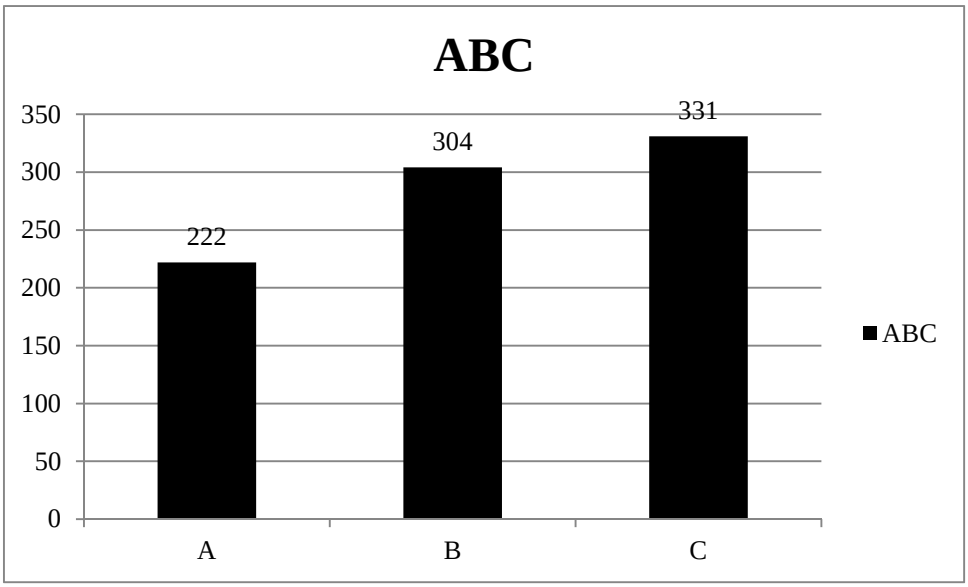

Figura 16 - Similaridade ABC

Através desse primeiro experimento, foi possível comprovar que as pessoas que avaliam o filme no intervalo C diferem menos a suas avaliações que nos outros intervalos. Além disso, demonstrou-se que no intervalo A possui a menor quantidade de filmes com o desvio padrão menor. Apenas 222 vezes dos 858, o desvio padrão no intervalor A é menor que no intervalo B e C.

Logo, as pessoas que avaliam no intervalo C são mais similares do que as que avaliaram em outros intervalos. Essa consideração pode, então, tornar-se um caminho a ser utilizado dentro dos sistemas de filtro colaborativo.

5.1 Similaridade baseada em nicho de mercado

Anderson (2004) relatou que conforme a distribuição da cauda longa, quanto mais o item estivesse na cauda e longe da cabeça, mais específico seria o nicho de pessoas que possuem preferências nesse item. A hipótese dessa dissertação é usar a ideia descrita por Anderson (2004) na dimensão do usuário. Logo, dois usuários que avaliam

na cauda do item (área C) possuem um grau de similaridade maior que os que avaliam em outras áreas, pois eles avaliam filmes de nichos de produtos, tornando-os clientes desses nichos.

O filtro colaborativo baseado em usuários considera que duas avaliações feitas por usuários distintos no mesmo item em intervalos distintos possuem o mesmo peso no cálculo da similaridade entre eles. Consequentemente, as pessoas que assistem ao filme no seu lançamento não são tratadas de forma diferenciadas das pessoas que assistiram após o filme tornar-se destinado a um nicho de mercado. Por exemplo, João, que gosta de filmes *Cult*, assistiu ao filme Poderoso Chefão no ano de 2012. Maria, que gosta de filmes populares, assistiu ao mesmo filme no dia do seu lançamento. Tanto o filtro colaborativo baseado em usuário quando baseado em item concebem o mesmo grau similaridade mesmo os dois tendo assistido em contextos diferentes.

De acordo com o algoritmo tradicional de filtragem colaborativa, se Maria e João avaliam bem o filme, eles são tratados como usuários similares, apesar deles gostarem de estilos de filmes diferentes.

Logo, o comportamento do usuário não sendo analisado pelos filtros colaborativos tradicionais. Abaixo serão listadas as técnicas abordadas a fim de obter melhores resultados na previsão de avaliação para nichos de mercado.

5.1.1 P. ABC 1 - User Based

A similaridade de Pearson será usada considerando um fator $f_{u,v,j}$ como um fator de ajuste da similaridade entre o usuário *u, v* por meio da avaliação em comum do item *j*. A técnica de similaridade abordada é utilizada no sistema de recomendação baseado em usuário.

A similaridade modificada de dois usuários *u* e *v* de Pearson será:

$$r_{u,v} = \frac{\Sigma_j (r_{u,j}(t) - M_j) * (r_{v,j}(t) - M_j) * f_{u,v,j}}{\sqrt{\Sigma_j (r_{u,j}(t) - M_j)^2 * \Sigma_j (r_{v,j}(t) - M_j)^2}}$$

Na fórmula acima, o valor M_j é a média das notas das avaliações no item j.

A variável $f_{u,v,j}$ será definido conforme o algoritmo abaixo:

calcularFatorSimilaridade(u , v)

Para cada filme avaliado em comum por u e v, i

Se(i for classificado como ABC)

Se($t_{u,i}$ E $t_{v,i} forem\ do\ intervalo\ C$)

$f_{u,v,j}$ = 3

Se($t_{u,i}$ E $t_{v,i} forem\ do\ intervalo\ B$)

$f_{u,v,j}$ = 2

Se($t_{u,i}$ E $t_{v,i} forem\ um\ do\ intervalo\ B\ e\ outro\ do\ C$)

$f_{u,v,j}$ = 2

Senão

$f_{u,v,j}$ = 1

Senão

$f_{u,v,j}$ = 1

Fim do para cada

Fim da função

5.1.2 P. ABC 2 - User Based

A segunda abordagem é semelhante a P. ABC 1, todavia o fator de similaridade é alterado pelas fórmulas abaixo.

$$r_{u,v} = \frac{\sum_j (r_{u,j}(t) - M_j * r_{v,j}(t) - M_j) * f_{u,v,j}}{\sqrt{\sum_j (r_{u,j}(t) - M_j)^2 * \sum_j (r_{v,j}(t) - M_j)^2}}$$

calcularFatorSimilaridade(u , v)

Para cada filme avaliado em comum por u e v, j

Se(j for classificado como ABC)

calcularIntervalodaAvaliacao($r_{u,j}(t)$)

calcularIntervalodaAvaliacao($r_{v,j}(t)$)

Se($|A_{u,j} - A_{v,j}| = 0$)

$f_{u,v,j}$ = 2

Se($|A_{u,j} - A_{v,j}| = 1$)

$f_{u,v,j}$ = 1

Se($|A_{u,j} - A_{v,j}| = 2$)

$f_{u,v,j}$ = 0,5

Fim do para cada

Fim da função

O $A_{u,j}$ será calculado da seguinte forma:

calcularIntervalodaAvaliacao($r_{u,j}(t)$)

Se(j for classificado como ABC)

Se($r_{u,j}(t)\ for\ do\ intervalo\ A$)

$A_{u,j}$ = 1

Se($r_{u,j}(t)\ for\ do\ intervalo\ B$)

$A_{u,j}$ = 2

Se($r_{u,j}(t)$ for do intervalo C)

$A_{u,j}$ = 3

Fim do se

Fim da função

5.1.3 P. Média Modificada - User Based

A terceira abordagem modifica a média das avaliações feitas pelo usuário ao considerar o tempo como fator de ajuste na média dele. Realiza-se uma média ponderada, caso o filme no qual usuário tenha avaliado. A técnica abordada é baseada no usuário. Segue abaixo o algoritmo para calcular a média das notas realizadas pelo usuário.

calcularMediaDoUsuario(u , v)

Média = 0, QtdDeAvaliações = 0;

Para cada filme avaliado pelo usuário u, i

Se(i for classificado como ABC)

Se($t_{u,i}\ for\ do\ intervalo\ C$)

Média = Média + 3 * $r_{u,I}(t)$

QtdDeAvaliações = QtdDeAvaliações + 3

Se($t_{u,i}\ for\ do\ intervalo\ B$)

Média = Média + 2 * $r_{u,I}(t)$

QtdDeAvaliações = QtdDeAvaliações + 2

Senão

Média = Média + $r_{u,I}(t)$

QtdDeAvaliações = QtdDeAvaliações + 1

Senão

Média = Média + $r_{u,I}(t)$

QtdDeAvaliações = QtdDeAvaliações + 1

Fim do para cada

M_u= Média / QtdDeAvaliações

Fim da função

5.1.4 P. Tempo Modificado - User Based

A quarta abordagem adiciona na equação de Pearson o tempo como variável para o cálculo da similaridade entre dois usuários.

A equação abaixo só é utilizada em caso do item no qual os usuários avaliaram sejam definidos como ABC. Se o filme não for definido como ABC, será usada a abordagem tradicional.

$$\rho_{u,v} = \frac{\sum_j (r_{u,j}(t) - M_j) * (r_{v,j}(t) - M_j)}{\sqrt{\sum_j (r_{u,j}(t) - M_j)^2 * \sum_j (r_{v,j}(t) - M_j)^2}} + \frac{\sum_j (y_{u,j}(t) - Y_j) * (r_{v,j}(t) - Y_j)}{\sqrt{\sum_j (y_{u,j}(t) - Y_j)^2 * \sum_j (y_{v,j}(t) - Y_j)^2}}$$

5.1.5 Pearson - User Based

Já para a quinta, utilizou-se a filtragem colaborativa baseado em usuário tradicional a fim de comparar com as outras técnicas.

Pearson (u,v)

5.1.6 ABC-Tail 1– Item Based

A técnica abaixo calcula a similaridade entre dois itens. Essa técnica é então inserida no sistema de recomendação baseado em itens. Calcula-se a similaridade entres dois itens por meio da similaridade Pearson entre a média das avaliações e desvio padrão em cada uma das três áreas de um filme que possui o padrão ABC.

```
calcularFatorSimilaridade( i , j)
    Para cada usuário u que avaliou os itens i e j
        Se( i e j foram classificado como ABC)
           Similaridade(i,j) = pearson( md_abc(i), md_abc(j) ) * pearson( i, j );
        Senão
            Similaridade(i,j) = Pearson( i, j );
        Fim se
    Fim para cada
Fim da função
```

Definindo:

m(a,i) = média do filme *i* na área A de um filme com padrão ABC.

d(a,i) = desvio padrão do filme i na área A de um filme com padrão ABC.

```
md_abc(i)
    pearson(m(a,i),d(a,i),m(b,i),d(b,i),m(c,i),d(c,i)
Fim da função
```

5.1.7 ABC-Tail 2 – Item Based

A técnica ABC-Tail 2 é similar a ABC-Tail 1, porém a equação que relaciona a equação de Pearson entre as áreas e avaliações é diferente.

```
calcularFatorSimilaridade(i , j)
    Para cada usuário u que avaliou os itens i e j
        Se( i e j foram classificado como ABC)
            Similaridade(i,j) = 2 * pearson( md_abc(i), md_abc(j) ) * pearson( i, j );
        Senão
            Similaridade(i,j) = Pearson( i, j );
        Fim se
    Fim para cada
Fim da função
```

5.1.8 Pearson - Item Based

Na última abordagem, utilizou-se a filtragem colaborativa baseada em item com a similaridade de Pearson tradicional a fim de comparar com as similaridades ABC-Tail1 e ABC-Tail 2 que são implementadas utilizando a filtragem colaborativa baseada em item.

5.2 Banco de Dados

O banco de dados do Netflix contém mais de 100 milhões de avaliações de filmes feitas pelos clientes do Netflix entre 31 de dezembro de 1999 a 31 de dezembro de 2005. São 480.189 clientes e 17.770 filmes disponíveis para avaliação. A decisão da escolha de uma base com tantas avaliações é pela quantidade de avaliações em um período de seis anos, facilitando a identificação de filmes que já foram populares e ao longo do tempo tornaram-se filmes avaliados por um nicho de mercado.

Os valores da avaliação ficam na faixa de um, indicando interesse baixo pelo item, a cinco, indicando um grande interesse pelo item. Não há possibilidade do usuário avaliar um item duas vezes.

Há outras bases de dados, tais como do MovieLens, 100k, 1M, 10M de dados que possuem avaliações feitas de usuários a filmes, porém o período entre a primeira e a última avaliação são de apenas seis meses, inviabilizando a possível utilização.

A base de dados do Netflix possui avaliações de usuários para item que estão na faixa de 1(não gostou) até 5 (gostou).

Para realização dos experimentos, definiu-se uma quantidade de filmes que serão utilizados na nossa amostra. Foram considerados os quatro mil filmes mais avaliados e seis mil usuários que mais avaliaram filmes.

O total de avaliações realizadas nessa amostra é de 8.629.121. Portanto, o índice de esparsidade, ou seja, a proporção de avaliações com relação ao total é de 35,95% Nessa amostra, um usuário avalia em média 1.438 itens e um item é avaliado em média por 2.157 usuários. Dentre os 858 filmes classificados anteriormente como ABC, 112 pertencem a esta amostra.

Os filmes classificados como ABC recebem em média 1.762 avaliações. Já os filmes que não possuem essa classificação, em média eles recebem 2.168 avaliações. Logo, os filmes classificados como ABC recebem aproximadamente 20% a menos de avaliações.

5.3 Metodologia e Organização do Experimento

A ideia dos experimentos é selecionar uma amostra e analisar a qualidade das previsões por intermédio dos resultados das oito técnicas descritas acima, por meio das métricas MAE e RMSE.

Dados de treinamento são as informações que serão usadas como conhecimento para que o sistema de recomendação seja capaz de prever avaliações aos usuários, enquanto que os dados de testes são as avaliações reais feitas pelos usuários aos filmes que servirão para prever a qualidade das previsões das avaliações.

Abaixo, a imagem explica o funcionamento do experimento. A amostra é armazenada em uma matriz que servirá de entrada para cada um dos três experimentos.

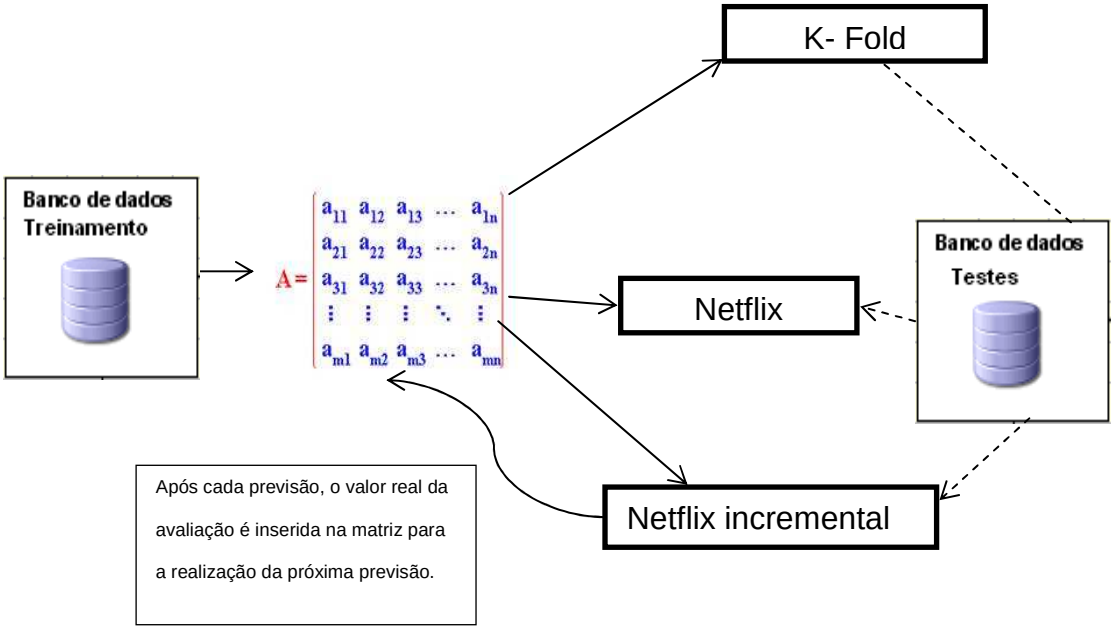

Figura 17 - Experimento

A definição dos dados para treinamento e teste será elaborada por três técnicas distintas por meio de três experimentos. No primeiro experimento, os dados serão separados por meio da técnica *K-fold cross-validation*. Nessa técnica, a amostra original é aleatoriamente dividida em k = 5 amostras a fim de demonstrar uma abordagem de 80% dos dados para treinamento e os 20% restantes para teste.

Uma única fatia da amostra é retida para testes enquanto que as outras serão usadas como dados de treinamento. Após a obtenção dos resultados, a amostra usada como dado de testes é inserida como dado de treinamento e outra fatia de dados dos testes, que não tenha ainda sido escolhida como dado de testes anteriormente, será eleita como dado de treinamento. Quando não há mais amostras que ainda não foram utilizadas como dados de teste, encerra-se a rodada do teste.

A cada rodada definem-se as avaliações que servirão de entrada para o processo de previsão de avaliações (dados de treinamento) e as avaliações que servirão para o

processo de avaliação da qualidade da previsão (dados de testes). A escolha dos dados (avaliações) que serão usadas como dados de treinamento ou testes, em cada rodada, será realizada de forma aleatória.

No primeiro experimento, a técnica K-fold, sendo k= 5, o tempo não é usado como restrição a ser considerado na escolha dos dados que serão usados para treinamento e teste. A fim de aumentar a confiabilidade do experimento, realizaram-se sete rodadas. Esse experimento será chamado de K-FOLD nas próximas referências.

No segundo experimento, as últimas nove avaliações de cada filme, que tenha o padrão ABC, definido anteriormente, serão usadas como dados de testes, o restante das avaliações realizadas serão usadas como dados de treinamento. Nesse experimento considera-se o tempo como fator de escolha dos dados que serão usados como teste e treinamento. Esse experimento será chamado de Netflix nas próximas referências.

O terceiro experimento é semelhante ao segundo, já que as últimas nove avaliações nos filmes realizadas pelos clientes do Netflix serão usadas como dados de testes. No entanto, a cada previsão feita por todos os algoritmos do sistema de recomendação, o dado contendo a avaliação real feita pelo usuário no item será incorporado como dado de teste. Os dados de testes são, então, ordenados de forma crescente em relação ao tempo, a fim de que as restrições temporais sejam consideradas no experimento em cada previsão. Esse experimento será chamado de Netflix incremental nas próximas referências.

Além disso, em todos os experimentos, somente os filmes classificados como ABC são utilizados para previsão das avaliações.

A mesma base de dados foi usada para os três experimentos em cada rodada. O algoritmo usado nesses experimentos foi o de filtragem colaborativa baseada em usuário e filtragem colaborativa baseada em item com 20 usuários de vizinhança.

5.4 Resultados

5.4.1 K-FOLD

Foram realizadas 197.421 previsões de avaliações em cada rodada. O experimento demonstrou que apesar de não considerar o tempo como fator de escolha das previsões, o método P. ABC2 – UB obteve o melhor resultado em todas as rodadas nas métricas MAE e RMSE, seguido pelo método ABC-Tail 1 – IB. Comparando os algoritmos tradicionais Person – UB e Pearson – IB, filtro colaborativo baseado em usuário e item respectivamente, o filtro colaborativo baseado em item obteve o melhor resultado, porém a diferença entre os dois resultados não foi tão significativa.

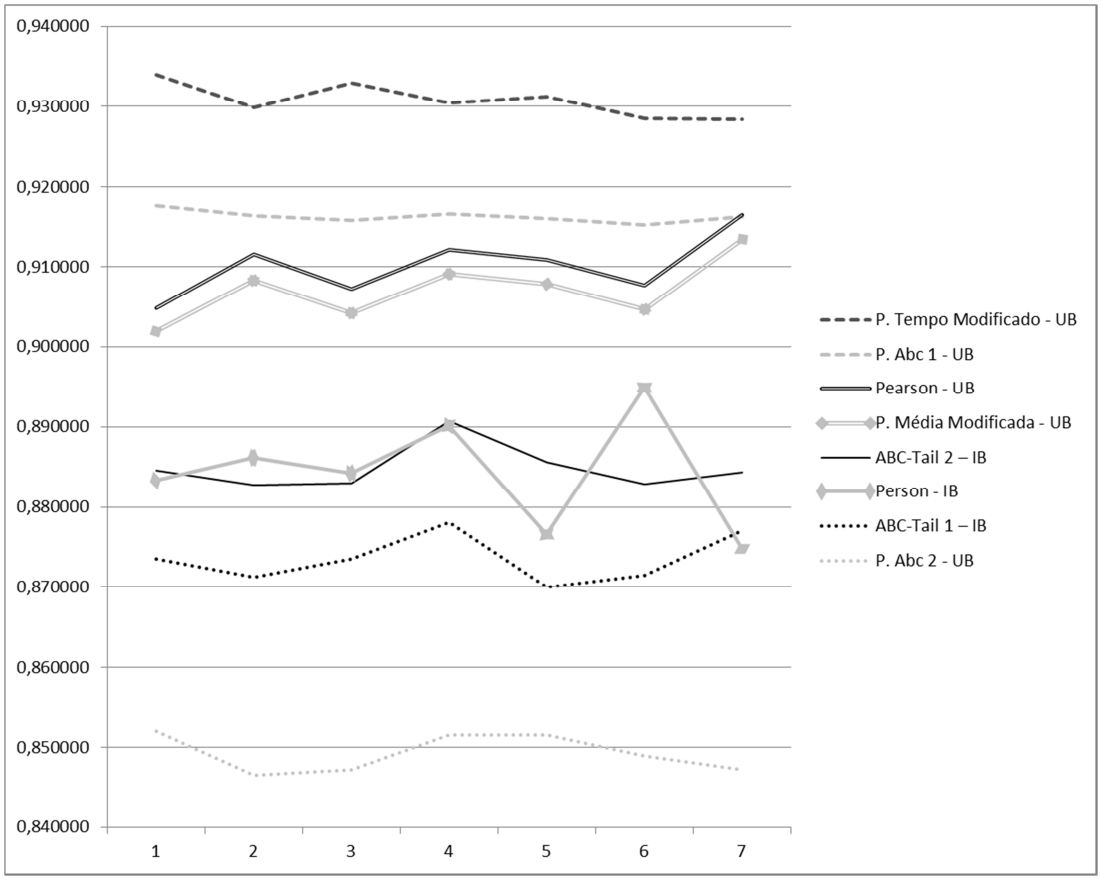

Figura 18 - MAE – K-FOLD

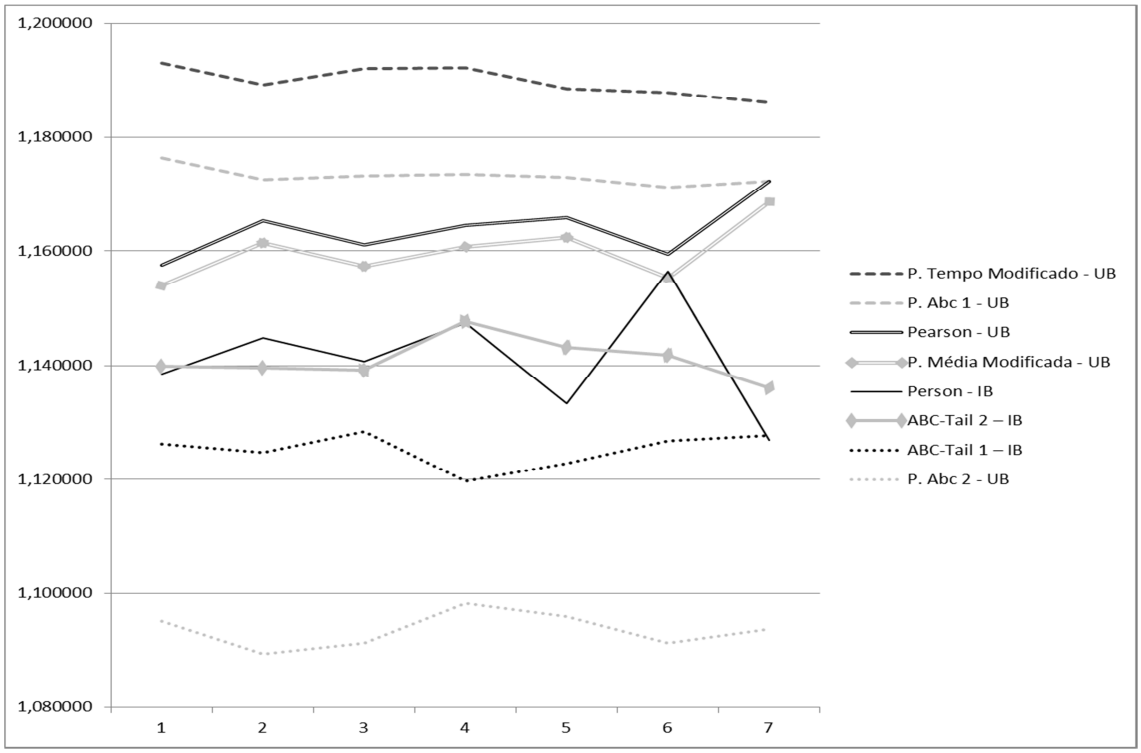

Figura 19 - RMSE – K-FOLD

Seguem abaixo os resultados dos experimentos, que demonstram a qualidade dos resultados obtidos pelas técnicas, no primeiro experimento.

1ª Rodada – K-FOLD		
Algoritmo	**MAE**	**RMSE**
Pearson – UB	0,904790	1,157573
P. Média Modificada - UB	0,901931	1,153945
P. Abc 1 – UB	0,917627	1,176346
P. Abc 2 – UB	0,851955	1,095039
P. Tempo Modificado - UB	0,933990	1,193020
ABC-Tail 1 – IB	0,873520	1,126301
ABC-Tail 2 – IB	0,884556	1,139879
Person – IB	0,883300	1,138490
2ª Rodada – K-FOLD		
Algoritmo	**MAE**	**RMSE**
Pearson – UB	0,911523	1,165321

P. Média Modificada - UB	0,908316	1,161447
P. Abc 1 – UB	0,916296	1,172534
P. Abc 2 – UB	0,846579	1,089325
P. Tempo Modificado - UB	0,929918	1,189217
ABC-Tail 1 – IB	0,871195	1,124707
ABC-Tail 2 – IB	0,882761	1,139611
Person – IB	0,886118	1,144812
3ª Rodada – K-FOLD		
Algoritmo	**MAE**	**RMSE**
Pearson – UB	0,907226	1,161050
P. Média Modificada - UB	0,904251	1,157302
P. Abc 1 – UB	0,915817	1,173304
P. Abc 2 – UB	0,847233	1,091318
P. Tempo Modificado - UB	0,932980	1,192037
ABC-Tail 1 – IB	0,873511	1,128481
ABC-Tail 2 – IB	0,883023	1,139201
Person – IB	0,884230	1,140670
4ª Rodada – K-FOLD		
Algoritmo	**MAE**	**RMSE**
Pearson – UB	0,912158	1,164535
P. Média Modificada - UB	0,909082	1,160726
P. Abc 1 – UB	0,916523	1,173508
P. Abc 2 – UB	0,851582	1,098250
P. Tempo Modificado - UB	0,930446	1,192255
ABC-Tail 1 – IB	0,878105	1,119742
ABC-Tail 2 – IB	0,890710	1,147715
Person – IB	0,890103	1,147474
5ª Rodada – K-FOLD		

Algoritmo	MAE	RMSE
Pearson – UB	0,910863	1,165840
P. Média Modificada - UB	0,907855	1,162295
P. Abc 1 – UB	0,915942	1,173015
P. Abc 2 – UB	0,851532	1,095913
P. Tempo Modificado - UB	0,931239	1,188598
ABC-Tail 1 – IB	0,869937	1,122802
ABC-Tail 2 – IB	0,885547	1,143143
Person – IB	0,876612	1,133397
6ª Rodada – K-FOLD		
Algoritmo	**MAE**	**RMSE**
Pearson – UB	0,907624	1,159426
P. Média Modificada - UB	0,904624	1,155295
P. Abc 1 – UB	0,915211	1,171200
P. Abc 2 – UB	0,848962	1,091266
P. Tempo Modificado - UB	0,928495	1,187865
ABC-Tail 1 – IB	0,871502	1,126753
ABC-Tail 2 – IB	0,882888	1,141835
Person – IB	0,894936	1,156498
7ª Rodada – K-FOLD		
Algoritmo	**MAE**	**RMSE**
Pearson – UB	0,916484	1,172293
P. Média Modificada - UB	0,913500	1,168698
P. Abc 1 – UB	0,916226	1,172220
P. Abc 2 – UB	0,847243	1,093780
P. Tempo Modificado - UB	0,928434	1,186074
ABC-Tail 1 – IB	0,877108	1,127801
ABC-Tail 2 – IB	0,884376	1,136074

Person – IB	0,874807	1,126924

Quadro 3 – K-Fold

Abaixo, segue a tabela contendo a média das métricas MAE e RMSE contabilizados após as sete rodadas.

K-Fold	**MAE**	**RMSE**
Pearson – UB	0,910095297	1,163719758
P. Média Modificada – UB	0,907079822	1,159958388
P. Abc 1 – UB	0,916234562	1,17316099
P. Abc 2 – UB	**0,84929813**	**1,093555833**
P. Tempo Modificado – UB	0,930786144	1,189866647
ABC-Tail 1 – IB	0,873554121	1,125226756
ABC-Tail 2 – IB	0,884837267	1,141065317
Person – IB	0,884300965	1,141180812

Quadro 4 – K-Fold-Resumo

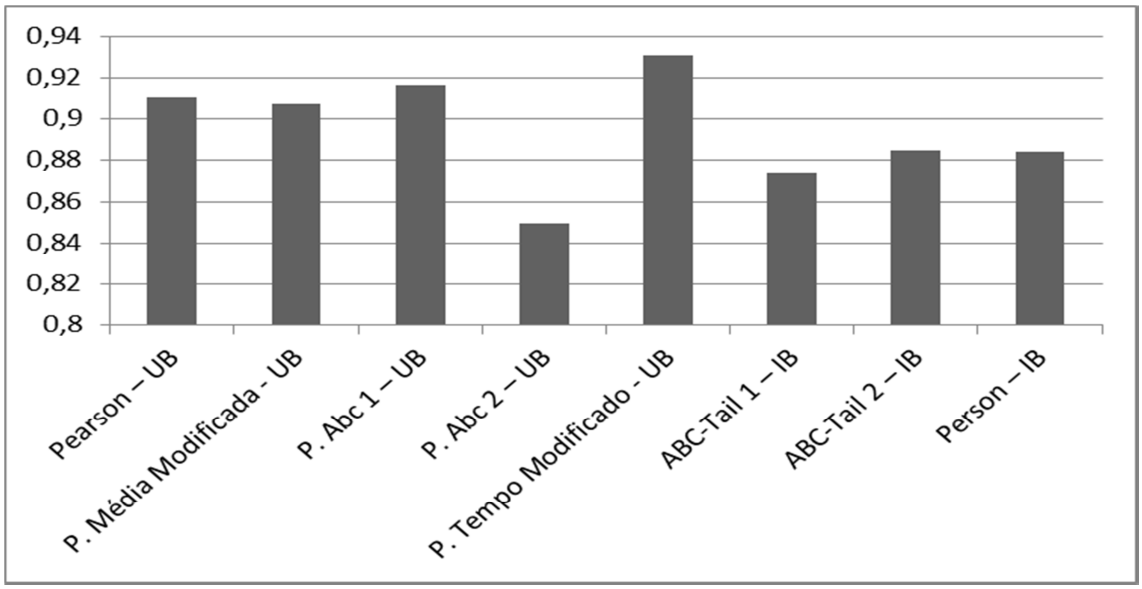

Figura 20 - MAE – K-Fold – Resumo

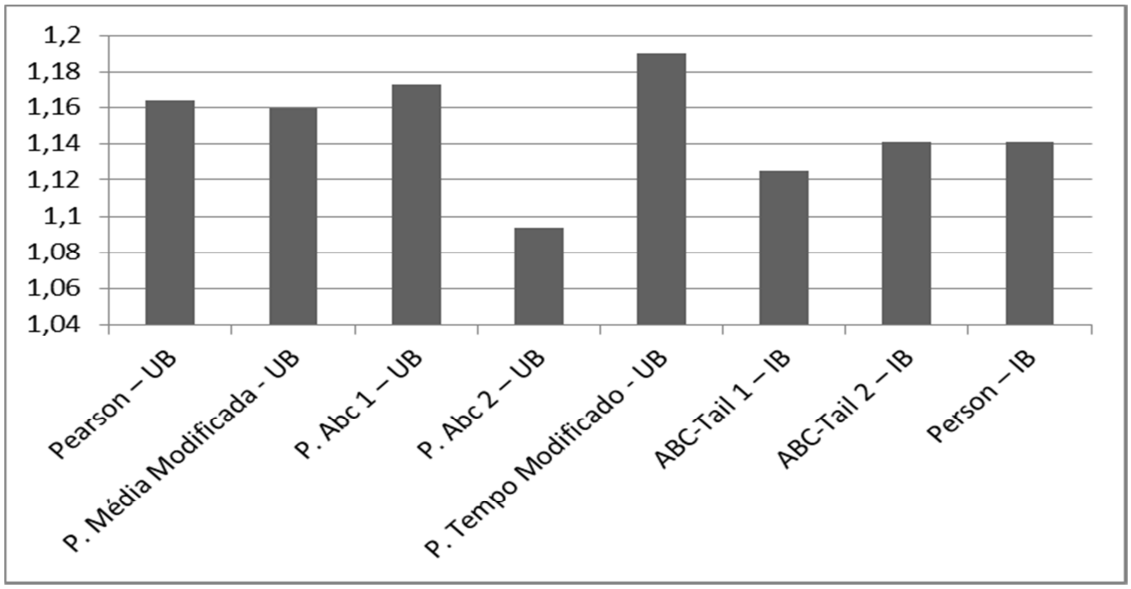

Figura 21 - RMSE – K-Fold- Resumo

5.4.2 Netflix

Nos experimentos Netflix e Netflix incremental, foram realizados 896 previsões. A tabela abaixo demonstrou que o método P. Abc 2 – UB que utiliza o filtro colaborativo baseado em usuário atingiu o melhor resultado. Verificou-se também que para os itens destinados a nichos de mercado, as técnicas baseadas em usuários alcançaram melhores resultados que as baseadas em item. Na métrica MAE, por exemplo, P. Abc 2 – UB foi aproximadamente 5% melhor que o P. Média Modificada – UB, classificado como o segundo melhor método. Em relação ao filtro tradicional baseado em usuário a melhora foi de 6%. Segue abaixo os resultados obtidos dos experimentos Netflix.

Netflix	**MAE**	**RMSE**
Pearson – UB	0,928919	1,193758
P. Média Modificada – UB	0,917843	1,180485
P. Abc 1 – UB	0,997224	1,298057
P. Abc 2 – UB	**0,861319**	**1,095872**

P. Tempo Modificado – UB	0,976053	1,233440
ABC-Tail 1 – IB	0,990862	1,278259
ABC-Tail 2 – IB	1,017510	1,309160
Person – IB	0,975799	1,296277

Quadro 5 – Netflix

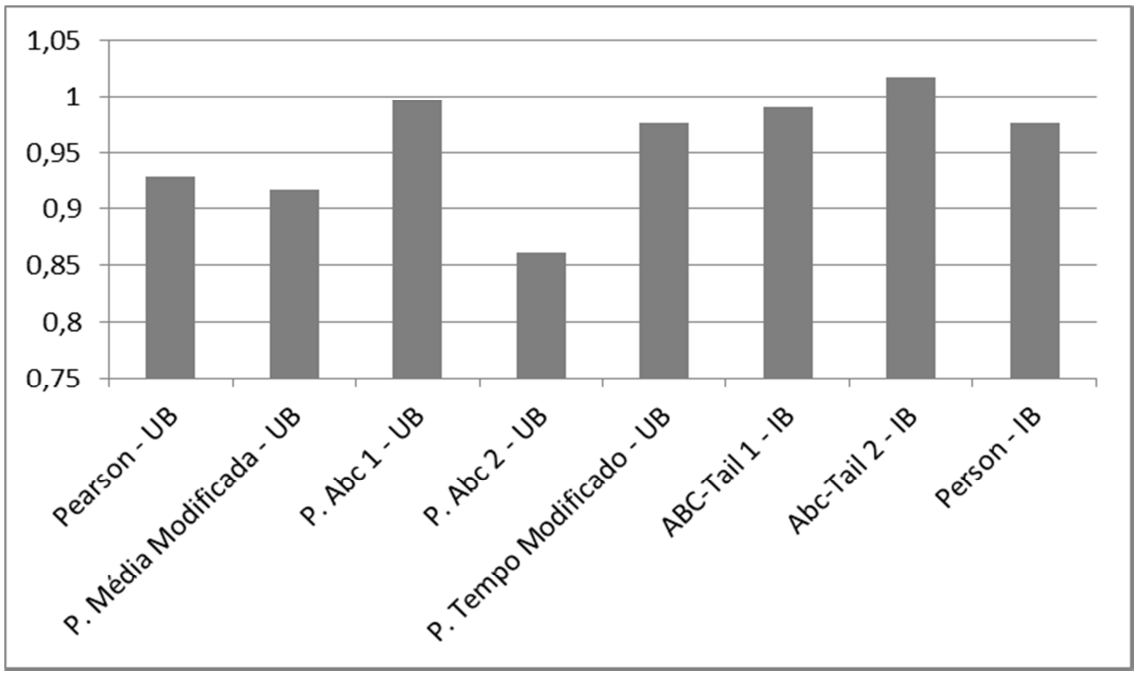

Figura 22 - MAE - Netflix

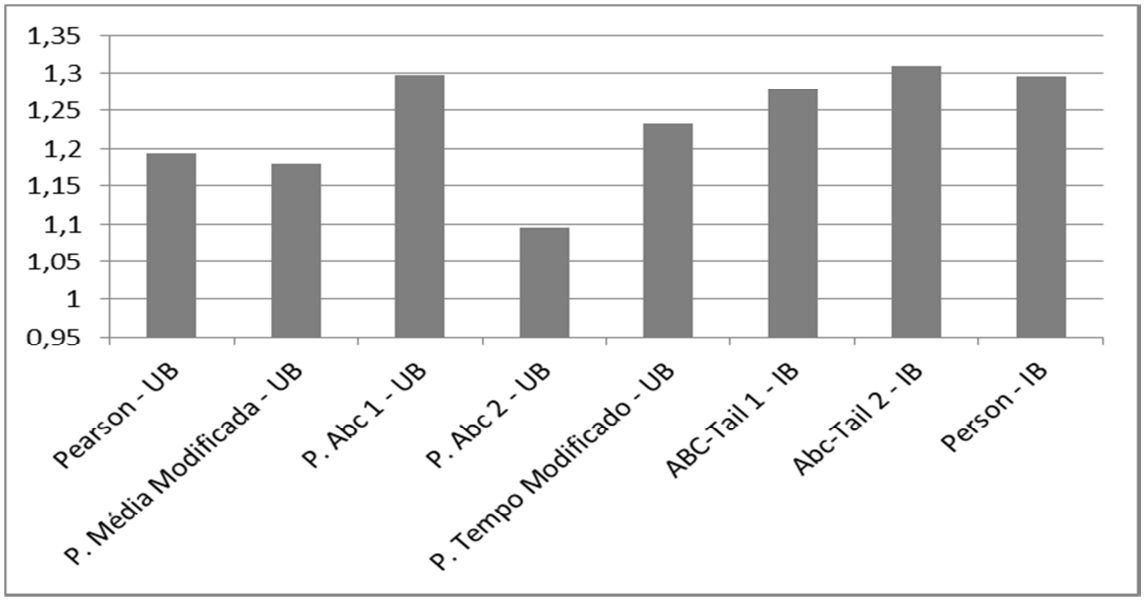

Figura 23 - RMSE – Netflix

5.4.3 Netflix incremental

Esse experimento obteve resultados distintos para cada uma das métricas. A métrica RMSE penaliza os erros grandes em comparando-os aos pequenos. Logo, apesar do método P.Abc 2 – UB obter a melhor métrica MAE , isso não aconteceu em relação ao método RMSE. Na métrica RMSE, o método que conseguiu o melhor resultado foi o P. Tempo Modificado – UB. Segue abaixo a tabela contendo os resultados na tabela e um gráfico das métricas MAE e RMSE.

Netflix incremental	**MAE**	**RMSE**
Pearson – UB	0,947247	1,199133
P. Média Modificada – UB	0,936807	1,187550
P. Abc 1 – UB	0,950563	1,211135
P. Abc 2 – UB	**0,904412**	1,211838
P. Tempo Modificado – UB	0,924119	**1,178421**
ABC-Tail 1 – IB	0,954165	1,223671
ABC-Tail 2 – IB	1,009274	1,285128
Person – IB	1,001597	1,298331

Quadro 6 - Netflix Incremental

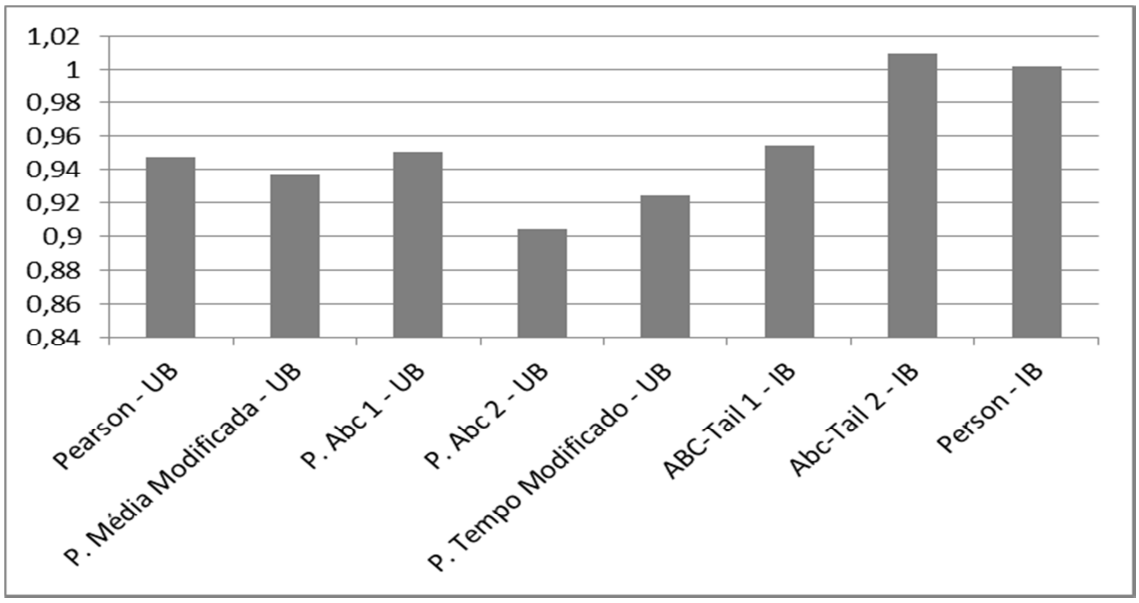

Figura 24 - MAE - Netflix incremental

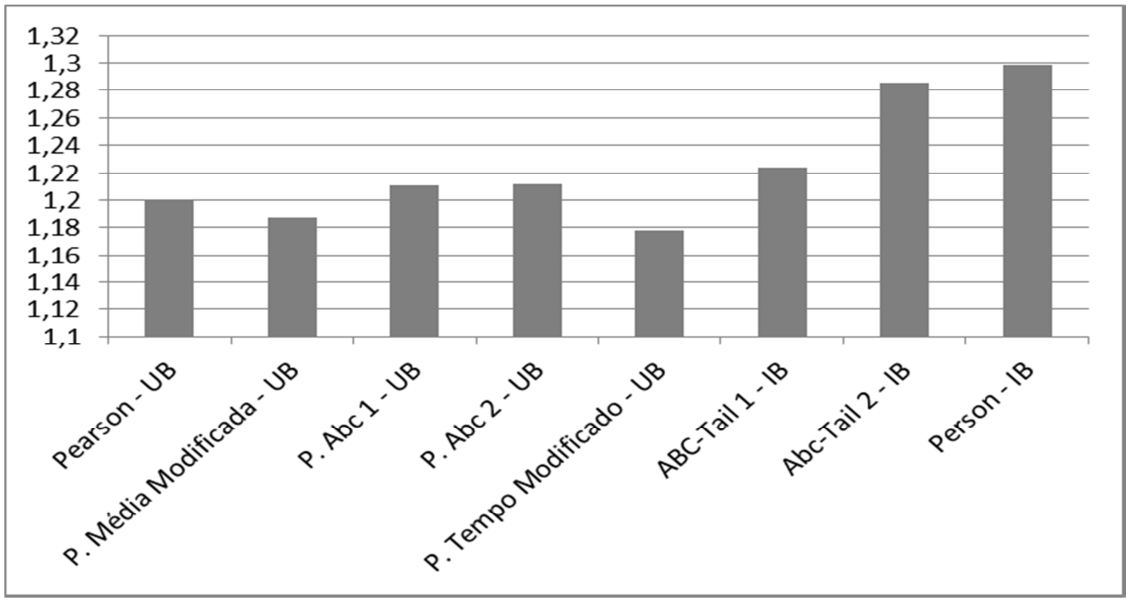

Figura 25 - RMSE - Netflix Incremental

Capítulo 6 – Conclusão e Trabalhos Futuros

Sistemas de recomendação são ferramentas poderosas que agregam valor ao negócio por meio da análise da base de dados. A Internet tem potencial para vender produtos para nichos de pessoas, uma vez que não há limitações de armazenamento, mostruário, logística de produtos como nos tradicionais meios de vendas.

Embora ambas as abordagens baseadas em conteúdo e filtragem colaborativa possuam suas vantagens, eles ainda apresentam problemas de qualidade nas previsões das recomendações em certas situações, especialmente para itens destinados a nichos de mercado.

Para os experimentos desta dissertação, foi utilizada a base de dados da empresa Netflix, que possui avaliações dos filmes realizadas pelos seus clientes. Essa base foi escolhida, pois reflete avaliações reais de clientes de uma empresa bem como uma grande quantidade de avaliações durante um longo período, sendo então possível avaliar mais adequadamente o padrão de avaliação dos itens destinados a nichos de mercado.

Foram realizados três experimentos para demonstrar três cenários diferentes a fim de avaliar os algoritmos propostos para melhorar a qualidade das previsões de avaliações para nichos de mercado. O primeiro cenário não considerou o tempo como uma restrição entre as avaliações que foram usadas para testes e treinamento. No segundo cenário, adicionou-se a restrição do tempo a fim de simular mais adequadamente as previsões das avaliações no mundo real. Já no terceiro experimento, o cenário é semelhante ao segundo, porém, após a cada previsão de uma avaliação de

um item realizada pelo sistema de recomendação, essa avaliação era incorporada na base de testes, gerando uma nova matriz de entrada a cada previsão realizada.

Para avaliar a qualidade das técnicas propostas, foi medida a diferença entre a previsão da avaliação do filme realizada pelo usuário e avaliação real feita por ele, por meio das técnicas *mean absolute error*(MAE) e *root mean square error*(RMSE). Além disso, foi medida a quantidade de itens que cada técnica conseguiu prever em cada um do experimento. No primeiro cenário, para melhorar a confiabilidade dos resultados das métricas abordadas, foram realizadas sete rodadas.

Nos algoritmos propostos baseados em usuários, dependendo do intervalo que ocorreu a avaliação, foi dado um peso maior na similaridade entre os usuários. Em cada um dos algoritmos foi proposto uma forma de estimar quais as avaliações deverão receber um peso maior em relação às outras. Já nas técnicas baseadas em item, foi avaliada a similaridade entre os itens conforme o intervalo que ocorreu a avaliação. Por meio dos resultados obtidos, o tempo demonstrou ser uma informação adequada para as técnicas utilizadas pelos sistemas de recomendação para obter previsões com maior qualidade para nichos de mercado.

Os resultados dos experimentos provaram que houve uma melhora nas previsões das avaliações de itens destinados para os usuários de nichos de mercado. Os resultados para diferentes estratégias demonstraram a dificuldade de conseguir melhores previsões de avaliações nos sistemas de recomendação. Além disso, foi possível identificar por intermédio dos testes realizados que pessoas de um mesmo nicho de mercado possuem gostos mais similares entre eles comparados aos outros usuários.

Concluiu-se também, que há ainda uma gama de melhorias nas recomendações e previsões realizadas pelos sistemas de recomendação que podem ser objetos de pesquisa

posteriores. Muitos dos problemas relatados nesta dissertação ainda não obtiveram uma solução ótima.

6.1 Contribuições

A seguir são apresentadas as contribuições mais relevantes deste trabalho:

- Identificação de um problema referente às previsões e às recomendações realizadas pelos sistemas de recomendação.
- Uma revisão bibliográfica sobre a área de Sistemas de Recomendação, realizada com base em trabalhos clássicos e incluindo os mais recentes desenvolvimentos e propostas na área;
- Definiu-se para o item um padrão de quantidade de avaliações realizadas pelos usuários, ao longo do tempo, a fim de que classificá-lo como sendo destinado a um nicho de mercado.
- Identificou nesse padrão de avaliações dos filmes, um intervalo de tempo no qual os usuários que o avaliam são mais similares entre eles.
- Definição e implementação de técnicas para melhorar a qualidade das previsões dos Sistemas de Recomendação, considerando o tempo e o comportamento dos usuários como fator dessa melhora.

6.2 Limitações

A técnica proposta neste trabalho é uma simplificação de uma realidade complexa e inerente ao contexto da aplicabilidade dos sistemas de recomendação. A seguir, são apresentadas algumas questões que podem comprometer a usabilidade da proposta.

- Essa técnica não trata da questão da esparsabilidade, partida-fria e escalabilidade.

- Os experimentos não consideraram todas as informações da base de dados do Netflix, pois houve uma limitação no tamanho da memória do computador usado para realizar o experimento.
- Definiu-se para o item um padrão de quantidade de avaliações realizadas pelos usuários, ao longo do tempo, a fim de que classificá-lo como sendo destinado a um nicho de mercado.
- Identificou também nesse padrão de avaliações dos filmes, um intervalo de tempo no qual os usuários que o avaliam são mais similares entre eles.
- Há possibilidades de melhorar a eficiência do processamento dos algoritmos usados nos testes realizados.

6.3 Trabalhos Futuros

Apesar das limitações expostas, a proposta oferece recursos para uma abordagem inicial de solução do problema. Entretanto, ela não possui o objetivo de definir e de criar uma técnica que contemple todas as suas nuances, mas apenas abordar seus principais aspectos.

- Há possibilidade de testar os algoritmos de similaridade apresentados na dissertação com outras técnicas abordadas pela área Sistemas de Recomendação, tais como SVD e o K-Means.
- Também há maneiras de elaborar estruturas de dados que possam melhorar a eficiência na detecção de padrões ABC por intermédio dessas estruturas.
- Além disso, há trabalhos a fim de que se possam detectar novos padrões que encontrem nichos de mercado por meio da mineração dos dados.
- Utilizar outras bases de dados para melhorar a confiabilidade dos resultados.

Referências Bibliográficas

ADOMAVICIUS, G., TUZHILIN, A.," Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions", *IEEE Trans. on Knowl. and Data Eng.*, v.17,n.6,pp. 734-749, Piscataway, NJ, USA, Jun 2005.

AGGARWAL, C. C., WOLF, J. L. , WU, K. L., *et al.*, Horting hatches an egg: A new graph-theoretic approach to collaborative filtering. In: *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 201–212, New York, NY, USA, 1999.

BOX, G., JENKINS, G. M., REINSEL, G. C., *Time Series Analysis: Forecasting and Control*, 3ed**,** New Jersey, USA, Prentice Hall, 1994.

BREESE, J. S., HEEKERMAN, D., KADIC, C., "Empirical analysis of predictive algorithms for collaborative filtering**",** In: *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*(UAI),pp. 43-52, Catalina Island, USA, 1998.

BRYNJOLFSSON, E., HU, Y. J., SIMESTER, D., "Goodbye Pareto Principle, Hello Long Tail: The Effect of Search Costs on the Concentration of Product Sales." *Management Science*, Forthcoming, v.57, n.8, pp. 1373-1386, Jan. 2011.

DING, Y., XUE, L., "Time weight collaborative filtering", *In: Proceedings of the 14th ACM international conference on Information and knowledge management"*, pp. 485-492, New York, NY, USA, Nov. 2005.

DING Y., XUE, L., MARIA, E. Orlowska.. "Recency-based collaborative filtering". *In Proceedings of the 17th Australasian Database Conference - Volume 49 (ADC '06), Gillian Dobbie and James Bailey (Eds.)*, V 49, pp. 99-107, Australia, 2006

ELBERSE, A., OBERHOLZER-GEE, F.," Superstars and Underdogs: An Examination of the Long Tail Phenomenon in Video Sales". *Journal: Harvard Business School Working Paper*, Maio 2009.

FLEDER, D.M., HOSANAGAR, K., "Blockbuster Cultures Next Rise or Fall: The Impact of Recommender Systems on Sales Diversity". *NET Institute Working Paper N. 07-10*, New York, NY, 2008

GREENING, D., *Building Consumer Trust with Accurate Product Recommendations***,** LikeMinds White Paper LWSWP-210-6966, 1997

GOLDBERG, D., D. NICHOLS, B.M. OKI, *et al.*, "Using collaborative filtering to weave an information tapestry**",** In: *Communications of the ACM*, v.35, pp. 61-70, Nova Iorque, NY, EUA, Dez. 1992.

GOLDBERG, K., ROEDER, T., GUPTA, D., *et al.*, Eigentaste: A constant time collaborative filtering algorithm. *Inf. Retr***,** *v.4, n2*, pp. 133-151, Hingham, MA, USA, Jul. 2001.

GONG, S., "A Collaborative Filtering Recommendation Algorithm Based on User Clustering and Item Clustering". *Journal of Software, v.* 5, n. 7,pp. 745–752, 2010.

HERLOCKER, J., KONSTAN, J., BORCHERS, *et al*., "An algorithmic framework for performing collaborative filtering." In: *SIGIR '99: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 230–237, New York, NY, USA, 1999

HERVAS-DRANE, A., "Word of Mouth and Recommender Systems: A Theory of the Long Tail**",** *NET Institute Working Paper N. 07-41*, New York University, New York, NY, Nov. 2007.

HILL, W., STEAD, L., ROSENSTEIN, M., *et al*. "Recommending and Evaluating Choices in a Virtual Community of Use", *In: Proceedings of CHI '95*, pp. 194-201, Denver, Colorado, USA, 1995.

HOFMANN, T., "Latent semantic models for collaborative filtering," In: *ACM Transactions on Information Systems*, v. 22, n. 1, pp. 89–115, New York, NY, USA, 2004.

JUNG, J. J., PHAM. X. H. "Attribute selection-based recommendation framework for long-tail user group: an empirical study on movieLens dataset". *In* Proceedings of the Third international conference on Computational collective intelligence: technologies and applications - Volume Part I *(ICCCI'11), Piotr Jedrzejowicz, Ngoc Thanh Nguyen, and Kiem Hoang (Eds.)*, V.1, Springer-Verlag, Berlin, Heidelberg,2011

KONSTAN, J. A., MILLER, B. N., MALTZ, D., *et al*., "GroupLens: applying collaborative filtering to Usenet News**"**, In: *Communications of the ACM* 40, v.40, n. 3, pp.77–87, New York, NY, USA, 1997.

KOREN, Y., "The bellkor solution to the netflix grand prize". *Netflix prize documentation*, 2009a.

KOREN, Y., "Collaborative filtering with temporal dynamics", *In:Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*,pp.447--456,Paris, France,2009b.

KRISHNAN, V., NARAYANASHETTY, P. K., NATHAN, M., *et al.*, "Who predicts better?: Results from an online study comparing humans and an online recommender system". In *Proceedings of the 2008 ACM conference on Recommender systems*, pp, 211–218, New York, NY, USA, 2008.

LAM, S. K., J. RIEDL. "Shilling recommender systems for fun and profit**".** In: *Proceedings of the 13th international conference on World Wide Web*, pp. 393–402, New York, NY, USA, 2004.

LEE, K., K. LEE. "My head is your tail: applying link analysis on long-tailed music listening behavior for music recommendation**".** In *Proceedings of the fifth ACM conference on Recommender systems*, pp. 213–220, New York, NY, USA, 2011.

LEE, T. Q., PARK, Y., PARK, Y. T., "A time-based approach to effective recommender systems using implicit feedback." *Expert systems with applications*, v.34, n.4, Mai. 2008.

LEKAKOS, G., GIAGLIS, G. M., "Improving the prediction accuracy of recommendation algorithms: Approaches anchored on human factors, Interacting with Computers**",** *Interacting with Computers***,** v.18, n.3 pp. 410-431,2006.

LI, B., ZHU, X., LI. R., "Cross-Domain Collaborative Filtering over Time**"**, In: *Proc. Proceedings of the Twenty-Second international joint conference on Artificial Intelligence*,v.3 pp.2293-2298, Barcelona, Catalonia, Spain,2011

LI, Y., LU, L., XUEFENG, L., "A hybrid collaborative filtering method for multiple-interests and multiple-content recommendation in E-Commerce"**,** *Expert Systems with Applications*, v. 28,n.1 pp. 67–77, 2005.

LIEBERMAN, H., "Autonomous Interface Agents", in *Proceedings of CHI'97*, pp.67-74, Atlanta GA, USA, ACM Press, Mar. 1997.

LINDEN, G., SMITH, B., YORK, J., "Amazon. com recommendations: Item-to-item collaborative filtering.", *Internet Computing***,** v.7, n. 1, pp. 76–80,2003.

MAES, P. "Agents That Reduce Work and Information Overload." *Communications of the* ACM v.37,n.7,pp. 31-40, New York, NY, USA, Jul. 1994.

MACQUEEN, J. B. ,"Some Methods for Classification and Analysis of MultiVariate Observations", *In: Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability*,v.1,pp. 281-297,Los Angeles,USA,1967

MELVILLE, P., MOONEY, R. J., R. NAGARAJAN, "Content-boosted collaborative filtering for improved recommendations"**.** In *Proceedings of the National Conference on Artificial Intelligence*, pp. 187–192, CA, USA, 2002.

PARK, Y. J., TUZHILIN, E. A, "The long tail of recommender systems and how to leverage it." In *Proceedings of the 2008 ACM conference on Recommender systems*,pp. 11-18, New York, NY, USA , Nov. 2008.

RASHID, A.M., G. KARYPIS, RIEDL, J., "Influence in Ratings-Based Recommender Systems: An Algorithm-Independent Approach", In: Proc. SDM, Philadelphia,USA, 2005.

RASHID, M. A., Shyong, K. L., KARYPIS, G., RIEDL, J., "ClustKNN: a highly scalable hybrid model-& memory-based CF algorithm". In *Proc. of WebKDD 2006: KDD Workshop on Web Mining and Web Usage Analysis, in conjunction with the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2006),* Philadelphia, PA, Ago. 2006.

RESNICK, P., IACOVOU, N., SUSHAK, M., *et al.*, "GroupLens: An open architecture for collaborative filtering of netnews**"**. In: *Proceedings of the computer supported cooperative work conference,* pp.175-186, New York, NY, USA, 1994

ROSEN, SHERWIN, "The Economics of Superstars,", *Journal: American Economic Review*,v.71,n.6,pp.845-58,Dez 1981.

SALTON, G., BUCKLEY, C., *Term weighting approaches in automatic text retrieval*. Technical Report, pp. 87-881, Ithaca, NY, USA, 1987.

SARWAR, B. M., JOSEPH, A., KONSTAN., *ET AL.*, "Using filtering agents to improve prediction quality in the GroupLens research collaborative filtering system." *In* Proceedings of the 1998 ACM conference on Computer supported cooperative work *(CSCW '98). ACM*, New York, NY, USA,pp.345-354,1998.

SARWAR, B., KARYPIS, G., KONSTAN, J., , *et al.*, "Application of dimensionality reduction in recommender system-a case study." In: *Proceedings of the ACM WebKDD Workshop*, p.158-167, Boston, Massachusetts, USA, Ago. 2000

SARWAR, B., KARYPIS, G., KONSTAN J., *et al.*, Item-based collaborative filtering recommendation algorithms. In: *Proceedings of the 10th international conference on World Wide Web* (WWW '01), pp 285-295, New York, NY, USA,2001.

SCHAFER, J. B., KONSTAN, J., RIEDI, J., "Recommender systems in e-commerce**"**. In *Proceedings of the 1st ACM conference on Electronic commerce*, pp.158–166, ACM, New York, NY, USA, 1999.

SCHEIN, A. I., POPESCUL, A., UNGAR, L. H., *et al.*, "Methods and metrics for cold-start recommendations**"**. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 253–260, New York, NY, USA, 2002.

SHARDANAND, U., MAES, P., "Social Information Filterng: Algorithms for Automating Word Mouth", *In*: *Proceedings of CHI ´95*, pp.210-217, Denver CO, 1995

TANG,T. Y. WINOTO,P., CHAN, K. C. C., "On the temporal analysis for improved hybrid recommendations". *In WI '03 Proceedings of the 2003 IEEE/WIC International Conference on Web Intelligence*, pp. 214, Washington, DC, USA, 2003.

TERVEEN, L. , HILL, W., "Beyond recommender systems: Helping people help each other', In *HCI In The New Millennium*, Jack Carroll, ed., Addison-Wesley,2001.

XIONG, L., CHEN, X., HUANG, T., *et al.,* "Temporal Collaborative Filtering with Bayesian Probabilistic Tensor Factorization**"**, In: *Proc. SDM* , pp. 211-222, Philadelphia,USA, 2010.

WU, L. JOUNG, Y., CHIANG, T. "Recommendation Systems and Sales Concentration: The Moderating Effects of Consumers' Product Awareness and Acceptance to Recommendations". *In* Proceedings of the 2011 44th Hawaii International Conference on System Sciences *(HICSS '11)*,pp.1-10 Washington, DC, USA,2011.