Universidade Federal do Amazonas
Instituto de Ciências Exatas
Departamento de Ciência da Computação
Programa de Pós-Graduação em Informática

Avaliação de Sistemas de Recomendação Baseados em Filtragem Colaborativa

Sheila da Nóbrega Silva

Manaus – Amazonas
Março de 2007

Universidade Federal do Amazonas
Departamento de Ciência da Computação

Autor: Sheila da Nóbrega Silva
Orientador: Prof. Dr. Edleno Silva de Moura

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Informática do Departamento de Ciência da Computação da UFAM, como parte dos requisitos necessários para a obtenção do título de Mestre em Informática. Área de concentração: **Recuperação de Informação**.

Banca Examinadora

Edleno Silva de Moura, Dr. ..... UFAM/PPGI
João Marcos Bastos Cavalcanti, Dr. ..... UFAM/PPGI
Wagner Meira Jr., PhD. ..... UFMG

Manaus, AM
Março/2007

Resumo

Os Sistemas de Recomendação surgiram para facilitar a vida do cliente, ajudando-o a encontrar itens que seriam do seu interesse diante de uma vasta variedade de produtos disponíveis. Vários algoritmos de recomendação têm sido propostos [2, 4, 6, 9, 17, 22] na literatura. Avaliar a qualidade deste algoritmos não é uma tarefa simples [15, 13].

A maioria dos experimentos encontrados na literatura foram avaliados exclusivamente com base em logs, principalmente pela praticidade de avaliar múltiplas técnicas simultaneamente [22, 18, 6]. Alguns autores questionam se este tipo de avaliação é ou não confiável, considerando que as avaliações com pessoas conseguem captar melhor a reação à recomendação e que um sistema com alta acurácia pode em algumas situações levar a resultados errôneos [13, 23, 15].

Nesta dissertação é apresentado um estudo de como os Sistemas de Recomendação Colaborativos são avaliados e é proposta uma forma de validar as avaliações baseadas em logs tornando o uso do log mais confiável. Para realização desta validação é executada além da avaliação baseada em logs, uma avaliação inicial com pessoas. A idéia é comparar os resultados das duas avaliações, identificar e filtrar possíveis ruídos que possam ser os responsáveis pela distância entre o resultado da avaliação baseada em logs e da avaliação com pessoas. O objetivo principal é que após o fim do processo, a base de logs possa ser utilizada como uma coleção de referência mais segura em novos experimentos.

Sumário

Resumo **iii**

Lista de Figuras **vii**

Lista de Tabelas **viii**

1 Introdução **1**

1.1 Motivação 1

1.2 Trabalhos Relacionados 3

1.2.1 Métodos de Avaliação de Sistemas de Recomendação 3

1.2.2 Avaliação de experimentos 5

1.3 Notações 6

1.4 Estrutura do Texto 7

2 Conceitos Básicos **8**

2.1 Sistemas de Recomendação 8

2.1.1 Processo de Recomendação 8

2.1.2 Classificação do Problema da Recomendação 11

2.2 Métodos de Recomendação 12

2.2.1 Baseada em Mineração de Regras de Associação 12

2.2.2 Baseada em conteúdo 14

2.2.3 Baseada em Filtragem Colaborativa 14

2.2.4 Abordagens híbridas 15

2.3 Sistemas de Recomendação Colaborativos 16

2.3.1 Sistema de Recomendação Colaborativo Baseado em usuário . 16

2.3.2 Sistema de Recomendação Colaborativo Baseado em Item . . 17

2.4 Métodos de Avaliação 18

2.4.1 Avaliação baseada em logs 18

2.4.2 Avaliação com pessoas 19
2.5 Métricas de Acurácia 20
2.5.1 MAE (Mean Absolute Error) 20
2.5.2 Hits 21
2.5.3 HR (Hit rate) 21
2.5.4 ARHR (average reciprocal hit-rank) 21
2.5.5 Precisão e Revocação 22

3 Validação de Avaliações Baseadas em Log 24
3.1 Geração das Recomendações 25
3.2 Comparação de Avaliações com Pessoas e Baseada em Logs 26
3.3 Identificação de ruídos no log 27
3.4 Filtragem do Log 28

4 Estudo de Caso 30
4.1 Implementação do Sistema de Recomendação 31
4.1.1 Algoritmo de Recomendação top-N Baseado em Item 31
4.2 Log de transações 35
4.3 Geração das Recomendações 36
4.4 Avaliações do Sistema de Recomendação 38
4.4.1 Avaliação baseada em logs 39
4.4.2 Avaliação com pessoas 39
4.5 Identificação de Ruídos e filtragem do log 42
4.5.1 Filtro dos populares do top-N 42
4.5.2 Filtro dos populares da base de treino 43
4.6 Comparação das avaliações 44
4.7 Alguns Resultados da Aplicação 46
4.7.1 Recomendação Não-Personalizada 48
4.7.2 Produtos não-similares 48

5 Conclusões e Trabalhos Futuros 52
5.1 Trabalhos Futuros 53

Referências Bibliográficas 53

Lista de Figuras

2.1 Exemplo de recomendação do tipo não-personalizada 10
2.2 Exemplo de recomendação do tipo personalizada 11

4.1 Avaliação baseada em Log: ARHR - formação da base de teste M1 . 39
4.2 ARHR - Avaliação com Pessoas 41
4.3 Resultado da avaliação baseada em log: ARHR - filtro dos populares do top-N 43
4.4 Resultado da avaliação baseada em log: ARHR - base treino com filtro dos populares 45
4.5 Resultado da avaliação com pessoas: ARHR - base treino com filtro dos populares 46
4.6 Recomendação não-personalizada: Exemplo - Marca 49
4.7 Recomendação não-personalizada: Exemplo - Cor 49
4.8 Recomendação não-personalizada: Exemplo - Dimensão 49
4.9 Recomendação não-personalizada: Exemplo - Compatibilidade 49
4.10 Gráfico de acurácia X satisfação por categoria de produtos recomendados 50

Lista de Tabelas

4.1 Cálculo da função de utilidade u(A,B) 34
4.2 Características das bases de teste obtidas com as formações M1, M2, M3 e M4 36
4.3 Avaliação baseada em log: Hit rate calculado em bases de teste diferentes. 37
4.4 Resultado do Hits, HR, ARHR e Cobertura de clientes utilizando as formações de base de teste M1 e M2 38
4.5 Resultado da avaliação com Pessoas: Hits, HR e ARHR 41
4.6 Resultado das avaliações baseadas em logs e pessoas 42
4.7 Identificao dos ruídos 43
4.8 Avaliação Baseada em Log:Hits, HR, ARHR e Cobertura de clientes - Base Treino com Filtro dos Populares 44
4.9 Avaliação com pessoas:Hits, HR e ARHR - Base Treino com Filtro dos Populares 44
4.10 Filtro dos populares 45
4.11 Cálculo da distância de Kendall 47

Capítulo 1

Introdução

1.1 Motivação

Os comércios varejistas possuem como característica o grande volume de clientes e também a oferta de uma elevada variedade de produtos. Empresas como submarino.com, por exemplo, possuem milhares de produtos disponíveis para venda, isso também é verdade em lojas varejistas convencionais.

Para os clientes, a definição do produto a ser comprado é um problema muitas vezes resolvido através de conversas com amigos e solicitações de recomendações sobre, por exemplo, que máquina digital comprar diante de tantos modelos disponíveis, ou ainda através de recomendações de novos produtos que pessoas parecidas com eles compraram e gostaram. Em geral, o cliente considera a recomendação baseando-se em três premissas sobre quem está recomendando [13]: Primeiro, o cliente acredita em quem recomenda; segundo: assume que quem recomenda conhece seus gostos ou conhece os gostos de pessoas parecidas com ele; e por último, entende que quem recomenda conhece boa parte das alternativas existentes.

Os Sistemas de Recomendação surgiram para facilitar a vida do cliente, ajudando-o a encontrar itens que seriam do seu interesse diante de uma vasta variedade de

produtos disponíveis através da recomendação de um conjunto de N itens que seriam do seu interesse. Do lado do varejista, estes sistemas passaram a tornar possível a personalização em vários canais de contato com o cliente, seja no site de e-commerce [14], em uma loja convencional, televendas ou através de mala direta.

Vários algoritmos de recomendação têm sido propostos [2, 4, 6, 9, 17, 22] na literatura. Avaliar a qualidade deste algoritmos não é uma tarefa simples [15, 13]. Em geral as avaliações são realizadas de duas formas:(1)Baseada em um log de transações de vendas: neste caso o log é dividido em duas partes, sendo uma para treino dos algoritmos de recomendação e a outra parte para avaliar a capacidade de predição dos mesmos; (2) Baseada em pessoas: neste tipo de avaliação, pessoas são convidadas a avaliar a recomendação com base em um histórico de compras utilizado como treino.

A maioria dos experimentos encontrados na literatura possuem avaliações baseadas em logs. Esta forma de realizar experimento é escolhida pelo seu baixo custo, praticidade e também pela dificuldade de serem conduzidos experimentos controlados no meio acadêmico. Embora possua muitas vantagens, a avaliação baseada em logs pode levar a resultados errôneos, considerando que um alto índice de predição não implica necessariamente em um sistema de recomendação de qualidade [23]. Por exemplo, em um estudo de caso que realizamos utilizando o log de transações de uma loja de departamentos, verificamos uma grande disparidade entre os resultados das avaliações realizadas com log de transações e o resultado das avaliações realizadas com pessoas.

Tal disparidade se deu em nosso caso porque a recomendação de itens populares, que o cliente compraria independentemente da qualidade da recomendação, produz bons resultados quando avaliamos os sistemas utilizando logs, mas produz resultados ruins quando os sistemas são avaliados por pessoas. Por exemplo, oferecer um cartão de celular (item popular) a quem está comprando uma cama de casal não faz sentido,

no caso o cliente normalmente esperaria uma sugestão de item complementar ao produto, como o colchão adequado às dimensões da cama, um criado-mudo com acabamento similar, etc. Apesar disso, se o cliente costuma comprar cartões na loja, a avaliação baseada em logs tenderá a considerar a recomendação do cartão como boa, ainda que os produtos não sejam considerados como relacionados por uma pessoa. Observe que neste caso a recomendação do cartão tende a ser ruim porque o cliente normalmente já compra o cartão na loja e porque não considera o contexto da compra.

Os objetivos deste trabalho são: implementar um Sistema de Recomendação e realizar um estudo sobre a forma de avaliação deste sistema; propor uma forma de validar as avaliações baseadas em logs e mostrar como o log utilizado no estudo de caso pôde ser filtrado para reduzir a distância entre a avaliação baseada em logs e a avaliação com pessoas.

1.2 Trabalhos Relacionados

1.2.1 Métodos de Avaliação de Sistemas de Recomendação

Os principais trabalhos relacionados ao estudo de formas de avaliar de sistemas de recomendação são:

Hayes et al.[13] propuseram um método on-line de avaliação de Sistemas de Recomendação para complementar a avaliação off-line (baseada em logs), considerando que a satisfação do usuário com a técnica de recomendação pode ser medida somente de forma on-line (com pessoas). Propuseram uma arquitetura similar a uma arquitetura padrão de Sistemas de Recomendação, a principal diferença é que ao invés de uma única máquina de recomendação existiriam duas máquinas de recomendação. No caso, o usuário receberia simultaneamente em um sistema on-line

as recomendações geradas por duas técnicas implementadas e então avaliaria a recomendação. Como trabalhos futuros, além da necessidade de um teste prático do método, os autores evidenciaram a necessidade da realização de uma comparação direta entre avaliações off-line e on-line. Em nosso trabalho nós realizamos esta comparação, mostramos uma forma de medir esta disparidade utilizando métricas de distância entre os resultados e apresentamos um exemplo de como um log pode ser filtrado para reduzir a distância no estudo de caso que realizamos.

McNee et al.[23] acreditam na premissa que os sistemas de recomendação com altas taxas de acurácia ,conforme as métricas utilizadas, em algumas situações não geram recomendações consideradas úteis pelos usuários. Os autores propuseram a seguinte maneira de realizar avaliação: Primeiro, julgar a qualidade das recomendações da forma em que os usuários vêem as recomendações, como uma lista de recomendações. Destacam a necessidade de métricas que atuem sobre a lista inteira, não apenas sobre um item que está aparecendo na lista. Segundo, ir além dos resultados das analises off-line utilizando o *feedback* dos próprios usuários do sistema de recomendação. Por último, entender o propósito da recomendação considerando o momento em que o usuário está iteragindo com o sistema de recomendação. Seria julgar as recomendações de cada usuário tendo como base se o sistema é ou não capaz de conhecer as suas necessidades daquele momento.

Os dois trabalhos a seguir, evidenciam a necessidade ainda atual de experimentos voltados à definição de uma metodologia para a avaliação de Sistemas de Recomendação:

Adomavicius et al.[8] escreveram um survey sobre o que seria o estado-da-arte em Sistemas de Recomendação e destacaram que ainda são necessários experimentos de qualidade elevada para entender os benefícios e limitações de uma técnica de recomendação proposta.

Herlocker et al.[15] fez um levantamento amplo dos fatores que são considerados

em avaliações de Sistemas de Recomendação e identificou fatores que influenciam nos resultados, como o tipo dos dados, o propósito da recomendação e outros. Mostrou como os sistemas tem sido avaliados e a diferença que os torna incomparáveis, mesmo quando as mesmas métricas são utilizadas em trabalhos distintos. São apresentados também resultados de avaliações empíricas sobre várias métricas, inclusive sobre as métricas chamadas de métricas de não-acurácia e que são voltadas para identificação da satisfação do usuário. A necessidade do desenvolvimento de mais métodos padronizados para avaliação de Sistemas de Recomendação faz parte da conclusão deste trabalho.

1.2.2 Avaliação de experimentos

Um extenso estudo feito por [15] sobre avaliação de Sistemas de Recomendação Colaborativos indica quais métricas são mais apropriadas considerando determinados objetivos da recomendação. O mesmo estudo também mostra a necessidade da utilização de métricas padronizadas para que seja possível realizar comparações entre os diversos algoritmos de recomendação que são propostos. Apesar deste estudo, ainda podem ser encontrados na literatura grupos de pesquisadores utilizando métricas distintas e várias propostas de novas métricas.

Encontramos vários trabalhos na literatura onde são comparados diferentes métodos de implantação de Sistemas de Recomendação e são propostos novos algoritmos. Os principais trabalhos e as métricas utilizadas para realização das avaliações são relacionados a seguir: Sarwar et al.[21] fizeram a análise de algorítmos de recomendação para e-commerce. O objetivo da análise era identificar escalabilidade e qualidade dos sistemas de recomendação. Os experimentos utilizaram dois tipos diferentes de dados, o histórico de compras da loja fingerhut.com e a base de dados do MovieLens[1]. A base de dados da loja fingerhut.com continha 6502 Clientes, 25.554

[1]http://www.movielens.org/

produtos e um total de 97.045 registros de compras. Para medir a qualidade dos experimentos utilizaram como métrica as curvas de precisão, revocação e F1. Huang et al.[24] propuseram a utilização de um grafo para modelar Sistemas de Recomendação para e-commerce. A avaliação foi feita utilizando-se curvas de precisão e revocação. Karypis et al.[6, 11, 18] utilizaram as seguintes métricas medir a qualidade dos algoritmos propostos:(1) Número de Hits , que é o número de transações na base de teste que possuem algum produto que coincide com a lista de top-N produtos recomendados para cada cliente. A lista de top-N recomendações é uma lista contendo N produtos em ordem decrescente de relevancia, onde N refere-se à quantidade de recomendações que será gerada pelo Sistema de Recomendação. (2)Hit-rate, que é o número de hits/n, onde n é a quantidade de clientes; ARHR (average reciprocal hit-rank), métrica que considera a posição do acerto. Sarwar et.al [21], Karypis et al. [11] e Ziegler et al.[4] utilizaram para medir a qualidade dos algoritmos propostos adaptações das curvas de precisão e revocação, métricas amplamente utilizadas em Recuperação de Informação.

1.3 Notações

Como a aplicação do Sistema de Recomendação estudada aqui está voltada para o contexto de um comércio varejista, nesta dissertação serão usados os termos *clientes* e *produtos* como sinônimos de usuários e itens, respectivamente. Os usuários para os quais deseja-se computar as recomendações serão referenciados como *clientes ativos*. Utilizaremos o termo *log* para denotar um conjunto de transações com os registros de compras dos clientes, sendo que cada cliente possui uma única transação contendo todos os produtos comprados em um determinado período e cada compra em particular convencionou-se chamar de *ticket*.

1.4 Estrutura do Texto

Esta dissertação está dividida em 5 capítulos. No Capítulo 2 são apresentados os conceitos básicos para o entendimento deste trabalho. No Capítulo 3 é apresentada uma proposta de como deve ser feita a validação de avaliações baseadas em logs. No Capítulo 4 é apresentado o estudo de caso realizado para validação de avaliações realizadas com logs. Neste capítulo estão descritas as características do Sistema de Recomendação implementado, as definições das formações das bases de treino e teste e os resultados e comparações entre as avaliações baseadas em logs e com pessoas. Neste mesmo capítulo também são apresentadas algumas observações sobre o Sistema de Recomendação implementado. Por fim, no Capítulo 5, são apresentadas as conclusões gerais do trabalho proposto bem como algumas perspectivas de trabalhos futuros.

Capítulo 2

Conceitos Básicos

2.1 Sistemas de Recomendação

Um Sistema de Recomendação pode ser definido como um sistema de filtragem de informação utilizado para identificar um conjunto de produtos que seriam de interesse de determinado cliente.

2.1.1 Processo de Recomendação

O primeiro passo do processo de recomendação é a entrada de dados, que podem ser transações de compras, dados demográficos dos clientes, características dos produtos, comentários, dentre outros. Esses dados são chamados de votos e podem ser obtidos de duas formas [19]:

- Explícita: O próprio cliente indica explicitamente a sua opinião sobre um determinado item, indicando o seu nível de satistação. Por exemplo, em uma escala de 1 a 5, uma estrela indicaria insatisfação com o produto e 5 estrelas o nível máximo de satisfação.
- Implícita: As avaliações ou votos são obtidos de forma implícita, através da ex-

tração de um perfil do cliente identificado conforme navegação no site, histórico de compras realizadas pelo cliente (assim o registro da compra de um produto já demonstra um forte interesse do cliente), etc.

Após a coleta dos dados, um método ou uma combinação de métodos são implementados para geração das recomendações. Em geral as recomendações são geradas considerando-se a correlação entre produtos, correlação entre clientes ou uma combinação destas informações.

De acordo com o grau de personalização utilizado será necessário que o cliente identifique-se para que as recomendações sejam apresentadas. O grau de personalização pode ser categorizado da seguinte forma [14]:

- **Não-Personalizado**: Neste caso, o cliente não precisa ser identificado e ao realizar uma consulta por determinado produto (em um site de e-commerce ou em uma loja física) a mesma recomendação que aparece para o cliente aparece para todos os outros clientes. O exemplo apresentado na Figura 2.1 é um exemplo típico de recomendação não personalizada onde os produtos fortemente relacionados são apresentados juntos com o objetivo principal de realizar uma venda adicional de um monitor, impressora ou nobreak a quem está olhando uma CPU. A recomendação é gerada com base no comportamento de compras dos clientes que no passado compraram o produto em questão.

- **Persistente ou Personalizado**: Na recomendação personalizada, como o próprio nome enfatiza, a recomendação é gerada para um cliente individualmente, sendo necessário coletar informações que identifiquem de forma única o cliente ativo. Por exemplo, na Figura 2.2 um cliente específico que havia comprado uma impressora, um cartão de recarga de celular, uma munhequeira e um cadeado, recebeu como recomendações um cartão de recarga de maior valor, cartuchos para impressora e um tênis.

- **Efêmero**: Os produtos que fizerem parte do carrinho de compras do cliente são considerados em conjunto para geração de recomendações, no caso de sites de e-commerce, os produtos selecionados, as buscas e outros comportamentos da navegação do cliente entre as categorias de produtos também são considerados durante o processo de geração das recomendações. Compras anteriores não são consideradas e o cliente não é reconhecido pelo sistema de recomendação.

Fig. 2.1: Exemplo de recomendação do tipo não-personalizada

A apresentação das recomendações pode ser feita de várias formas. O cliente pode indicar explicitamente que quer visualizar as recomendações ou as recomendações podem ser mostradas sem que haja uma solicitação do cliente, por exemplo, em um site de e-commerce através da leitura do *cookie*[1], a página inicial já é aberta mostrando as recomendações específicas para o cliente ativo. Outra forma também é apresentar as recomendações sem que esteja explicitado para o cliente que são recomendações geradas especificamente para ele.

Assim como em outros sistemas, existem duas preocupações a serem consideradas no projeto de um Sistema de Recomendação [21], a primeira é o falso positivo, que é

[1]Grupo de dados trocados entre o navegador e o servidor de páginas, colocado num arquivo (ficheiro) de texto criado no computador do usuário.

Fig. 2.2: Exemplo de recomendação do tipo personalizada

a recomendação de produtos que não são de interesse do Cliente; A segunda, é o falso negativo, que seria não recomendar produtos que seriam de interesse do Cliente. O mais crítico é o falso positivo, pois impacta diretamente na credibilidade do Sistema de Recomendação e pode levar a não utilização do sistema.

2.1.2 Classificação do Problema da Recomendação

O problema da recomendação pode ser visto de duas formas distintas:

1. **Problema da Predição** [8, 11]: consiste em predizer se um cliente em particular irá gostar de um produto aleatório estimando um voto para este produto com base em votos anteriormente atribuídos pelo cliente a outros produtos. Ou seja, se reduz a estimar votos aos produtos ainda não vistos pelo cliente [8]. No caso, são recomendados os produtos com maiores estimativas de voto.

O Sistemas MovieLens, por exemplo, utiliza informações explícitas quando ao gosto de um usuário por determinados filmes para em seguinda predizer o quanto o mesmo usuário iria gostar de outros filmes.

2. **Problema da Recomendação top-N** [11, 21]: neste caso, o problema é identificar o conjunto de N itens que seriam de interesse de determinado cliente. Em geral, o problema da recomendação top-N pode ser formulado da seguinte forma [8]: Seja C o conjunto de todos os clientes e seja P o conjunto de todos os produtos candidatos a recomendação, existe uma função u que mede a utilidade do produto p para o usuário c, u:C X $P \rightarrow R$, onde R é um conjunto ordenado. Para cada cliente $c \in C$ pretende-se obter os N produtos $p' \in P$ com maior taxa de utilidade. A utilidade de um item é representada por um voto, obtido de forma implícita ou explícita, evidenciando algum interesse do usuário por um determinado item.

Como no estudo de caso que realizamos a base de dados disponível é o log de transações de compras e não votos atribuídos pelos cliente indicando explicitamente uma preferência ou não por determinados produtos, daremos ênfase neste trabalho ao problema da recomendação top-N.

2.2 Métodos de Recomendação

Muitas abordagens têm sido propostas, a seguir relacionamos as principais abordagens utilizadas em Sistemas de Recomendação de produtos.

2.2.1 Baseada em Mineração de Regras de Associação

A técnica mais utilizada em mineração de dados para geração de recomendação é a extração de regras de associações entre produtos e clientes a partir de um con-

junto de históricos de transações. Basicamente, esta técnica consiste em descobrir associações entre dois conjuntos de produtos de forma que a ocorrência de algum produto em uma determinada transação implique que outros produtos de outras transações também estarão presentes. Ou seja, a associação $X \Rightarrow Y$, é a representação da probabilidade condicional de ocorrer Y dado que X ocorre, onde X e Y são conjuntos de produtos.

As recomendações usando regras de associação podem ser geradas da seguinte forma [21]:

1. Considerar como entrada um conjunto de transações, onde cada transação contém todos os produtos comprados por um cliente. Chamaremos de transação o histórico de compras de um cliente.

2. Executar um algoritmo para encontrar todas as regras de associações entre conjuntos de produtos, considerando um limiar de suporte (percentual de transações que contém X e Y) e confiança (percentual das transações que contém X e que também contém Y). Onde:

$$\text{suporte} = \frac{X \cap Y}{\text{total de transações}} \tag{2.1}$$

$$\text{confiança} = \frac{X \cap Y}{\text{quantidade de transações que contém X}} \tag{2.2}$$

3. Para gerar as top-N recomendações de um cliente m, deve-se selecionar todas as regras que contenham no conjunto X os itens do histórico de compras do cliente.

4. Selecionar todos os conjuntos Y das regras selecionadas no passo anterior e descarte produtos que já tenham sido comprados pelo Cliente.

5. Ordenar os produtos em ordem crescente de *confiança* e recomendar os top-N produtos.

2.2.2 Baseada em conteúdo

Nesta abordagem são recomendados produtos similares aos produtos comprados anteriormente pelo cliente. Formalmente , a utilidade $u(p, c)$ do produto p para o cliente c é estimada com base nas utilidades $u(p_i, c)$ atribuídas pelo cliente c aos produtos $p_i \in P$ que são similares ao produto p [8]. Por exemplo, para recomendar um CD para um cliente, o sistema de recomendação busca produtos similares aos comprados anteriormente pelo cliente considerando puramento o texto (descrição do item) e características, como artista e gênero musical. Para determinadas categorias de produtos esta abordagem funciona, mas para outras categorias não. Um exemplo seria: Em uma recomendação personalizada qual seria a utilidade de recomendar geladeiras similares a geladeira que o Cliente acabou de comprar? Já em uma recomendação não personalizada, pode ser útil apresentar recomendações de geladeira similares (características como cor, preço e funcionalidades, por exemplo) a geladeira que o cliente está consultando. A principal vantagem do Sistema de Recomendação baseado em conteúdo é não necessitar de um avaliador ou uma quantidade de compras expressiva para gerar recomendações. Por outro lado, possui a desvantagem de especializar as recomendações de produtos, recomendando sempre produtos similares aos produtos comprados ou vistos anteriormente pelo cliente.

2.2.3 Baseada em Filtragem Colaborativa

Consiste em recomendar produtos, ainda não comprados pelo cliente ativo, que outros clientes com preferências similares compraram no passado. O sistema Tapestry [12] foi a primeira implementação de Sistema de Recomendação baseado em

filtragem colaborativa, em uma pequena comunidade de usuários, o sistema permitia que os usuários registrassem reações aos documentos lidos e essas anotações das reações eram vistas por outros usuários, sendo possível filtrar documentos avaliados como releventas por um usuário determinado. Atualmente, a filtragem colaborativa é a técnica mais utilizada para construir sistemas de recomendação. Como esta dissertação é sobre avaliação de Sistema de Recomendação baseado em filtragen colaborativa, na Seção

2.2.4 Abordagens híbridas

As abordagens híbridas são utilizadas combinando a abordagem baseada em conteúdo e a abordagem colaborativa com o objetivo de reduzir as limitações das duas abordagens. Em [8] as diferentes formas de combinar os métodos colaborativo e basedo em conteúdo dentro de um sistema de recomendação híbrido são classificadas em uma das categorias a seguir:

- Implementação dos métodos colaborativo e baseado em conteúdo separadamente e posteriomente combinação das predições.
- Incorporação de algumas características do modelo baseado em conteúdo dentro do modelo colaborativo.
- Incorporação de algumas características do modelo colaborativo dentro do modelo baseado em conteúdo.
- Construção de um modelo unificado que incorpora características do modelo colaborativo e do modelo baseado em conteúdo.

2.3 Sistemas de Recomendação Colaborativos

Sistemas de recomendação colaborativos utilizam a opinião de uma comunidade de usuários para ajudar usuários dentro da própria comunidade a identificarem conteúdos, produtos, serviços e até mesmo pessoas, de seu interesse, de forma mais eficiente.

Para gerar recomendações com base em informações colaborativas é necessário coletar opiniões de clientes a respeito dos produtos. Em Sistemas de Recomendação para comércio, a opinião dos clientes pode ser obtida através de votos (avaliações) atribuídos aos produtos pelos clientes. Estes votos são em geral, histórico de compras realizadas pelos clientes ou um valor numérico onde o cliente expressa seu gosto pelo produto.

Na literatura, encontramos duas abordagens para construção de Sistemas de Recomendação baseados em filtragem colaborativa: Baseado em usuário e baseado em Item, também chamados de baseado em memória e baseado em modelo [22].

2.3.1 Sistema de Recomendação Colaborativo Baseado em usuário

Em linhas gerais consiste em: Identificar k clientes que são mais similares ao cliente ativo; Computar a união entre os produtos comprados por todos os k clientes e associar um peso para cada produto baseado na sua importância dentro do conjunto; Por último, dentro desta união, gerar a recomendação dos N produtos de maior peso e que ainda não foram comprados pelo cliente ativo. Desta forma, para construir um Sistema de Recomendação baseado em usuário é necessário [11]:

1. Contruir uma matriz R de dimensão $m \times n$, onde a dimensão m representa os clientes e a dimensão n os produtos e $r_{ij} = 1$ caso o cliente i tenha comprado o produto$j$ e $r_{ij} = 0$, caso contrário.

2. Computar a similaridade entre clientes. Esta similaridade é comumente calculada utilizando a correlação de Person ou o conseno. Para cada cliente c, encontrar uma lista de x clientes C=$c_1, c_2, ..., c_x$ em ordem crescente de similaridade, tal que $\text{sim}(c, c_1) > \text{sim}(c, c_2) >.. > \text{sim}(c, c_x)$.

3. Gerar a lista de recomendações. Basta selecionar os produtos comprados por todos os clientes em C mais similares ao cliente ativo, acumulando a frequência dos produtos candidatos a recomendação e descartando os produtos já comprados anteriormente pelo cliente ativo. Serão recomendados os N mais frequentes. Uma outra forma seria gerar regras de associação entre os produtos comprados pelos clientes mais similares, mas tem como desvantagem ser em muitas situações uma base pequena tornando difícil a identificação de um comportamento de compra.

Uma grande vantagem do Sistema de Recomendação utilizando filtragem colaborativa baseado em cliente é justamente a capacidade de detectar um comportamento de compra e consequentemente gerar recomendações "não óbvias". Mas, possui fortes limitações relacionadas a: esparsidade, pois enquanto a matriz m x n é bastante esparsa a matriz m x m é muito densa; e escalabilidade, já que a complexidade de tempo deste método cresce linearmente com o número de clientes, o que dificulta a sua utilização por grandes empresas varejistas onde existem milhões de clientes e produtos.

2.3.2 Sistema de Recomendação Colaborativo Baseado em Item

A abordagem de recomendação colaborativa baseada em item surgiu para resolver o problema de escalabilidade existente na abordagem baseada em usuário. Várias

abordagens tem sido desenvolvidas utilizando similaridades entre itens, [1, 22, 18, 17] que em linhas gerais consistem em [18, 17]:

1. Construir um modelo que captura o relacionamento entre itens.

2. A partir deste modelo pré-computado determinar as top-N recomendações para um cliente ativo.

Conforme estudos realizados por [22, 11], algoritmos de recomendação baseado em item apresentam melhor performance computacional e qualidade equivalente ou superior aos algoritmos baseados em usuário. Estes foram o principais motivos pela escolha desta abordagem para implementação do Sistema de Recomendação e utilização no estudo de caso. No Capítulo 4, apresentamos em detalhes o algoritmo de recomendação top-N baseado em item que implementamos.

2.4 Métodos de Avaliação

Um estudo realizado por [8], mostra a importância do tipo da recomendação para definição da métrica mais apropriada a ser utilizada na avaliação da qualidade da recomendação. Nesta sessão, explicaremos as métricas mais utilizadas na literatura em avaliações de Sistema de Recomendação.

2.4.1 Avaliação baseada em logs

Este tipo de avaliação é a mais utilizada na literatura principalmente pela praticidade, considerando-se que com pouco esforço é possível realizar avaliações de múltiplas técnicas e também repetir experimentos. Em geral, o log é dividido em duas partes, uma parte para treino (chamada de base de treino) dos algorítmos e a outra para teste (chamada de base de teste). No caso, as recomendações são geradas

tendo como entrada os dados da base de treino e para medir a qualidade destas recomendações são utilizadas as compras do cliente na base de teste, sendo que neste tipo de avaliação quanto mais o sistema conseguir predizer as compras da base de teste melhor será o seu resultado. Este tipo de avaliação vem sendo questionado por vários autores [13, 15, 23] partindo-se da premissa de que um sistema com acurácia não significa necessariamente um sistema que gera recomendações úteis para os clientes. O principal ponto deficitário desta forma de avaliação é que não é possível captar a reação do cliente a uma recomendação e consequentemente medir a sua real satisfação com a utilidade do produto recomendado.

2.4.2 Avaliação com pessoas

Na avaliação com pessoas, clientes reais ou voluntários são convidados a opinarem sobre a qualidade do Sistema de Recomendação. As avaliações com pessoas são mais difíceis de serem realizadas pelo custo da avaliação quando comparadas à avaliação baseada em logs. A grande vantagem da avaliação com pessoas é a capacidade de captar a satisfação do cliente e identificar ocorrências de falso positivo, problema bastante crítico em um sistema de recomendação. A situação ideal seria que os próprios clientes avaliassem as recomendações durante o processo de compra, mas não é uma situação viável por deixar exposto um sistema experimental que em determinados momentos poderá gerar recomendações ruins. Uma desvantagem da avaliação em um ambiente experimental é que não existe a intenção de compra.

Na avaliação com pessoas é necessário definir como as recomendações geradas pelos algoritmos serão apresentadas. Em [13] são identificadas três alternativas de apresentação das recomendações:

- **Misturadas**: neste caso, o avaliador irá visualizar um conjunto uníco de recomendações, ou seja, a união das recomendações geradas por todas as técnicas

que fazem parte do experimento. Desta forma, os produtos duplicados serão visualizados uma única vez e o avaliador não saberá qual técnica está avaliando.

- **Separadas**: cada conjunto de recomendações resultante de uma técnica será apresentado separadamente para o avaliador, de tal forma que o avaliador saiba exatamente que está avaliando técnicas diferentes.

- **Em cascata**: a recomendações resultantes das diferentes técnicas são apresentadas de forma alternada.

2.5 Métricas de Acurácia

Na literatura encontramos diversas métricas que são utilizadas para avaliar a qualidade de um Sistema de Recomendação. As métricas de acurácia são em geral categorizadas como, métricas de acurácia estatística e métricas de acurárica de suporte a decisão. A primeira avalia a acurácia de um sistema comparando um *score* numérico de recomendação e o voto atribuído pelo cliente ao par cliente-item dentro da base de teste. A segunda avalia o quão efetivamente uma máquina de predição ajuda um cliente a selecionar produtos de qualidade entre os demais produtos disponíveis [21].

2.5.1 MAE (Mean Absolute Error)

Mede a capacidade que o algoritmo tem de predizer o voto que seria dado por um cliente a um produto selecionado aleatoriamente. É classificada como métrica de acurácia estatística.

$$MAE = \sum_{i=1}^{N} |p_i - v_i|/N \tag{2.3}$$

onde, p_i é o valor da predição e v_i é o valor do voto.

Esta métrica é aplicada aos sistemas de recomendação que tratam o problema da predição do voto que o cliente daria a um produto específico, mas para o sistema de recomendação que tiver como objetivo gerar o ranking dos melhores produtos que seriam de interesse do cliente, esta métrica não é apropriada [15].

2.5.2 Hits

Para computar o número de hits, dividi-se o log de transações em duas partes, sendo uma para treino, chamada de base de treino, e a outra para teste, chamada de base de teste. As recomendações são geradas com base nos dados contidos na base de treino. O número de Hits, ou acertos, é o total de produtos na base de teste que também estão presentes na lista de produtos recomendados para cada cliente. É computado no máximo um acerto por cliente.

2.5.3 HR (Hit rate)

A métrica mais recentemente usada na literatura para medir a qualidade dos Sistemas de Recomendação cujo objetivo é gerar uma lista com top-N recomendações, é o HR (Hit-rate). Seja n o número de clientes, o hit-rate é calculado da seguinte forma:

$$HR(hit - rate) = \text{número de hits}/n \tag{2.4}$$

2.5.4 ARHR (average reciprocal hit-rank)

Como o hit-rate trata igualmente os acertos independentemente da posição em que o item aparece dentro da lista de recomendações, escolhemos o $ARHR$ por considerar a ocorrência do acerto dentro do top-N gerado para cada cliente. Desta

forma, quanto mais perto do topo da lista forem os hits maior será o seu peso. O cálculo do $ARHR$ é:

$$ARHR = 1/n \sum_{i=1}^{h} 1/p_i \tag{2.5}$$

Onde h é o número de acertos nas posições p_i para cada lista top-N gerada. Sendo $1 <= p_i <= N$.

2.5.5 Precisão e Revocação

São métricas amplamente utilizadas em Recuperação de Informação. Em Sistemas de Recomendação, a precisão mede a acurácia de predição do comportamento de compra do cliente e a revocação, mede a relevância da recomendação para um cliente. Em [15], existe a seguinte definição destas métricas no contexto de Sistemas de Recomendação:

$$Precisão = \frac{N_{RL}}{N_R} \tag{2.6}$$

onde, N_{RL} é o total de produtos relevantes recomendados e N_R é o total de produtos recomendados.

$$Revocação = \frac{N_{RL}}{N_{RLD}} \tag{2.7}$$

onde, N_{RL} é o total de produtos relevantes recomendados e N_{RLD} é o total de produtos relevantes disponíveis.

Algumas adaptações destas métricas foram sugeridas [21, 11] e a proposta por [11] acabou sendo adotada em vários trabalhos encontrados na literatura e é igual ao hit rate, ou seja, a revocação é igual ao número de produtos contidos na base de teste e que também fazem parte das top-N recomendações retornadas para cada

cliente.

$$Revocação = \frac{Teste \cap top - N}{n} \tag{2.8}$$

onde n é o número de clientes.

Capítulo 3

Validação de Avaliações Baseadas em Log

A grande maioria das avaliações de Sistemas de Recomendação encontradas na literatura é baseada em logs. Essa forma de avaliação é escolhida principalmente pela praticidade, considerando-se que com pouco esforço é possivel avaliar um grande número de algoritmos. Já a avaliação com pessoas possui um custo elevado e por isso muitas vezes acaba não sendo considerada, apesar de ser a avaliação que melhor consegue captar a reação do cliente à recomendação [13, 15, 23].

A questão é se a avaliação baseada em log é suficiente ou não para indicar se um sistema de recomendação é melhor ou pior que outro. A situação ideal seria a realização de avaliações baseadas em log com resultados os mais próximos possíveis das avaliações com pessoas, ou sabendo-se qual a distância entre as avaliações. Neste trabalho é proposta uma forma de verificar se a avaliação baseada em um dado log pode ser utilizada de forma confiável e apresentando resultados próximos de avaliações com pessoas. Para realização desta validação deverá ser executada necessariamente, além da avaliação baseada no log, uma avaliação inicial com pessoas. A idéia principal é comparar os resultados das duas avaliações, identificar e filtrar

possíveis ruídos que possam ser os responsáveis pela distância entre o resultado da avaliação baseada em log e a avaliação com pessoas. O objetivo principal é que, uma vez realizada a validação, novos experimentos possam ser realizados apenas com logs, mas de maneira confiável. Desta forma, após o fim do processo, a base de logs possa ser utilizada como uma coleção de referência mais segura em novos experimentos.

Neste capítulo serão apresentados os passos necessários para validação de experimentos realizados com logs. Na Sessão 3.1 são apresentados alguns pontos a serem considerados na geração das recomendações para que os ruídos possam ser evidenciados. Na Sessão 3.2 é definido o modo como as avaliações serão comparadas. Na Sessão 3.3 serão mostradas algumas formas de identificar quais são os produtos que podem gerar ruídos na avaliação baseada em logs. Por fim, na Sessão 3.4 é apresentada uma sugestão de como pode ser feito o filtro de um log e quais os impactos que essa filtragem pode trazer para o Sistema de Recomendação.

3.1 Geração das Recomendações

Alguns pontos na geração das recomendações deverão ser considerados para realização da validação de uma avaliação baseada em log :

- É necessário que seja explorado um número significativo de possíves recomendações. A geração de recomendações as mais variadas possíveis facilita a identificação de ruídos que possam distorcer o resultado da avaliação baseada em log quando comparada com a avaliação com pessoas.

- Os algoritmos devem ser executados sobre a mesma base de treino para que as recomendações geradas sejam comparáveis entre si. As recomendações que serão avaliadas por pessoas deverão ser uma amostragem das recomendações

geradas para avaliação baseada em log. As recomendações devem ser armazenadas em uma base de recomendações de tal forma que seja possível identificar qual algoritmo originou cada recomendação. A partir desta base de recomendações única podem ser realizadas em paralelo as avaliações baseadas em log e as avaliações com pessoas.

3.2 Comparação de Avaliações com Pessoas e Baseada em Logs

Após as avaliações com pessoas e avaliações baseadas em log serem concluídas é necessário compará-las para identificar se as avaliações baseadas em log produzem resultados que refletem a percepção de qualidade indicada pelas pessoas.

Uma forma de comparar a avaliação com pessoas e a avaliação baseada em log seria simplesmente medir as distâncias entre os resultados, mas como os resultados não são compatíveis, considerando-se que a primeira mede satisfação e a segunda capacidade de predição, optou-se por comparar a ordem de qualidade atribuída aos vários sistemas estudados nas duas avaliações. Com estas listas sabe-se exatamente quais algoritmos geram recomendações melhores ou piores na avaliação baseada em log e na avaliação com pessoas. Quanto maior a diferença entre as listas menor é a utilidade da avaliação baseada em log.

Para comparar a distância entre estas listas é proposta a utilização da distância de Kendall [16], por ser de fácil entendimento e uma métrica amplamente utilizada na literatura. Sendo L_1 a lista com os algoritmos em ordem decrescente de acurácia conforme avaliação baseada em log e L_2 a lista com os algoritmos em ordem decrescente de avaliação da utilidade das recomendações conforme avaliação com pessoas. Utilizando-se a fórmula para medir a distância de Kendall entre duas listas L_1 e L_2

temos:

$$K(L_1, L_2) = \frac{\sum_{i,j \in P} K'_{i,j}(L_1, L_2)}{(n(n-1)/2)} \tag{3.1}$$

onde, P é o conjunto dos pares dos distintos elementos em L_1 e L_2;

$K'_{i,j}(L_1, L_2) = 0$, se i e j estão na mesma ordem dentro das listas L_1 e L_2;

$K'_{i,j}(L_1, L_2) = 1$, se i e j não estão na mesma ordem dentro das listas L_1 e L_2;

A distância é normalizada dividindo-se o resultado por $(n(n-1)/2)$, onde n é o tamanho da lista. Assim, o valor 1 indica o nível máximo de divergência entre L_1 e L_2.

Se a distância entre as listas não for satisfatória, ou seja, próxima de 1, uma estratégia para reduzir esta distância é filtrar entradas do log que possam ser os responsáveis pela distância entre as avaliações.

3.3 Identificação de ruídos no log

Uma forma de identificar entradas no log responsáveis pela distância entre o resultado da avaliação com pessoas e o resultado da avaliação baseada em logs é analisar os produtos não relevantes na avaliação com pessoas e que são classificados como relevantes na avaliação baseada em logs e tentar encontrar características genéricas entre estes produtos. É importante identificar estes produtos pois são eles que certamente geram os falsos positivos, problema que pode comprometer a credibilidade do sistema de recomendação.

Em um Sistema de Recomendação para um comércio varejista os produtos populares certamente serão fortes candidatos a gerarem falsos positivos. Em geral são produtos que o cliente compraria independentemente da recomendação. Recomendá-los pode parecer óbvio demais e algumas vezes sem sentido. Um ponto a ser destacado é que em um Sistema de Recomendação onde o log de transações

possui produtos de uma única categoria, tais como Livros, CDs e DVDs, não é tão evidente o impacto dos populares, pois acaba sendo indicado um popular da mesma categoria. Mas, quando há muitas categorias, os produtos populares podem gerar recomendações avaliadas como estranhas ou sem utilidade pelas pessoas. Por exemplo, o cliente seleciona uma geladeira e o sistema recomenda o CD da Banda Calypso.

Após identificação dos produtos responsáveis pelos falsos positivos na avaliação com log, deve-se retirar esta entrada do ranking das recomendações geradas e realizar novo cálculo da eficácia de predição do algoritmo, comparando-se o resultado com a avaliação com pessoas. Quanto mais próximos forem os resultados, melhor é a idenficação dos ruídos, pois significará que a simulação da avaliação com pessoas está sendo realizada com maior exatidão pela avaliação baseada em logs.

3.4 Filtragem do Log

Uma vez identificados os produtos que podem estar ocasionando ruídos, o log pode ser filtrado de duas formas:

1. Antes de Gerar as Recomendações: Neste caso, a parte do log de transações que compõe a base de treino deve ter um filtro dos produtos identificados como ruídos. Deve-se observar que a filtragem do log pode resultar em uma cobertura menor de produtos recomendados e de clientes com recomendações, consideran-se que os produtos fitrados não serão recomendados.

2. Na geração da lista top-N: No momento de gerar o ranking das recomendações são descartados os produtos geradores de ruídos. Mesmo nesta situação ainda há uma diminuição na cobertura de clientes que recebem recomendações. Uma solução para amenizar este problema seria não recomendar os produtos gera-

dores de ruído somente em situações em que hajam outros N produtos a serem recomendados.

Capítulo 4

Estudo de Caso

Este estudo de caso tem como principal objetivo realizar a validação de avaliações de Sistemas de Recomendação top-N baseadas em log seguindo os passos descritos no Capítulo 3. Para realização deste estudo de caso teve-se acesso a um log de transações de vendas de uma empresa varejista que possui lojas convencionais, um site de e-commerce e lojas híbridas (com venda por catálogo). Foi implementado no site[1] de e-commerce desta empresa um sistema de recomendação personalizado e não personalizado. A principal característica deste log é a diversidade, sendo cerca de 35.000 produtos ativos em mais de 100 categorias diferentes. Os dados obtidos são implícitos, sendo que a compra de um determinado produto por um cliente é uma forte demonstração de interesse no produto e o fato do cliente não comprar um produto em particular não significa que o cliente não goste deste produto.

Na Sessão 4.1 são apresentados os detalhes de implementação do Sistema de Recomendação. Na Sessão 4.2 são descritas as características do log de transações utilizado neste estudo de caso. Na Sessão 4.3 é apresentada a forma na qual as recomendações foram geradas e um detalhamento de como foi escolhida a formação das bases de treino e teste. Na Sessão 4.4 são apresentados os resultados das ava-

[1] http://www.bemol.com.br

liações baseadas em logs e avaliações com pessoas. Na Sessão 4.5 são identificados e filtrados os ruídos encontrados no log de transações. Na Sessão 4.6 são comparados os resultados das avaliações e finalmente na Sessão 4.7 são apresentadas algumas observações

4.1 Implementação do Sistema de Recomendação

Na implementação do Sistema de Recomendação optou-se por utilizar o Algoritmo de Recomendação top-N baseado em Item pela escalabilidade, por obter bons resultados quando comparado à recomendação colaborativa baseada em usuários [22, 11] e por já ser um algoritmo implementado em sites comerciais, como por exemplo o site da Amazon.com [17].

4.1.1 Algoritmo de Recomendação top-N Baseado em Item

Este algoritmo parte do princípio de que os produtos comprados no passado podem determinar uma preferência de compra no futuro, considerando o comportamento de outros clientes que compraram os mesmos produtos e também outros produtos. Os outros produtos comprados são os produtos candidatos à recomendação.

A construção do modelo pode ser definida como: Seja C o conjunto de todos os clientes, P o conjunto de produtos e T o conjunto de transações, onde cada transação t' é composta por um cliente c' e os respectivos produtos p' comprados. Contruir uma matriz R de dimensão $n \times n$, onde n é o número de produtos e $r_{ij} = u(p_i, p_j)$, sendo $u(p_i, p_j)$ uma função que para cada produto $p_i \subset P$ mede a utilidade de recomendar um produto $p_j \subset P$, considerando a ocorrência simultânea de p_i e p_j em cada transação $t' \subset T$.

Foi implementado o mesmo algoritmo descrito em [17], conforme pode ser visto no Algoritmo 1, primeiramente nas linhas 02 a 08 registra-se na matriz $item \times item$

Algoritmo 1 Cálculo da matriz *item* × *item*

```
01 início
02   para cada produto p_i ⊂ P
03     para cada cliente c que comprou p_i
04       para cada produto p_j comprado por c
05         Atualizar Matriz item × item (p_i, p_j)
06       fim
07     fim
08   fim
09   para cada produto p_j
10     Computar função de utilidade u(p_i, p_j)
11   fim
fim
```

a ocorrência simultânea de produtos dentro da mesma transação de compras, em seguida, nas linhas 09,10 e 11, faz-se o cálculo da utilidade de recomendar um produto p_j dado que o cliente demonstrou algum interesse pelo produto p_i.

Uma vez computada a matriz que calcula a utilidade entre produtos $u(p_i, p_j)$, a recomendação não-personalizada pode ser realizada simplesmente indicando para um produto p_i os N produtos p_j relacionados em ordem decrescente de utilidade $u(p_i, p_j)$. A recomendação personalizada é realizada com base nos produtos comprados anteriormente pelo cliente ativo. Conforme descrito no Algoritmo 2, para gerar a recomendação personalizada é passado como parametro o cliente ativo c e o seu histórico de compras t', nas linhas 03 a 05 é feita a união dos produtos candidatos à recomendação de cada $p \subset t'$, em seguida na linha 06 a função GeraTopN(r,N) retorna os top-N mais relacionados já excluindo-se ocorrências de produtos comprados anteriormente pelo cliente.

Função de Utilidade

Embora na literatura seja bastante utilizado o termo similaridade entre produtos, neste trabalho é utilizado o termo utilidade de um produto dada a acorrência de outro, ou simplesmente utilidade. Um exemplo prático seria: qual a utilidade

Algoritmo 2 Recomenda(c,t'): Recomendação personalizada

```
01 r: matriz[1,n]{n é a quantidade de produtos distintos}
02 início
03    para cada produto p ⊂ t'
04        r = SomaLinha(p_i)
05    fim
06    GeraTopN(r,N)
07 fim
```

de se recomendar um depurador de ar considerando-se que um fogão foi comprado, no caso, A e B não são produtos similares. Para que fossem geradas diversas recomendações distintas o calculo da função de utilidade $u(A, B)$ foi feito de 8 formas distintas (vide Tabela 4.1). A maioria das fórmulas utilizadas é baseada na probabiliade condicional de compra de um item considerando que outro item foi comprado anteriormente. Em [11] é possível encontrar um estudo detalhado sobre a utilização de probabilidade condicional para cálculo da função de utilidade. Sendo $P(B/A)$ a probabilidade condicional do cliente comprar um produto B, considerando que um produto A foi comprado anteriormente, então $P(B/A)$ pode ser definida [11] como a quantidade de clientes que compraram os produtos A e B, dividido pela quantidade de clientes que compraram o produto A, sendo

$$P(B/A) = \frac{Freq(A, B)}{Freq(A)}, \tag{4.1}$$

onde Freq(X) é definido como o número de clientes que compraram produtos dentro do conjunto X. Este cálculo claramente tende a beneficiar os itens populares, ou seja, $P(A, B)$ resultará em valores elevados que não refletirão exatamente a ocorrência simultânea de A e B e sim a popularidade de B.Existem algumas soluções apresentadas em [11] para reduzir este problema. Experimentos realizados por [18] chegaram à seguinte fórmula para computar a utilidade de recomendar um produto B dado que um produto A foi comprado anteriomente:

$$u(A,B) = \frac{Freq(AB)}{Freq(A) * Freq(B)^{\alpha}} \tag{4.2}$$

onde α é um valor entre 0 e 1. Quando $\alpha = 0$ a Equação 4.2 fica idêntica a Equação 4.1.

Como suspeita-se que os itens populares distorcem os resultados das avaliações com log, algumas das fórmulas que serão utilizadas neste estudo claramente beneficiam os populares e outras penalizam os populares. Conforme apresentado na Tabela 4.1, as fórmulas $F4$, $F5$ e $F6$ são idênticas à fórmula apresentada na Equação 4.2, com $\alpha = 0$, $\alpha = 1$, $\alpha = 0,5$, respectivamente. A fórmula $F3$ é $P(A/B)$ multiplicado por $log_2 P(B)$ e foi proposto em [10]. As fórmulas $F1$, $F7$ e $F8$ foram propostas como alternativas para complementar as possibilidades de geração de recomendações distintas, onde $F7$ nada mais é do que $P(A/B)$ multiplicado por $log_2 P(A)$, $F8$ é $P(B/A)$ multiplicado por $P(A/B)$ e por último a fórmula $F1$ é a fórmula $F3$ modificada, sendo igual a $F3$ multiplicada por $Freq(AB)$.

u(A,B)
$F1 = (Freq(AB)/Freq(B)) * (Freq(AB) * log_2(P(B))$
$F2 = Freq(AB)$
$F3 = (Freq(AB)/Freq(B)) * log_2 P(B)$
$F4 = Freq(AB)/(Freq(A) * Freq(B)^1)$
$F5 = Freq(AB)/(Freq(A) * Freq(B)^0)$
$F6 = Freq(AB)/(Freq(A) * Freq(B)^5)$
$F7 = (Freq(AB)/Freq(A)) * log_2(P(B))$
$F8 = (Freq(AB)/Freq(A)) * (Freq(AB)/Freq(B))$

Tab. 4.1: Cálculo da função de utilidade u(A,B)

Complexidade

A computação da matriz com a utilidade entre os produtos é $O(N^2M)$ no pior caso, onde N é a quantidade de produtos e M a quantidade de clientes. Na prática a complexidade é próxima de $O(NM)$, pois a grande maioria dos clientes compra

um número pequeno de produtos comparado com a quantidade disponível.

4.2 Log de transações

Teve-se acesso a um log de transações onde estão registrados 9 meses de compras realizadas por 29.664 clientes identificados. O critério para seleção dos clientes foi buscar somente os clientes que possuíam pelo menos uma compra no mês seguinte aos três primeiros meses iniciais. Para facilitar o entendimento da formação das bases de treino e teste, o log de transações foi dividido em três bases conforme descrito a seguir:

1. Base1: 3 meses de compras (abril, maio, junho)
2. Base2: 2 meses de compras (julho e agosto)
3. Base3: 6 meses de compras (julho, agosto, setembro, outubro, novembro e dezembro)

Na literatura encontram-se vários tipos de formações da base de teste. A ausência de um padrão torna difícil a comparação dos resultados dos experimentos. Neste estudo de caso utilizou-se inicialmente as seguintes formações de bases de treino e teste:

- M1: Base de teste composta pela primeira compra realizada por cada Cliente logo após o período de treino. A base de treino é a Base1 e a base de teste é o primeiro ticket de compra do cliente registrado na base2. Desta forma será medida a acurácia da recomentação dada uma compra (ticket) do Cliente.
- M2: Define-se um produto comprado por cada cliente para compor a base de teste. Para esta formação de base, foi retirado aleatoriamente um produto

do último ticket de compra do cliente da base 1 e estes produtos passaram a compor a base de teste.

- M3: Todas as compras realizadas durante 2 meses após o período utilizado na base de treino irão compor a base de teste. A base de treino é a Base1 e a base de teste é a Base2.

- M4: Todas as compras realizadas durante 6 meses após o período utilizado na base de treino irão compor a base de teste.A base de treino é a Base1 e a base de teste é a Base3.

Na Tabela 4.2 são apresentadas as características de cada base de teste, quantidade de clientes, número de tickets de compra, o número médio de tickets por cliente e a número médio de produtos distintos comprados por cliente.

	Formação da base de teste			
Características	M1	M2	M3	M4
Quant. de clientes	29.664	25.353	29.664	29.664
Número de Tickets	29.664	25.353	91.842	148.708
Média de Tickets	1	1	3,10	5,02
Média de Produtos	1,91	1	4,97	10,22

Tab. 4.2: Características das bases de teste obtidas com as formações M1, M2, M3 e M4

4.3 Geração das Recomendações

As recomendações foram geradas a partir da execução do algoritmo de recomendação sobre a base de treino. O resultado da execução é a lista top-N de itens que seriam de interesse do cliente, onde N é igual a 10. Foram geradas 8 listas top-N para cada cliente, sendo cada uma das listas geradas utilizando uma das funções de utilidade $u(A, B)$ descritas na Tabela 4.1.

Para definição de qual formação da base de treino e teste seria utilizada neste estudo de caso, calculou-se o HR sobre as 4 formações da base de teste indicadas na Tabela 4.2. Na Figura 4.3 pode ser visto o hit rate (HR) obtido em cada uma das funções de utilidade u(a,b) avaliadas utilizando as bases de teste M1, M2, M3 e M4. Analisando-se o HR obtido sobre as bases M1, M3 e M4, é fácil observar que quanto maior a base de teste, maior o hit rate, entretanto o número médio de tickets por cliente também aumenta.

	HR (Hit Rate)			
u(a,b)	M1	M2	M3	M4
F1	0,081	0,147	0,162	0,218
F2	0,124	0,156	0,237	0,331
F3	0,023	0,062	0,047	0,067
F4	0,044	0,094	0,086	0,121
F5	0,124	0,159	0,235	0,329
F6	0,094	0,154	0,179	0,244
F7	0,120	0,162	0,224	0,311
F8	0,088	0,157	0,172	0,233

Tab. 4.3: Avaliação baseada em log: Hit rate calculado em bases de teste diferentes.

As formações de base M1 e M2 parecem ser as mais indicadas por refletirem a acurácia do Sistema dada uma compra específica. Por este motivo, as avaliações utilizando estas duas possibilidades de formações de bases serão analisadas mais detalhadamente para então definir-se qual formação será utilizada neste estudo de caso.

Os resultados do hits e demais métricas utilizando a formação da base de teste M2 foran melhores, como pode ser visto na Tabela 4.4, por fazer da base de teste um único produto por cliente, que compõe um ticket de compra que possui outros produtos comprados juntos e que fazem parte da base de treino. Supondo-se que um cliente possui um único ticket de compras com dois produtos, uma cama e um

colchão, o processo normal de um Sistema de Recomendação seria a partir deste histórico tentar predizer quais produtos seriam de interesse deste cliente. Portanto, dividir um ticket de compra em treino e teste certamente aumenta a acurácia pois em inúmeras situaçoes a base de treino já possui outros produtos fortemente relacionados, entretanto não segue o processo normal de recomendação.

Optou-se por utilizar a base M1 por parece ser a mais adequada por três motivos: primeiro, reflete a qualidade da recomendação na primeira compra (ticket) realizada pelo cliente logo após a geração da recomendação; segundo, com este tipo de formação a avaliação com pessoas é feita simulando-se exatamente um sistema onde as recomendações são apresentadas ao cliente considerando todo o seu histórico de compras; e por último, como consequência do segundo motivo, torna possível comparar o resultado da avaliação baseada em logs e avaliação com pessoas.

	Hits		**HR**		**ARHR**		**Cob.Cliente**	
u(A,B)	M1	M2	M1	M2	M1	M2	M1	M2
F1	2429	**3742**	0,081	0,147	0,0385	0,0824	29664	25353
F2	3692	**3963**	0,124	0,156	0,0605	0,0882	29664	25353
F3	697	**1580**	0,023	0,062	0,0083	0,0288	29658	25348
F4	1304	**2395**	0,044	0,094	0,0146	0,0422	29658	25348
F5	3707	**4042**	0,124	0,159	0,0598	0,0919	29663	25352
F6	2809	**3916**	0,094	0,154	0,0411	0,0854	29653	25342
F7	3568	**4115**	0,120	0,162	0,0568	0,0933	29661	25350
F8	2623	**3999**	0,088	0,157	0,0410	0,0890	29661	25350

Tab. 4.4: Resultado do Hits, HR, ARHR e Cobertura de clientes utilizando as formações de base de teste M1 e M2

4.4 Avaliações do Sistema de Recomendação

Após geradas as recomendações foram realizadas 2 avaliações de qualidade das recomendações: a primeira, baseada em logs e a segunda, com pessoas. O resultado

de cada avaliação é uma lista contendo as funções de utilidade em ordem decrescente do resultado da métrica ARHR.

4.4.1 Avaliação baseada em logs

Na avaliação baseada em logs utilizou-se exclusivamente a base de teste. Utilizando-se a base de teste M1, para cada função de utilidade foram computados o hits, HR, ARHR e cobertura de clientes com recomendações, os resultados podem ser vistos na Tabela 4.4. O próximo passo foi gerar um ranking a partir de uma determinada métrica, no caso, a métrica ARHR. Seja r_1 o ranking da qualidade das funções que geraram as recomendações na avaliação com log utilizando a métrica ARHR, conforme resultado apresentado na Figura 4.1 tem-se $r_1 = F2, F5, F7, F6, F8, F1, F4, F3$.

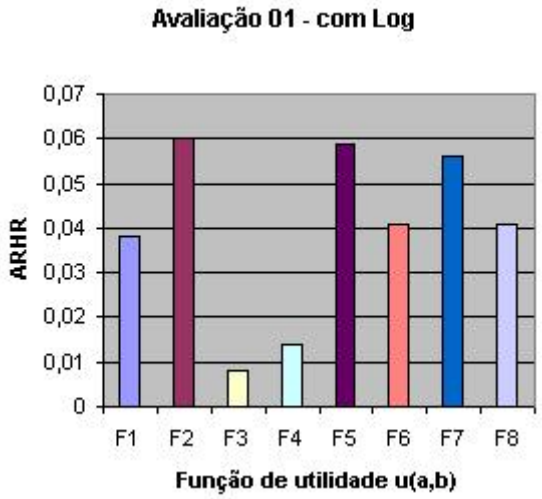

Fig. 4.1: Avaliação baseada em Log: ARHR - formação da base de teste M1

4.4.2 Avaliação com pessoas

Vinte pessoas foram convidades a participar do experimento, sendo 15 pessoas do grupo GTI da UFAM, dentre eles alunos de graduação e mestrado, e 5 pessoas com conhecimento em varejo. Foram selecionados aleatoriamente 120 clientes da base

de treino e as suas 8 respectivas top-10 recomendações. Cada pessoa teve acesso ao histórico de compras e as recomendações geradas para 12 clientes diferentes. As recomendações foram apresentadas em ordem crescente do código do item e sendo a união das oito top-10 recomendações geradas para cada cliente.

Para o domínio do estudo de caso, que é para uma empresa de venda de produtos no varejo, as recomendações podem ser apresentadas de forma diferente conforme o objetivo da avaliação. Dois exemplos seriam:

1. Existência de muitos algoritmos de recomendação e necessidade de identificar os melhores. Situação típica quando se está implantado um sistema de recomendação e uma avaliação inicial será realizada. Neste caso, o ideal é que as recomendações sejam apresentadas misturadas para que o avaliador não saiba qual técnica está avaliando e concentre-se nas recomendações, não em comparar qual técnica é melhor ou pior.

2. Dois ou três algoritmos com resultados muito próximos. Quando se está em uma fase de refinamento do sistema de recomendação, o total conhecimento dos avaliadores quanto a qual algoritmo está sendo avaliado deverá ajudar na obtenção de avaliações onde as pessoas já apontam, através de uma comparação explícita dos resultados, qual a melhor técnica na percepção das pessoas.

Neste estudo de caso, as recomendações forma apresentadas misturadas. Cada produto recomendado foi avaliado indicando-se o grau de relevância, se muito relevante (nota 3), relevante (nota 2), pouco relevante (nota 1) ou não relevante (nota 0). Cada cliente foi avaliado por 3 pessoas, sendo considerada a avaliação da terceira pessoa somente na ocorrência de divergência entre o primeiro e o segundo avaliador. Para calcular o hits,HR e ARHR foram consideradas as notas 3 e 2 como relevantes e as notas 1 e 0 como não relevantes, o resultado pode ser visto na Tabela 4.5.

Seja r_2 o ranking da qualidade das funções que geraram as recomendações na avaliação com pessoas, utilizando a métrica ARHR, conforme resultado apresentado na Figura 4.2, tem-se:

$r_2 = F8, F4, F6, F1, F3, F7, F5, F2$

u(A,B)	Hits	HR	ARHR
F1	117	0,975	2,0248
F2	114	0,950	1,3992
F3	117	0,975	1,8573
F4	119	0,991	2,0934
F5	118	0,983	1,5004
F6	119	0,991	2,1131
F7	119	0,991	1,7760
F8	119	0,991	2,1450

Tab. 4.5: Resultado da avaliação com Pessoas: Hits, HR e ARHR

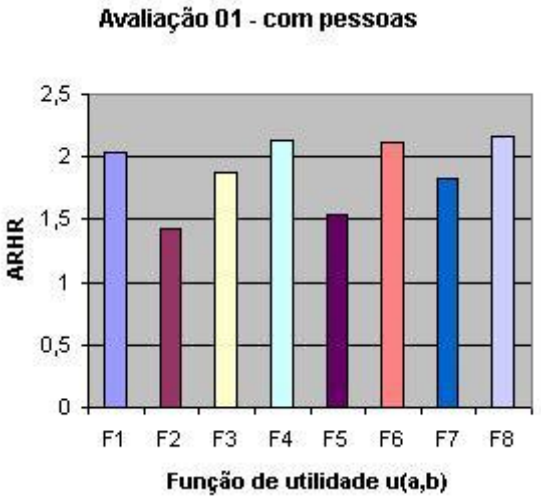

Fig. 4.2: ARHR - Avaliação com Pessoas

4.5 Identificação de Ruídos e filtragem do log

Conforme pode ser visto na Tabela 4.6, o resultado da avaliação baseada em log (r_1) e o resultado da avaliação com pessoas (r_2) são praticamente opostos, as recomendações geradas com as funções de utilidade F2, F5 e F7 classificadas na avaliação baseada em log como primeira, segunda e terceira melhores funções foram a oitava, sétima e sexta na avaliação com pessoas.

Neste caso a avaliação baseada em log não foi suficiente para indicar de forma correta qual a função de utilidade que gerou de fato as melhores recomendações.

Lista	1	2	3	4	5	6	7	8
r_1 (Aval. baseada em log)	F2	F5	F7	F6	F8	F1	F4	F3
r_2 (Aval. com pessoas)	F8	F6	F4	F1	F3	F7	F5	F2

Tab. 4.6: Resultado das avaliações baseadas em logs e pessoas

4.5.1 Filtro dos populares do top-N

Através da observação das recomendações não-personalizadas geradas pelo algoritmo e das avaliações com pessoas, identificou-se que os itens populares diminuem a qualidade da recomendação. Este problema ficou mais evidente por tratar-se de um log com produtos diversificados, onde existem transações de vendas de produtos de diversas categorias diferentes.

Para confirmar se realmente os produtos populares eram os geradores do falso positivo, realizou-se um novo experimento filtrando do top-N os produtos populares. Este experimento foi conduzido de forma similar ao experimento baseado em log. Como resultado obteve-se a lista $r_3 = F8, F6, F1, F4, F7, F5, F3, F2$, que é o ranking da qualidade das funções que geraram as recomendações na avaliação com log excluindo os populares do top-N, conforme valores da métrica ARHR apresentados na Figura 4.3.

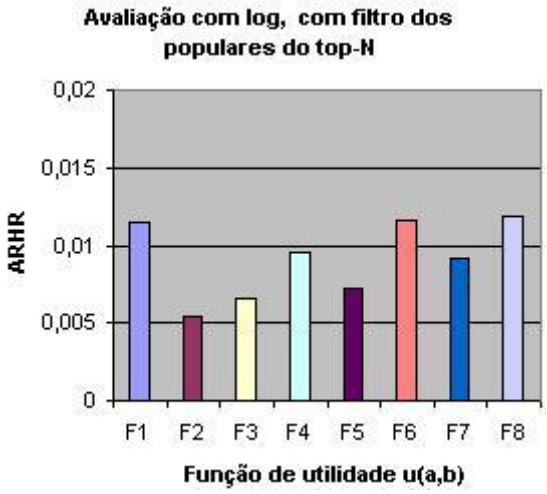

Fig. 4.3: Resultado da avaliação baseada em log: ARHR - filtro dos populares do top-N

Lista	1	2	3	4	5	6	7	8
r_2 (Aval. com pessoas)	F8	F6	F4	F1	F3	F7	F5	F2
r_3 (Aval. com filtro dos populares do top-N)	F8	F6	F1	F4	F7	F5	F3	F2

Tab. 4.7: Identifiçao dos ruídos

4.5.2 Filtro dos populares da base de treino

Para confirmarmos que um tratamento no log reduziria a distância entre as avaliações baseada em log e avaliações com pessoas, realizamos outro experimento, agora retirando os itens populares antes da geração das recomendações. As avaliações foram realizadas baseada em log e com pessoas, como já existe uma base das avaliações com pessoas, neste segundo experimento as pessoas avaliaram somente as recomendações que não tinham sido avaliadas anteriormente, sendo assim reduzido o esforço de execução da avaliação. Nas Tabelas 4.8 e 4.9 são apresentamos os resultados dos valores de Hits, HR e ARHR obtidos nas avaliações baseadas em logs e com pessoas respectivamente. Nas Figuras 4.4 e 4.5 é possível visualizar mais claramente o ranking das recomendações utilizando as diversas funções de utilidade. Desta forma tem-se como resultado as listas, r_4 que é o ranking da avaliação baseada

em log e r_5 qué o ranking da avaliação com pessoas.

u(A,B)	Hits	HR	ARHR	Cobertura
F1	796	0,0323	0,0143	24607
F2	957	0,0388	0,0174	24607
F3	627	0,0254	0,0106	24602
F4	700	0,0284	0,0111	24602
F5	943	0,0383	0,0170	24605
F6	844	0,0343	0,0142	24597
F7	920	0,0373	0,0166	24603
F8	831	0,0337	0,0146	24605

Tab. 4.8: Avaliação Baseada em Log:Hits, HR, ARHR e Cobertura de clientes - Base Treino com Filtro dos Populares

u(A,B)	Hits	HR	ARHR
F1	107	0,8916	1,8752
F2	107	0,8916	1,8616
F3	107	0,8916	1,8298
F4	107	0,8916	1,8599
F5	107	0,8916	1,8818
F6	107	0,8916	1,9087
F7	107	0,8916	1,9253
F8	107	0,8916	1,9125

Tab. 4.9: Avaliação com pessoas:Hits, HR e ARHR - Base Treino com Filtro dos Populares

4.6 Comparação das avaliações

Para medir a distância entre as avaliações utilizou-se a métrica proposta por Kendall. Conforme detalhado no Capítulo 3, esta métrica conta o número de pares divergentes nas duas listas e quanto maior a distância menor é a similaridade entre as listas. Na Tabela 4.11 é apresentado o cálculo da distância de Kendall entre as

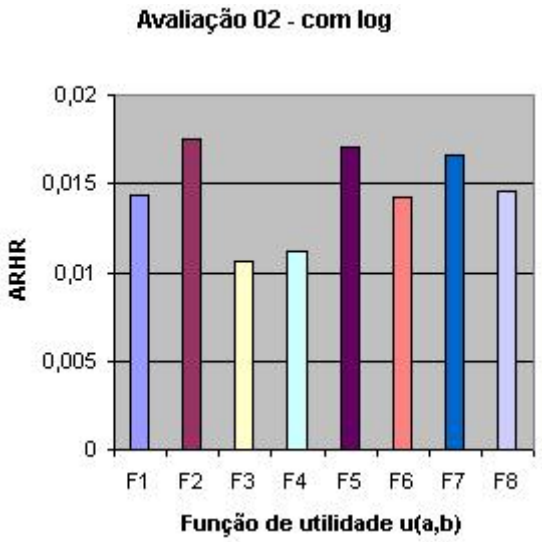

Fig. 4.4: Resultado da avaliação baseada em log: ARHR - base treino com filtro dos populares

Lista	1	2	3	4	5	6	7	8
r_4 (Aval. baseada em log)	F2	F5	F7	F8	F1	F6	F4	F3
r_5 (Aval. com pessoas)	F7	F8	F6	F5	F1	F2	F4	F3

Tab. 4.10: Filtro dos populares

seguintes listas: r_1 e r_2, que representam o resultado da avaliação baseada em log e avaliação com pessoas; r_3 e r_2, que são os resultados das avaliações com filtro dos populares do top-N e avaliações com pessoas; e por fim r_4 e r_5, que foram as avaliações baseadas em log e com pessoas cujas recomendações foram geradas sobre o log ja' com o filtro dos populares.

A maior distância é entre as listas r_1 e r_2. É a distância mais próxima de 1, indicando listas na ordem praticamente oposta. Esta distância de $0,7142$ é uma forte evidência de que realizar avaliações somente com logs pode levar a resultados errôneos, como os obtidos neste estudo de caso. Após avaliar o log, houve a suspeita de que os produtos populares eram os principais responsaveis pela distância entre as avaliações. Então, filtrou-se os populares do top-N e fez-se uma nova avaliação baseada em log com o objetivo de identificar se realmente estes produtos estavam

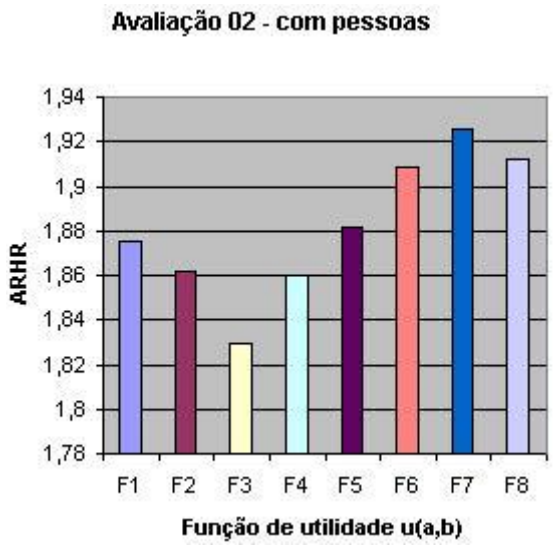

Fig. 4.5: Resultado da avaliação com pessoas: ARHR - base treino com filtro dos populares

gerando o falso positivo. Calculou-se a distância entre esta última avaliação realizada com log e a avaliação originalmente feita com pessoas e obteve-se listas bastante similares com distância de Kendall igual a $0,1078$. Desta forma ficou claro não somente o quão distante da avaliação com pessoas está a avaliação com log, mas também que a recomendação de produtos populares, na opinião das pessoas, não é relevante. Por fim, as recomendações foram geradas após tratamento do log e a distância entre os resultados da avaliação baseada em log e com pessoas de fato reduziu consideravelmente de $0,7142$ para $0,1875$.

4.7 Alguns Resultados da Aplicação

Nesta sessão apresentaremos alguns resultados da implementação da aplicação e que podem ser fonte de estudo em futuros trabalhos.

Pares	(r_1, r_2)	(r_3, r_2)	(r_4, r_5)
1,2	1		1
1,3			
1,4	1	1	
1,5	1		
1,6			1
1,7	1		
1,8			
2,3	1		
2,4	1		
2,5	1		1
2,6	1		1
2,7	1		1
2,8	1		1
3,4			
3,5	1	1	
3,6			
3,7	1	1	
3,8			
4,5	1		
4,6			
4,7	1		
4,8			
5,6	1		1
5,7	1		1
5,8	1		1
6,7	1		
6,8	1		
7,8	1		
Kendall	0,7142	0,1071	0,1875

Tab. 4.11: Cálculo da distância de Kendall

4.7.1 Recomendação Não-Personalizada

Através da implementação do Sistema de Recomendação Não-Personalizado observamos que o algoritmo consegue identificar:

- Itens com características similares. Características como marca (ver Figura 4.7.1), cor (ver Figura 4.7), dimensões (ver Figura 4.8) e compatibilidade (ver Figura 4.9) são relacionadas pelo próprio histórico de compras.

- Itens complementares. Com um grande volume de histórico de transações, o algoritmo mostra resultados muito bons de vendas adicionais não personalizadas.

- Itens da mesma linha de preço. O histórico também evidência o comportamento de Clientes que compram produtos da linha baixa, média ou alta e recomenda novos produtos na mesma linha de preço.

4.7.2 Produtos não-similares

Uma das características do log de transações utilizado neste estudo de caso é a diversidade. Os experimentos encontrados na literatura são em sua grande maioria realizados com grupos de produtos específicos como CDs, DVDs e Livros. Pode-se conceituar produtos similares como produtos que possuem as mesmas características básicas de funcionamento e a diferença entre eles está em atributos visuais (ex.cor, dimensão,etc), performance ou funcionalidades adicionais. Como exemplo destes tipos de produtos temos, geladeira, liquidificador, TV e muitos outros. Produtos não similares são produtos únicos, sendo necessário especificar o produto para identificar as suas características básicas (ex. Livro, CD, DVD,etc).

Com base nesta classificação dos produtos foi possível observar algumas relações entre a acurácia (avaliação baseada em log) e a satisfação (avaliação com pessoas).

Fig. 4.6: Recomendação não-personalizada: Exemplo - Marca

Fig. 4.7: Recomendação não-personalizada: Exemplo - Cor

Fig. 4.8: Recomendação não-personalizada: Exemplo - Dimensão

Fig. 4.9: Recomendação não-personalizada: Exemplo - Compatibilidade

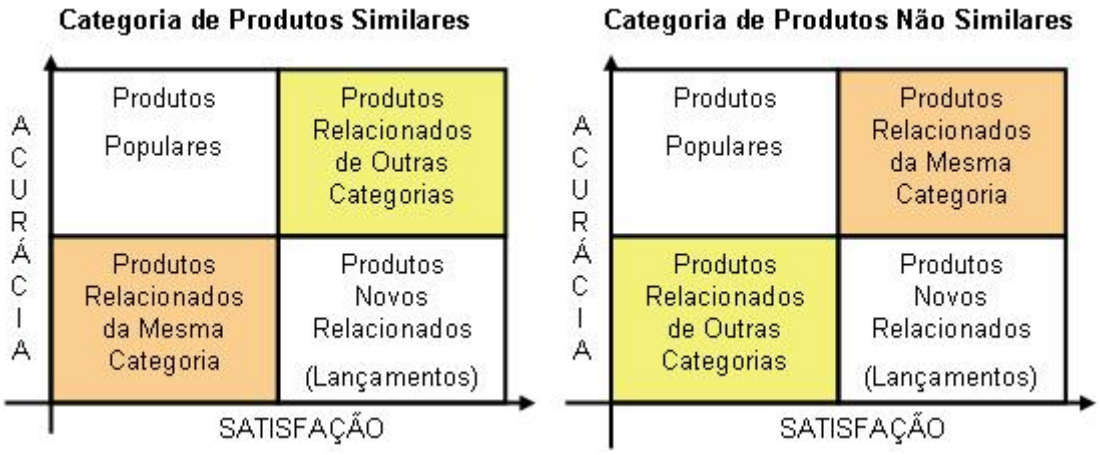

Fig. 4.10: Gráfico de acurácia X satisfação por categoria de produtos recomendados

Cada quadrante apresentado na Figura 4.10 é detalhado a seguir:

Alta Acurária e Alta Satisfação: Em produtos classificados como similares, quando recomenda-se produtos relacionados ao histórico de compras do cliente e pertencentes a outra categoria (ex. o cliente comprou um criado mudo e o sistema recomenda um abajour), em geral a acurária e satisfação coincidem, sendo ambas mais elevadas. Já quando tratam-se de produtos não-similares, a acurária e a satisfação são elevadas quando recomenda-se produtos da mesma categoria (ex. se o cliente está consultando um CD ou comprou um CD anteriormete, ele espera que um outro CD de seu interesse seja recomendado e não um outro tipo de item.

Alta Acurácia e Baixa Satisfação: A recomendação de produtos populares certamente produzem altas taxas de acurácia, mas possuem baixo nível de satisfação e pode comprometer fortemente a credibilidade do sistema de recomendação.

Baixa Acurária e Alta Satisfação: Certamente, outra forma de alcançar um nível elevado de satisfação do cliente é recomendando um produto novo relacionado aos produtos comprados anteriormente pelo cliente. Este tipo de recomendação é possível utilizando outras técnicas de recomendação, como a técnica baseada em conteúdo onde as características do produto podem ser exploradas para geração de recomendação de novos produtos.

Baixa Acurária e Baixa Satisfação: Em produtos classificados como similares, a recomendação de produtos pertencentes a mesma categoria dos produtos comprados anteriormente, para muitos categorias existentes, possui baixa acurácia e não é considerada uma recomendação útil pelas pessoas. (ex. recomendar um novo modelo de lavadoura de roupa a um cliente que acabou de comprar este produto). O inverso ocorre com itens não-similares onde a baixa Acurária e baixa Satisfação ocorrem quando são recomendados itens de outras categorias.

Capítulo 5

Conclusões e Trabalhos Futuros

Este trabalho realizou estudos sobre como os Sistemas de Recomendação são avaliados. Identificou-se que a grande maioria dos experimentos encontrados na literatura foram avaliados exclusivamente com base em logs, principalmente pela praticidade de avaliar múltiplas técnicas simultaneamente [22, 18, 6]. Também verificou-se que existem estudos questionando se este tipo de avaliação é ou não confiável, considerando que as avaliações com pessoas conseguem captar melhor a reação à recomendação e que um sistema com alta acurácia pode em algumas situações levar a resultados errôneos.

Com base nesta pesquisa, a proposta desta dissertação foi apresentar uma forma de validar as avaliações baseadas em logs tornando o uso do log confiável em avaliações de sistemas de recomendação colaborativos. Foi realizado um estudo de caso partindo do princípio que algumas entradas do log de transações poderiam ocasionar uma distância considerável entre a simples verificação da acurácia de predição do sistema de recomendação e a real percepção de qualidade das recomendações identificadas por pessoas.

Neste estudo de caso verificou-se uma grande disparidade entre os resultados das avaliações baseadas em log e avaliações com pessoas. Tal disparidade ocorreu

porque a recomendação de produtos populares, que o cliente compraria independente de haver uma recomendação, produziu bons resultados na avaliação baseada em logs e o mesmo não ocorreu na avaliação com pessoas, onde os resultados das avaliações foram ruins. Esta evidência confirma a suspeita levantada por alguns autores [13, 23, 15] de que sistemas de recomendação com alta acurácia, conforme as métricas existentes, podem não gerar as recomendações mais úteis para os clientes

Pode-se concluir deste trabalho que não é suficiente realizar avaliações de Sistemas de Recomendação somente baseadas em logs de transações. Conclui-se também que é possível realizar avaliações baseadas em log obtendo resultados próximos das avaliações realizadas com pessoas desde que sejam identificados e filtrados ruídos do log que possam ser os principais responsáveis pelas divergências entre as avaliações.

5.1 Trabalhos Futuros

Nesta sessão são apresentadas sugestões de trabalhos futuros que poderão dar continuidade aos experimentos realizados nesta dissertação.

Sugere-se o estudo de uma métrica que penalize a recomendação de produtos populares, ocorrência comum em logs de transações de empresas varejistas, não havendo com isso a necessidade de filtragem do log. Seria interessante também estudos mais aprofundados de avaliações de recomendação não-personalizada a partir da consulta de um produto categorizado como similar e não-similar.

Referências Bibliográficas

[1] Brendan Kitts, David Freed, Martin Vrieze. *Cross-sell: A Fast Promotion-Tunable Customer-Item Recommendation Method Based on Conditionally Independet Probabilities.* KDD, 2000, Boston, USA.

[2] Cai-Nicolas Ziegler, Lars schmidt-Thieme, and Gerorge Lausen. Exploiting Semantic Product Descriptions for Recommender Systems. In *Proceedings of the 2nd ACM SIGIR Semantic Web and Information Retrieval Workshop*, Sheffield, UK, July 2004.

[3] Cai-Nicolas Ziegler, Gerorge Lausen, and Lars schmidt-Thieme. Taxonomy-driven Computation of Product Recommendations. In *Proceedings of the 2004 ACM CIKM - Conference on Information and Knowledge Management*, Washington, D.C., USA, November 2004.

[4] Cai-Nicolas Ziegler, Sean M. McNee, Joseph A. Konstan, and Gerorge Lausen. Improving Reommendation Lists Through Topic Diversification. In *Proceedings of WWW 2005*, Chiba, Japan.

[5] Cosley D., Lam, S.K., Albert, I.,Konstan, J.A., and Riedl,J. Is Seeing Believing? How Recommender Interfaces Affect User's Opinion. In *CHI*, Florida, USA, 2003.

[6] Eui-Hong Han and George Karypis. *Feature-Based Recommendation System.* CIKM, november 2005.

[7] G. Adomavicius and A. Tuzhilin. Multidimensional Recommender System: a data werehousing approach. In *Proceedings of the Second International Workshop on Eletronic Commerce.* Lecture Notes in Computer Science, vol 2232, Springer 2001.

[8] G. Adomavicius and A. Tuzhilin. *Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions.* IEEE Transactions on Knowledge and Data Engineering, vol.17, NO.6, June 2005.

[9] G. Adomavicius, R. Sankaranarayanan, S. Sen, and A. Tuzhilin. Incorporating Contextual Information in Recommender System Using a Multidimensional Approach. In *ACM Transactions on Information Systems (TOIS), Volume 23 Issue 1* , january 2005.

[10] G. Salton. Automatic Text Processing: The Transformation Analysis and Retrieval of Information by Computer. Addison-Wesley, 1989.

[11] George Karypis. *Evaluation of Item-Based Top-N Recommendation Algorithms.* CIKM, 2001.

[12] Goldberg, D., Nichols, D., Oki, B. M., and Terry, D. *Using Collaborative Filtering to Weave an Information Tapestry.* Communications of ACM, 1992.

[13] Hayes, C., Massa, P., Avesani, P., and Cunningham, P. *An on-line evaluation framework for recommender systems.* In Workshop on Recommendation and Personalization Systems, Springer Verlag, 2002. Lecture Notes in Computer Science, vol 2232, Springer 2001.

[14] J. B. Schafer, J.A.Konstan, and J. Riedl. E-commerce Recommendation Applications. *Data Mining and Knowledge Discovery*, vol.5, pages 115–153,2001.

[15] Jonathan L. Herlocker and Joseph A. Konstan, Loren G. Terveen, and John T. Riedl. *Evaluating Collaborative Filtering Recommender Systems.* In ACM Transactions on Information Systems (TOIS), Volume 22 Issue 1, pages 5-53, january 2004.

[16] Kendall, M. *Rank Correlation Methods.* Charles Griffin and Company Limited, (1948)

[17] Linden, G., Smith, B., and York, J. *Amazon.com Recommendations: Item-to-Item Collaborative Filtering.* IEEE Internet Computing, january 2003.

[18] Mukund Deshpande and George Karypis. *Item-Based Top-N Recommendation Algorithms.* TOIS, january 2004.

[19] Nichols, D. Impliciting Rating and Filtering. In *Proceedings of the Fifth DELOS Workshop on Filtering and Collaborative Filtering*, page 31–36. Budabeste, 1998.

[20] Rashmi R. Sinha and Kirsten Swearingen. *Comparing Recommendations Made by Online Systems and Friends.* DELOS Workshop: Personalisation and Recommender Systems in Digital Libraries 2001.

[21] Sarwar, B., Karypis,G., Konstan J., and Riedl, J. Analysis of Recommendation Algorithms for e-commerce. In *Proceddings of the 2nd ACM Conference on Electronic Commerce*, pages 158–167, Minneapolis, USA, 2000.

[22] Sarwar, B., Karypis,G., Konstan J., and Riedl, J. Item-Based Collaborative Filtering Recommendation algorithms. In *WWW10, may, Hong Kong* 2001.

[23] Sean M. Mcnee, J.A.Konstan, and J. Riedl. Being Accurate in Not Enough. How Accuracy Metrics have hurt Recommeder Systems. *CHI'06 - Conference on Human Factors in Computing Systems* , pages 1097 - 1101, Montréal, Québec, Canada, 2006.

[24] Z. Huang, W. Chung, and H. Chen. A Graph Model for E-Commerce Recommender System. *Journal of the American Society for Information Science and Technology*, pages 259-274, 2004.