DAS Departamento de Automação e Sistemas
CTC Centro Tecnológico
UFSC Universidade Federal de Santa Catarina

# Framework de Avaliação Off-line para Sistemas de Recomendação

*Relatório submetido à Universidade Federal de Santa Catarina como requisito para a aprovação da disciplina:*

***DAS 5501: Estágio em Controle e Automação Industrial***

***Thiago Belluf Inoue***

*Florianópolis, Agosto de 2012*

# Framework de Avaliação Off-line para Sistemas de Recomendação

***Thiago Belluf Inoue***

**Orientadores:**

***Leopoldo da Silva Xavier / Engenheiro de Controle e Automação***

_____

Assinatura do Orientador

**Prof. *José Ricardo Rabelo***

_____

Assinatura do Orientador

Este relatório foi julgado no contexto da disciplina

**DAS 5501: Estágio e Controle e Automação Industrial**

e aprovado na sua forma final pelo

**Curso de Engenharia de Controle e Automação**

## Resumo

Com o avanço e popularização da internet, dos computadores pessoais e de dispositivos móveis, a quantidade de informação acessada diariamente aumentou em proporções consideráveis no decorrer dos últimos anos. Essas diferenças impactaram diversos setores e transformou a rotina das pessoas, desde a forma como as interações sociais são construídas até o comportamento de consumo. A consequência final de toda esta integração, virtualização e alta disponibilização é que a quantidade de informações apresentadas para as pessoas aumentou de maneira que existe grande dificuldade em filtrar e obter o que é de fato relevante.

Os Sistemas de Recomendação são ferramentas e técnicas de software desenvolvidas com o objetivo de auxiliar a navegação dos usuários frente a toda essa informação, através da promoção das entidades que apresentem o maior potencial de relevância e utilidade dentro do contexto de navegação. Estes sistemas hoje atuam em diferentes segmentos, desde sugestão de produtos em ambientes de comércio eletrônico, sugestão de músicas e vídeos no YouTube até recomendações de possíveis amigos em redes sociais como o Facebook[1].

A aplicação nesses diferentes segmentos é possível graças à grande variedade de abordagens conhecidas, o que possibilita a criação dos mais diversos tipos de sistemas, que se aplicam aos mais diversos e específicos casos. As preocupações com o desenvolvimento e adequação destes sistemas motivou um dos sub-ramos das pesquisas no tema, relacionado à avaliação da das características, qualidades e resultados destes sistemas. Esse processo de avaliação é um grande desafio, pois os sistemas estudados são altamente dinâmicos e dependentes de contexto, o que torna difícil a tarefa de definir o que é bom e suficiente, definições bases para o processo de avaliação.

Este projeto visa avançar as pesquisas em avaliação de Sistemas de Recomendação dentro da Chaordic, provendo tanto conhecimento quanto ferramentas para auxiliar o processo.

[1] YouTube (www.youtube.com) / Facebook (www.facebook.com)

## Abstract

Due to the rapid advance and popularization of the internet, personal computers and mobile devices, the amount of information available and daily accessed increased significantly for the past years. These differences had major impact on many industry segments and transformed people's routines, from the way social interactions are built to their consumption behavior. The ultimate consequence of this integration, virtualization and high availability effects is the increased difficulty in filtering and obtaining what really is relevant information.

Recommender Systems are software tools and techniques developed in order to help user navigation through all these information available, by promoting items that presents the highest potential of relevance and utility within a given context. These systems act today with significantly different purposes, ranging from products suggestions on e-commerce platforms, video and music suggestions in media platforms like Youtube up to friends suggestions in social networks like Facebook.

This diversity of Recommender Systems is only made possible by the variety of available approaches for generating recommendations, which allows their use in a wide range of very specific domains. Worries with development and adequacy motivated the expansion of some research sub-areas related to the evaluation of these systems, and by evaluation is meant the processes of describing characteristics, quality and results. These processes are now a great challenge mainly because recommender systems are highly dynamic and context dependent, turning difficult the task of building the definitions of what is good and where to get, which are the baselines for the evaluation process.

This project aims to contribute with the advance of researches related to evaluating Recommender Systems within the Chaordic Systems, providing both theory expansion and tools to support the evaluation processes.

## Sumário

## Índice de Ilustrações

# 1. Introdução

## 1.1. Contextualização na Empresa e Motivação

A Chaordic Systems é uma empresa no ramo de tecnologia que provê serviços para sites de comércio eletrônico de médio e grande porte do mercado brasileiro. O serviço básico da empresa consiste em uma plataforma de fornecimento de recomendações personalizadas.

De maneira geral, Sistemas de Recomendação são sistemas computacionais usados para reconhecimento de padrões. Mais especificamente, usam informações sobre usuários, itens e interações entre eles, para fazer previsões sobre possíveis itens capazes de agradar um usuário em um determinando momento de sua experiência.

Graças à rápida evolução e popularização da internet estes sistemas vêm se tornando muito expressivos e sendo aplicados nos mais diversos contextos. Um de seus objetivos é auxiliar os usuários de um sistema a navegar em meio à grande e crescente quantidade de informações existentes.

Atualmente existem várias estratégias para se trabalhar com recomendações automáticas e personalizadas, entre elas podemos citar as abordagens baseadas nos conteúdos dos itens, em perfis de usuário, e, informações demográficas ou em informações de interações de usuários. Em meio a todas essas possibilidades de ferramentas e conceitos, um grande campo de pesquisa vem se desenvolvendo, tanto no meio acadêmico quanto na indústria.

Essa diversidade de sistemas de recomendação com diferentes características e potenciais torna complexo o processo de escolha do mais adequado. Os critérios que definem qualidade, além de serem ligados aos objetivos desejados e ao contexto aplicado, também dependem de decisões arbitrárias. Métodos para sistematizar essas escolhas ainda são escassos e estão sendo desenvolvidos.

Com o objetivo de auxiliar na caracterização e escolha dentre diversos sistemas de recomendação e suas possíveis variações está a área da pesquisa voltado para métodos de avaliação.

Atualmente a equipe de pesquisa e desenvolvimento da Chaordic possui maneiras de melhorar a qualidade e impacto do produto, porém o processo de avaliação é pouco formal e consequentemente peca em determinar os impactos das mudanças feitas em seus produtos.

As propostas deste projeto são a realização de uma pesquisa sobre propriedades e métodos de avaliação de sistemas de recomendação, a contextualização dos resultados da pesquisa aos produtos e propósitos da empresa e, por fim, a construção do protótipo de um framework de avaliação que auxilie a equipe de pesquisa da empresa no processo de experimentação.

Visto que os processos de avaliação hoje existentes se baseiam em avaliações quase exclusivamente qualitativas, conseguem avaliar apenas os comportamentos dos sistemas em casos muito específicos. Por essa falta de abrangência as generalizações feitas apresentam grandes riscos de não captar o comportamento geral dos sistemas, e consequentemente escolhas são feitas a partir de premissas potencialmente falhas.

Espera-se com o proposto sistema de avaliação alavancar as atividades de pesquisa, dando suporte para um entendimento não só qualitativo, mas principalmente quantitativo às iniciativas de inovação.

Outra motivação para a realização deste projeto é prover condições de, a partir de um processo de realimentação de informações, desenvolver e maximizar os impactos dos produtos já existentes melhorando o que é chamado internamente de processo de tunagem. O processo de tunagem consiste na escolha dos valores de parâmetros e possibilidades de filtros a serem usados nos algoritmos nos diversos contextos. Um dos passos almejados é a evolução do chamado processo de *tunagem*, que consiste na escolha dos conjuntos de parâmetros usados nos algoritmos de cálculo de similaridades usados pelo sistema de recomendação.

## 1.2. Objetivos

O processo de avaliação de sistemas de recomendação proposto por *Ricci et al.* em “*Recommender Systems Handbook*” é entendido como sendo constituído de três abordagens: avaliações off-line, estudos com usuários e avaliações on-line. Cada etapa é vista como complementar sobre a outra e

juntas compõem um processo de avaliação valioso para o desenvolvimento da pesquisa em Sistemas de Recomendação [11].

Os objetivos deste trabalho serão focados em processos de avaliação off-line, usando indicadores de desempenho para auxiliar no entendimento das características e comportamentos expressados pelos sistemas avaliados.

A seção 1.2.1 descreve os objetivos gerais do projeto, sumarizando o que se espera como valor final agregado do projeto. A seção 1.2.2 discute os objetivos específicos, explicitando as etapas do projeto e suas respectivas entregas.

### 1.2.1. Objetivos Gerais

Tendo em vista o problema apresentado na seção 1.1, e considerando o processo de avaliação proposto, é preciso avançar do ponto de vista teórico na complementação do conhecimento sobre como avaliar corretamente sistemas de recomendação, relacionando o processo de avaliação com as expectativas esperadas para os produtos da empresa. Precisamos entender quais critérios são importantes e aplicáveis ao contexto, bem como a maneira com que podemos mensurar a adequação dos nossos sistemas dentro desses critérios.

Do ponto de vista prático, é preciso criar processos e ferramentas que facilitem análises e que tornem viáveis suas realizações durante a rotina de trabalho, apresentando principalmente agilidade e flexibilidade, uma vez que o processo de mensurar é essencial para os processos de avaliação, pesquisa e desenvolvimento.

### 1.2.2. Objetivos Específicos

Nesta seção são apresentadas as etapas definidas para o projeto, juntamente com uma breve descrição e com a definição da respectiva entrega planejada. Como objetivos específicos para o projeto foram levantados os seguintes pontos:

1. Fazer um estudo sobre propriedades de sistemas de recomendação e métodos de avaliação off-line baseados em indicadores de desempenho, com

ênfase na tarefa de relacionar e priorizar quais dimensões e indicadores relacionados são aplicáveis ao contexto da Chaordic;

2. Criar uma descrição do principal algoritmo usado na geração de recomendações, enfatizando quais parâmetros são usados por ele e quais propriedades são relevantes para sua avaliação;

3. Definir uma lista de indicadores, descrevendo para cada um deles a propriedade de sistema relacionada e o significado da medida. Esta relação será usada para decidir quais indicadores serão no protótipo do sistema;

4. Definir um documento de especificação funcional e uma proposta de arquitetura para o protótipo do framework de avaliação;

5. Concretizar o protótipo do framework, que servirá para auxiliar na realização de experimentos e estudos de caso para validar a utilidade da teoria estudada;

6. Planejar e realizar experimentos, usando o protótipo feito, e validar se as propriedades e indicadores usados podem de fato auxiliar no processo de escolha, e se o processo de avaliação apresenta a agilidade e praticidade pretendida.

Os objetivos descritos aqui são apresentados e discutidos nas seções de 4 a 7, com exceção do objetivo 2, que é discutido na seção 2.3.1, por melhor se adequar.

## 1.3. Metodologia

Esta seção apresenta a metodologia usada para o desenvolvimento do trabalho. A metodologia seguida para o desenvolvimento do projeto foi:

1. Pesquisa e documentação. Com foco no tema de avaliação de Sistemas de Recomendação, com o intuito de expandir os conhecimentos existentes na Chaordic e explorar o estado da arte nas pesquisas deste ramo;

2. Desenvolvimento de uma estrutura para suportar a execução de experimentos. Com essa estrutura, realizar experimentos para verificar a teoria. Levantar o que é adequado e o que não é ao contexto do projeto e verificar possíveis lacunas que precisam ser preenchidas para garantir que a etapa de análise possa ser realizada.

3. Realização de estudos práticos, sobre problemas reais, para verificar a utilidade da pesquisa e das ferramentas desenvolvidas.

Espera-se que a partir desta metodologia, que garante embasamento teórico forte e experimentação prática, os objetivos do projeto possam ser atendidos e verificados.

# 2. Chaordic Systems

## 2.1. Apresentação da Empresa

A Chaordic Systems é uma empresa focada no desenvolvimento de soluções web de personalização em massa, tendo como principais clientes, empresas que atuam no ramo de comércio eletrônico no Brasil.

A Chaordic é localizada no Parque Tecnológico Alfa, em Florianópolis, Santa Catarina, e faz parte da incubadora de empresas do grupo Certi, chamada Celta.

A empresa foi fundada em 2009, fruto dos trabalhos de mestrado e doutorado de estudantes na área de Inteligência Artificial da Universidade Federal de Santa Catarina. Os resultados dos trabalhos de pesquisa desenvolvidos deram origem a ideia de um produto fortemente baseado em conceitos de sistemas de recomendação, data mining e estatística, que mais tarde viria a ser uma plataforma baseada no modelo de software conhecido como SaaS, sigla do inglês que significa Software as a Service de fornecimento automático de recomendação personalizadas.

Inspirado pela sigla SaaS, o principal produto da Chaordic ficou conhecido como RaaS (*Recommendation as a Service*) e trata-se de uma plataforma voltada quase exclusivamente para recomendação de produtos de sites de comércio eletrônico. O sistema de recomendação é baseado em conceitos de filtragem colaborativa e usa as interações de usuários de *e-commerces* com os produtos dos mesmos para fazer sugestões com o objetivo final de melhorar as vendas.

Hoje a Chaordic possui diversos produtos disponíveis e em operação nos maiores sites de comércio eletrônico do país, além de várias ideias para explorar novos nichos do mercado eletrônico e outros aspectos da WEB.

## 2.2. Estrutura Interna

Desde sua criação em 2009 até hoje a Chaordic desenvolveu uma estrutura de produtos e serviços bastante ramificada e teve sempre seu crescimento pautado em princípios de trabalho bem definidos, tais como autonomia,

liberdade e responsabilidade. Esses princípios definem o perfil das pessoas componentes da organização e serviram para moldar a atual divisão interna de trabalho da empresa.

Hoje, contando com aproximadamente 35 colaboradores a empresa é dividida em 10 times de trabalho: Comercial, Pessoas, Administrativo, Serviços, OnSite, Mail, Pesquisa (também chamado RecSys), Engine, Operações e Dashboard. Cada time possui responsabilidades bem definidas e autonomia para escolher tanto seus objetivos e metas quanto o modo de gerenciar a rotina e definir atividades.

O desenvolvimento e funcionamento dos times são sempre pautados por definições estratégicas gerais à empresa, concebidas e acordadas através do envolvimento de todos os colaboradores. A partir das definições gerais e específicas os times funcionam como pequenas empresas fornecendo serviços uns aos outros, apresentando sempre com uma postura de colaboração mútua.

Este projeto foi desenvolvido dentro do time de Pesquisa, responsável principalmente por manter e desenvolver o módulo de geração de recomendações e as recomendações fornecidas, além de desenvolver pesquisas em sistemas de recomendação com os objetivos de desenvolver os algoritmos existentes, explorar novas técnicas e explorar novos mercados.

# 3. Sistemas de Recomendação

## 3.1.1. Definição

Sistemas de Recomendação são sistemas computacionais e conjunto de técnicas cujo principal objetivo é identificar itens com potencial utilidade para um determinado usuário. Estas sugestões normalmente estão relacionadas a algum processo de decisão, que podem variar em diferentes contextos. Por exemplo: decidir por qual filme escolher em uma locadora de vídeo, qual livro comprar ou em que ações da bolsa de valores investir [10].

Para a geração das recomendações personalizadas, esses sistemas usam técnicas e conceitos diversos, baseados nos ramos de inteligência artificial, data mining, estatística, entre outros [5].

O princípio básico destes sistemas é usar informações disponíveis de usuários, itens e interações entre eles, em um determinado domínio, para então, através da identificação de relações entre estas entidades, tentar fazer previsões sobre quais novos itens podem interessar um determinando usuário em um determinado momento. As recomendações geradas são ditas personalizadas, pois o conjunto de itens recomendados é calculado a partir do conjunto específico das interações de um usuário, o que implica em diferentes recomendações para usuários com diferentes perfis, tornando assim a recomendação feita praticamente única e altamente contextualizada.

## 3.1.2. Importância e Aplicação

Com o avanço e popularização da internet na vida das pessoas, e com o advento de diversos dispositivos que promovem com extrema facilidade a conexão das pessoas com a rede, a disponibilização de informações dos mais diversos tipos e nos mais diversos formatos atingiu níveis antes nunca experimentados. Essa revolução mudou o padrão de vida e comportamento de grande parte das pessoas no mundo, redefinindo o que se entendia por globalização e trazendo seus efeitos para níveis cada vez mais próximos das ações diárias e naturais das pessoas [5].

Graças a essa grande quantidade de informação disponível hoje existe uma dificuldade por parte dos consumidores em filtrar as informações relevantes e úteis, e por parte dos provedores de informação em fazer com que a informação correta atinja os públicos corretos e nos momentos oportunos. Frequentemente "soterrados" de informações, os usuários dos sistemas tendem a fazer decisões pobres, que nem sempre suprem as necessidades, o que por vezes causa frustrações e pode levar a consequências negativas, tais como cancelamentos de compras ou até mesmo abandono do sistema. Por estes motivos, os Sistemas de Recomendação são uma importante parte dos ecossistemas de informação e comércio eletrônico, e representam uma poderosa ferramenta que permite a filtragem e organização de uma grande quantidade de informações e produtos em um domínio.

Estes sistemas ganharam grande importância se transformando em uma área de pesquisa nos meados dos anos noventa, com a aparição das primeiras publicações relacionadas à filtragem-colaborativa. Como pioneiro no uso de Sistemas de Recomendação no setor de comércio eletrônico está a plataforma da Amazon. Outros setores que também se destacaram pelo uso de sistemas similares foram os setores de música e filmes disponibilizados via web. Destacam-se nestes contextos as plataformas de mídia digital Last.fm, iTunes e Netflix, sendo as duas primeiras plataformas de fornecimento de música e o último um site de comercialização de filmes.[2]

Em outubro de 2006, a empresa Netflix anunciou um concurso em âmbito mundial que premiaria o grupo capaz de conceber um algoritmo de recomendação capaz de melhorar o algoritmo usado por eles na época, chamado Cinematch™, em pelo menos 10%, usando uma métrica relacionada ao erro médio associado às previsões. O concurso só teve seu fim em julho de 2009, quando o grupo Bellkor's Pragmatic Chaos, formado por profissionais de diversas empresas importantes de tecnologia, tais como AT&T e Yahoo, venceu o desafio. Este concurso foi muito importante para a história e desenvolvimento de Sistemas de Recomendação, pois deu visibilidade e fez crescer as iniciativas de pesquisa no setor [6].

[2] Last.fm (www.lastfm.com) / iTunes (www.apple.com/iTunes) / Netflix (www.netflix.com)

Outros possíveis domínios de aplicação de Sistemas de Recomendação são, entre os muitos existentes, por exemplo [10]:

- **Sistemas de entretenimento** tais como sites de compartilhamento de músicas (Pandora, Lastfm e Spotify)[3], locação online de filmes (Netflix) e televisão, com recomendação de programação (IPTV);
- **Serviços**, por exemplo, na área de turismo com recomendação de lugares para visitas, locais para hospedagem ou mesmo programas turísticos (Trippy);[4]
- **Redes Sociais** (Facebook e Twitter[5]), que usam Sistemas de Recomendação para filtrar atualizações, sugerir amigos e determinar quais banners de propagandas serão mostrados para um determinado usuário.

### 3.1.3. Dados e Fontes de Informação

Para o cálculo das recomendações em um sistema genérico é preciso considerar dois pontos do processo: os dados a serem apresentados ao sistema e a técnica de recomendação usada para a realização da predição.

Sistemas de Recomendação são fortemente dependentes da quantidade e qualidade dos dados apresentados, assim como vários outros sistemas de identificação de padrões, como, por exemplo, redes neurais, sistemas adaptativos, sistemas com aprendizagem, data mining, entre outros. A qualidade e integridade das informações apresentados tem grande papel na determinação da qualidade do produto final.

Além disso, devemos considerar também o tipo, formato e como os dados serão usados. Existem várias fontes de dados e existe certo desacoplamento entre o dado de entrada do sistema de recomendação e a técnica usada. Abaixo segue uma abordagem de classificação dos tipos de informação mais comumente usados:

1. Relacionadas aos Usuários:

[3] Pandora (www.pandora.com) / Spotify (www.spotify.com)
[4] Trippy (www.trippy.com)
[5] Twitter (www.twitter.com)

- Informações de perfil, tais como preferências informadas pelos usuários a respeito de conjuntos de produtos como filmes, livros, música, atividades, comunidades e etc.

2. Relacionadas aos Itens:

- Informações relacionadas ao conteúdo do item, que pode ser manifestada através de uma descrição textual, um conjunto de tags associadas, ou mesmo pela classificação em categorias e subcategorias;
- Informações adicionais a respeito do item, não relacionadas ao conteúdo, tais como as datas de lançamento, de inclusão no catálogo, da primeira interação de compra ou visualização, do valor do item, o valor de desconto sobre o item. entre outros;

3. Relacionadas às Interações:

- Interações de usuários com outros usuários
- Interações de usuários com itens, que podem ser relacionadas a visitas, compras, avaliação do item por uma opinião quantitativa ou qualitativa, entre outros;

4. Relacionadas à Navegação:

- Informações geográficas, que podem ser adquiridas explicitamente por informações passadas pelo usuário ou através de identificações da navegação, quando aplicado ao contexto web;
- Informações de navegação, como horários e locais navegados em um determinado site, ou banners clicados;

### 3.1.4. Técnicas de Recomendação

Uma vez levantados e apresentados os dados de entrada de maneira adequada ao sistema, é preciso manipulá-los para extrair a relação entre eles, gerando como produto final recomendação de um item com potencial utilidade ao usuário.

Existem diversas abordagens de tratamento dos dados de entrada, cada qual com vantagens, desvantagens e necessidades diferentes para sua utilização. O tipo e formato dos dados, o tipo de técnica de recomendação usada e o tipo de saída determinam o sistema de recomendação.

A escolha destes fatores é extremamente dependente de domínio e deve ser feita com muito critério. É preciso um estudo sobre quais informações estarão disponíveis no ambiente, o tipo de domínio onde se está inserido e o tipo de usuário final ao qual se deseja recomendar, além do objetivo esperado com a recomendação.

Sobre a base do processo de recomendação, os sistemas podem se basear em modelos construídos a partir de dados (coletados e classificados como mostrado anteriormente), ou em modelos observacionais representados, por exemplo, na forma de ontologias ou conjuntos de restrições.

A seguir apresentamos um pequeno resumo das abordagens mais comuns encontradas na literatura [10]:

- **Baseada em Conteúdo**: Está técnica de recomendação se baseia no conteúdo dos itens para construir as relações de similaridades. O uso desta técnica de recomendação requer maneiras específicas de se representar o conteúdo do item. Algumas técnicas de inteligência artificial e processamento de linguagem natural são usadas no sentido de mapear o conteúdo dos itens e relacionar de maneira hábil os mesmos. Essa abordagem apresenta algumas vantagens, como por exemplo, não depender de um período de coleta de informações de interações dos usuários com os itens, processo esse que pode levar um tempo considerável até que uma quantidade suficiente seja coletada. Porém, como desvantagem pode ser citada a dificuldade de se relacionar itens aparentemente diversos, mas que poderiam apresentar um alto potencial de interação, por exemplo, o caso da alta frequência de venda conjunta de fraldas e cerveja nas sextas feiras durante a noite, como constatadas pela rede de supermercado Wal-Mart ao realizar uma mineração dos dados de compras das lojas da rede[6].
- **Filtragem Colaborativa**: Esta técnica de determinação de recomendações é fundamentada em um conceito simples, que se baseia nas interações feitas entre os usuários e os itens do sistema. Usa basicamente as informações de transações feitas pelos usuários do sistema, que podem ser de

[6] Exemplo retirado da revista Exame Online, da editora Abril. (http://exame.abril.com.br/revista-exame/edicoes/0633/noticias/o-que-cerveja-tem-a-ver-com-fraldas-m0053931)

diversos tipos dependendo do contexto (informações de visualizações, de compras, de notas atribuídas à itens e etc.). A recorrência de interações entre itens e usuários aumenta a correlação entre as entidades do sistema, e tal informação é usada para o cálculo das recomendações. A matriz que registra as interações e similaridades é o modelo do sistema, que é constantemente atualizado conforme novas informações são incluídas. Os mecanismos de cálculo das recomendações normalmente requerem um poder computacional alto e que cresce com a inclusão de usuários e inclusão de itens no sistema.

- **Abordagens menos usuais**: Baseados em conhecimento, baseados em dados demográficos, baseados em comunicação (por exemplo, interações em redes sociais) e sistemas híbridos [5, 11];

As técnicas apresentadas compõem as diferentes abordagens mais comuns de exploração das informações disponibilizadas aos sistemas de recomendação. Além disso, outro ponto importante é o agrupamento destas informações e a maneira como o resultado será apresentado. A partir disto, temos duas maneiras básicas de organizar as informações dentro do sistema, sendo elas: centrada no item e centrada no usuário. Cada uma destas abordagens de organização é descrita nos itens a seguir:

- **Centrada no item**: Para cada item do sistema sendo o item de referência calcula-se a similaridade deste com os outros itens do sistema. Essa similaridade é calculada com base na interação de usuários com pares de itens. Quanto maior a frequência de interações comuns entre o item referência e algum outro item, maior a similaridade entre eles. A lista de recomendações para o item referência é construída a partir da ordenação de todos os itens similares a ele, usando os valores de similaridade calculados.
- **Centrada no usuário**: consideram a semelhança entre usuários para efetuar as recomendações. O cálculo de similaridade entre usuários é feito considerando-se o grupo de itens que apresentam interações feitas por um par de usuários. Quanto mais parecido for o grupo de itens com interações comuns, maior a similaridade entre os usuários.

#### 3.1.4.1. Filtragem Colaborativa

A filtragem colaborativa é dos métodos mais populares de geração de recomendações e identificação de padrões de interações. Baseia-se principalmente na consideração de interações conjuntas entre entidades como indicativo de similaridade entre elas [5]. Entidades podem ser usuários ou itens dentro de um contexto. Interações mútuas podem ser, por exemplo, para o caso de usuários, compras de um mesmo determinado produto, e para o caso de itens, as compras feitas em conjunto com outros itens. Essa maneira de usar as frequências conjuntas para determinar as similaridades entre entidades do sistema é chamada de filtragem colaborativa.

As abordagens de filtragem colaborativa centrada no usuário normalmente geram recomendações com qualidade melhor do que as centradas no item, porém, existem várias questões associadas à escalabilidade e desempenho em tempo real que deixam a desejar desta abordagem, o que torna um pouco impraticável sua aplicação em situações na indústria [10]. A abordagem mais comum para o uso de filtragem-colaborativa é a centrada no item, que é uma abordagem bastante usada na Chaordic e que é o foco deste projeto.

Esta seção explica um algoritmo de filtragem colaborativa que usa métodos probabilísticos para fazer o cálculo da similaridade, e que toma como entrada as informações de usuários de um sistema e as interações (compras, visualizações, atribuição de notas e etc.) associadas a ele.

O conjunto de informações de entrada pode ser representado por uma matriz, chamada matriz Usuário-Item, de tamanho M x N, onde M é o número de usuários da matriz e N é o número de itens, denotada por UI e representada na equação 1. A partir da matriz UI constrói-se a matriz Item-Item, onde cada linha e cada coluna indicam um item do universo considerado, e os elementos da matriz correspondem ao número de interações mútuas apresentadas pelo par de itens. A matriz Item-Item é denotada por II, apresenta tamanho N x N, onde N é o número de itens, e é representada na equação 2.

$$UI_{MxN} = \begin{bmatrix} a_{11} & \dots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \dots & a_{mn} \end{bmatrix} \quad \text{(Eq. 1)}$$

$$II_{NxN} = \begin{bmatrix} b_{11} & \dots & b_{1n} \\ \vdots & \ddots & \vdots \\ b_{n1} & \dots & b_{nn} \end{bmatrix} \quad \text{(Eq. 2)}$$

O cálculo das recomendações para um determinado item segue dois passos: construção do modelo e aplicação.

A construção do modelo é a etapa em que as informações de entrada do sistema são combinadas para gerar um grau de similaridade para toda relação entre dois itens. Esta similaridade é considerada posteriormente na etapa de aplicação do modelo para a geração das listas de recomendações.

Na etapa de construção do modelo usam-se diferentes algoritmos para a composição dos dados e cálculo das similaridades. Dois dos algoritmos mais frequentes na literatura são o cálculo baseado no cosseno e o cálculo baseado em probabilidade condicional.

A similaridade baseada no cosseno de vetores é calculada a partir da representação de um item como um vetor no espaço de usuários. O sistema de recomendação recebe como dados de entrada a matriz UI, e então usa essas informações para compor um vetor para cada item do universo. Cada vetor contém em uma determinada posição o número de compras feitas por um dado usuário. A similaridade entre dois itens é o valor do cosseno entre estes dois vetores representando dois itens. O cosseno entre dois vetores é dado pela equação 3, onde i e j são os itens para os quais se deseja calcular a similaridade e R são os respectivos vetores de interações [5, 10].

$$sim(i,j) = \cos\left(\vec{R}_{*,i}, \vec{R}_{*,j}\right) = \frac{\vec{R}_{*,i} \cdot \vec{R}_{*,j}}{\|\vec{R}_{*,i}\|_2 \cdot \|\vec{R}_{*,j}\|_2} \quad \text{(Eq. 3)}$$

A similaridade baseada em probabilidade condicional é calculada a partir das informações de frequências de interação individuais de cada item e das frequências de interações mútuas entre pares de itens. O algoritmo de cálculo de similaridade considerado no projeto foi inspirado por este método de cálculo, e por isso ele é explorado com mais detalhes.

A base deste método para o cálculo da similaridade entre itens é a probabilidade condicional de se ocorrer a interação com um item referência j

dado que a interação com outro item i. Esta probabilidade é dada pela relação da frequência mútua de interação os dois itens, *Freq(i, j)* e da frequência de interação individual do item já interagido *Freq(i)*. Esta relação é expressa na equação 4.

$$P(j|i) = \frac{Freq(i,j)}{Freq(i)} \quad \text{(Eq. 4)}$$

O valor desta probabilidade é a medida de similaridade entre os itens. Sabe-se, entretanto que usando esta medida existe uma tendência em apontar alta similaridade quando o item não referência possui um número alto de interações. Existem algumas correções conhecidas para atenuar este efeito e elas são discutidas no capítulo 5, onde detalhes são explorados sobre a implementação do algoritmo de similaridade estudado especificamente neste projeto.

## 3.2. Provendo Recomendações na Chaordic

Nesta seção serão descritas as principais características do processo de fornecimento de recomendações da Chaordic, com foco especial no algoritmo usado como objeto de estudo deste projeto, descrito com mais detalhes na seção 2.3.1.

O processo de fornecimento de recomendações da Chaordic pode ser dividido em quatro etapas: captação dos dados de interações dos usuários, geração das recomendações elementares, disponibilização das recomendações na plataforma, apresentação das recomendações ao usuário final.

A primeira etapa, de captação dos dados de interações dos usuários, consiste em registrar todas as interações feitas pelos usuários com o website dos clientes. As interações atualmente registradas são: visualizações, compras, cliques em produtos e recomendações e finalização de pedidos.

A segunda etapa consiste na geração das recomendações, usando os dados coletados na primeira etapa como entrada do sistema de geração das recomendações, conhecido internamente na empresa como Engine. O Engine é responsável por encapsular a inteligência do sistema. Através do cálculo das

similaridades entre itens usando um algoritmo de similaridade chamado internamente de “Similar Items”, este módulo é responsável por gerar as chamadas recomendações elementares, que eventualmente podem ser compartilhadas com diferentes produtos, mas que dependendo da maneira como são apresentadas, podem ter objetivos diferentes.

A terceira etapa do processo consiste em disponibilizar essas recomendações geradas nos servidores da Chaordic que são acessados pelos websites dos clientes através de uma arquitetura orientada a serviço. Esses servidores estão hospedados e distribuídos usando a tecnologia de computação em nuvem[7]. A computação em nuvem fornece uma abstração da característica de distribuição do sistema, onde os websites se comunicam com os servidos da Chaordic usando o conceito de WebService, não precisando se preocupar com a localização dos dados.

A quarta e última etapa do processo é a apresentação das recomendações ao usuário final, através da interface do site do cliente da Chaordic. Nesta etapa as recomendações passam por vários filtros, que garantem a qualidade e adequação do produto ao contexto de apresentação. As recomendações são apresentadas por meio de vitrines, localizadas em páginas e locais específicos do site do cliente. Esta etapa de apresentação é crucial para a qualidade do serviço, uma vez que determinam como será a interação entre o usuário final e o produto da Chaordic.

Este projeto se enquadra mais especificamente na etapa dois, e tem como objetivo auxiliar na avaliação das recomendações elementares geradas pelo Engine. A seção 3.2.1 irá apresentar e discutir as particularidades do algoritmo de geração de recomendações usado pela Chaordic e que é objeto de estudo deste projeto.

### 3.2.1. Algoritmo Similar Items

O algoritmo de recomendação usado pela Chaordic tem como base o conceito de filtragem-colaborativa, definido na seção 3.1.4.1, e é um sistema do tipo centrado no item. Sendo assim, usa informações de interação de usuários

[7] Computação em nuvem consiste na utilização de computadores distribuídos, compartilhados e interligados via internet.

com diversos produtos para atribuir a cada produto uma lista de recomendação de itens similares.

O algoritmo é alimentado com informações de dois tipos: compras e visualizações. Dependendo do tipo de entrada, tipos de recomendação diferentes são produzidas. Para este projeto considera-se apenas alimentação com dados de compras, e sendo assim, espera-se que o algoritmo seja capaz de gerar listas de recomendações com itens que apresentem grande probabilidade de serem comprados pelos usuários.

O algoritmo usado apresenta como medida de similaridade a similaridade baseada em probabilidade condicional, muito próxima à apresentada na seção 3.1.4.1. Diferenças significativas foram incluídas no algoritmo da Chaordic para melhorar seu desempenho, porém por questões de confidencialidade estas modificações não são apresentadas.

Como fatores que influenciam as recomendações geradas pelo algoritmo são citados os três principais:

- **Tamanho máximo da lista de recomendações**: O sistema de recomendação pode gerar listas de recomendação tão grandes quanto o número de itens que apresentem interações mútuas com o item referência. Porém, do ponto de vista prática é extremamente custoso gerenciar listas de recomendações com tamanhos muito grandes. É necessário maior tempo para o cálculo das recomendações além de maior capacidade de memória para armazenar todas as informações. Além disso, os itens localizados em posições muito inferiores das listas não apresentam relevância suficiente que justifique sua manutenção. Por relevância entende-se o potencial apresentado pelo item para motivar uma compra pelo usuário. Sendo assim, determinou-se como tamanho máximo das listas de recomendação geradas pelo sistema o número de 40 itens. Itens de referência que possuem listas de recomendação maiores do que este número terão suas listas fixadas neste tamanho e apenas os 40 primeiros itens que apresentarem maior similaridade com o item de referência serão considerados.
- **Filtros de Frequência:** Dada a maneira como a similaridade entre itens é calculada pelos algoritmos de filtragem-colaborativa, todas as informações de interações mútuas são usadas na composição das listas de recomendação.

Entretanto, sabemos também que itens que apresentam pouca frequência mútua de interação podem, às vezes, ser considerados como altamente similares e isto nem sempre é verdadeiro. Quanto maior a frequência mútua maior o indício de similaridade entre os itens. Em algumas situações é preferível não fornecer recomendação a fornecer recomendações de baixa qualidade e, por isso, filtros de frequência são usados, onde pares de itens que apresentem frequência mútua de interações abaixo de patamar escolhido são retirados das listas de recomendação. Apesar de este procedimento diminuir a cobertura (porcentagem do catálogo para o qual é possível fazer recomendações) das recomendações, é considerado essencial para tornar possível o uso do sistema. O valor dos patamares determinados para os filtros depende de vários fatores, e vária para cada situação de uso do sistema.

- **Penalização de itens populares**: Na apresentação do algoritmo de filtragem-colaborativa feita na seção 3.1.4.1 discutiu-se o cálculo de similaridade entre itens baseado na probabilidade condicional. Ao final da seção foi apresentado o problema atribuído à tendência de indicação de alta similaridade quando o produto em referência apresenta um grande número de interações. O algoritmo em estudo resolve este problema através da inclusão de um mecanismo de penalização. Itens com grande número de interações frente à maioria podem ser classificados como itens muito populares. Uma forma simples de penalizar estes itens é diminuir a similaridade entre este item e outro proporcionalmente à quantidade de compras apresentada por ele. A variação da similaridade decresce com o aumento do número de interações do item de referência, e a taxa de decrescimento pode ser controlada por um parâmetro que será chamado a partir de agora por parâmetro de penalização de itens muito populares [5]. As seções 5.2 e 6.2 discutem um estudo de caso onde a influência dos valores deste parâmetro na assertividade e novidade do sistema será apresentada.

## 3.3. Avaliação de Sistemas de Recomendação

O processo de avaliação de sistemas, independente do tipo e propósito é de extrema importância, pois sem a análise do que está sendo entregue não é

possível determinar a qualidade do produto final e garantir que as qualidades e funcionalidades desejáveis estão sendo corretamente entregues.

A avaliação é requerida em diferentes estágios do ciclo de vida do sistema, e em cada estágio ela pode apresentar um propósito e medir coisas distintas. Tomando como base os Sistemas de Recomendação, como etapas do ciclo de vida destacam-se [10]:

- Período de **design do sistema**, quando são escolhidas as diretrizes principais do sistema. Nesta etapa ocorre um estudo sobre quais dados estão disponíveis para servir como entrada, e como esses dados serão tratados, ou seja, quais técnicas de recomendação se mostram mais adequadas (baseada em conteúdo, filtragem-colaborativa, baseada em conhecimento e etc.). Nesta etapa várias opções podem ser experimentadas. Uma vez determinada a técnica são definidos os detalhes sobre a inteligência do sistema, com, por exemplo, o algoritmo a ser usado para determinar as similaridades. Nesta etapa o processo de avaliação é importante para determinar entre as diversas opções disponíveis qual é a mais adequada para o contexto. De posse da técnica e dos algoritmos, ocorrem os primeiros ajustes, chamados na Chaordic de tunagem, que consiste na determinação dos diversos parâmetros a serem usados no processo de geração da recomendação.
- Período do **sistema em produção,** onde a solução é colocada em produção e os resultados são avaliados. Nesta etapa pode-se verificar o impacto real do sistema sobre o ambiente, sendo possível então a coleta de dados on-line para alimentar processos de avaliação mais confiáveis. Está etapa é constantemente reavaliada sujeita a varias alterações, principalmente nos primeiros períodos de uso. Aqui, o custo de errar é maior do que na etapa de design, pois as interações acontecem com usuários reais.

Para cada etapa do ciclo de desenvolvimento existem processos de avaliação adequados para captar as características desejadas. Na etapa de design, onde existem muitas opções para resolver o problema proposto é preciso um processo de avaliação rápido, que auxilie na escolha das melhores opções para serem colocadas em produção. Já na etapa de produção, onde apenas os melhores candidatos foram concretizados, a avaliação ocorre no

sentido de coletar resultados e verificar o real valor agregado pelo sistema. Normalmente nesta etapa o processo de avaliação é mais lento e centrado no usuário e nas interações do mesmo com o sistema.

Mais comumente relatados na literatura, os processos de avaliação de Sistemas de Recomendação divide-se em três tipos [11]:

- **Avaliação Off-line**: Este tipo de avaliação é dito off-line, pois se trata da observação do sistema antes da sua colocação em produção. Trata-se de um processo de avaliação ágil, onde a avaliação ocorre sobre as características das recomendações geradas e não dos resultados produzidos por elas. Nesta etapa determinam-se os comportamentos e características esperados para o sistema de avaliação e a análise é feita com a ajuda de indicadores de desempenho relacionados a estas características. É um processo de avaliação com a vantagem de ser ágil, podendo ser repetido facilmente para diversas variações, porém sua confiabilidade é baixa, pois um bom desempenho dos indicadores não necessariamente é traduzido em valor de negócio;
- **Estudos com usuário**: Este tipo de avaliação conta com a opinião de um grupo de usuários específico que avalia as características do sistema e a utilidade do mesmo. Normalmente é feita através de uma série de entrevistas e questionário e tem como objetivo captar a percepção do usuário sobre o que o sistema está oferecendo. Pode-se avaliar não só o conteúdo das recomendações geradas, mas também o formato de apresentação das mesmas e outras características do sistema. Este processo de avaliação é mais custoso do que o processo off-line, tanto em tempo como em investimento financeiro, e por isso é muitas vezes inviável. A determinação do grupo de avaliação é uma escolha que pode influenciar os resultados, e muitas vezes conseguir um grupo representativo é difícil. Além disso, o fato do próprio grupo saber que está sendo avaliado pode incluir vieses nos resultados;
- **Avaliação On-line**: Este tipo de avaliação tem como objetivo captar o resultado da verdadeira medida de valor do negócio, ou seja, consegue expressar o real ganho de valor que o sistema de recomendação pode trazer. Normalmente é feito também pela análise de indicadores de desempenho, e pode usar algumas técnicas de experimentação online conhecidas, como os testes AB, ou testes de divisão em grupos de tratamento e controle [8].

Em cada etapa de avaliação deseja-se analisar o sistema de diferentes maneiras, sendo que para cada sistema e para cada propósito de uso do mesmo algumas características são mais importantes em detrimento de outras.

A seção 3.3.1 apresenta maiores detalhes sobre o processo de avaliação off-line, que é o processo escolhido como escopo deste projeto. Na seção 3.3.2 descreve o processo de avaliação dos sistemas de recomendação da Chaordic e a visão da empresa sobre um processo completo de avaliação. A seção 3.3.3 apresenta as principais propriedades associadas aos sistemas de recomendação, e que servem de base para os processos de avaliação.

### 3.3.1. Avaliação Off-line

Experimentos off-line são experimentos feitos usando conjuntos de dados pré-coletados. A partir destes dados procura-se simular o comportamento dos usuários com o sistema. O objetivo é identificar características relevantes do sistema em estudo de maneira a caracterizá-lo. As análises feitas e os métodos usados dependem do que se pretende com o experimento.

A vantagem deste tipo de experimento é que grande quantidade de informação pode ser descoberta com pouco esforço e tempo. Novos experimentos podem ser feitos sobre os dados e pequenas variações podem ser testadas agilmente. Este processo é bastante adequado quando dispomos de um conjunto amplo de casos a serem avaliados.

Entre desvantagens podemos citar a limitação dos aspectos do sistema possíveis de serem levantados. O que se deseja, em última instância, durante a avaliação de um sistema de recomendação é verificar sua utilidade para o usuário final. Experimentos off-line promovem a investigação de certas características que supostamente levariam os sistemas a atingir bons resultados, mas não há garantia que um bom desempenho off-line se reflita em bom desempenho do sistema em produção (on-line) . Sendo assim, processos de avaliação off-line são apenas uma das etapas do processo total de avaliação de um sistema de recomendação [7, 11].

Em um sentido mais amplo, o que se espera da realização de experimentos off-line é a filtragem de um conjunto de sistemas de recomendação potencialmente candidatos a um bom desempenho, para um conjunto menor,

que possa ter seu processo de avaliação continuado e testado com o objetivo de avaliar real retorno.

A determinação do conjunto de dados a serem usados para a realização deste tipo de experimento é a primeira etapa. Os dados necessários e o formato dependem diretamente de que características deseja-se descrever. O método de coleta e apresentação também deve ser observado, uma vez que, dependendo da organização dos dados e de suas características gerais, variações dos casos de análises podem ser privilegiadas. O importante nesta etapa é que os dados sejam os mais próximos possíveis de dados de um sistema real, e que a quantidade e conteúdo sejam coerentes e suficientes para a avaliação.

Dados que apresentam características temporais, como sazonalidade, deve ter atenção especial. O desempenho constatado para um determinado período pode ser bem diferente do constatado para outro. Por exemplo, para um sistema de recomendação de filmes, influências de determinadas épocas do ano, como a realização da premiação do Oscar, ou as férias de verão, podem causar influencias significativas sobre o resultado. Para estes casos, sempre que possível, as comparações devem ser feitas usando os mesmos períodos de tempo, ou fazer considerações de comparações entre pontos semelhantes de um ciclo, como por exemplo, comparar dados de dois anos diferentes, porém usando sempre os mesmos meses de referência.

### 3.3.2. Processo de Avaliação na Chaordic

Os projetos formais relacionados à avaliação dos sistemas de recomendação na Chaordic começaram há cerca de um ano. Três projetos estão sendo atualmente desenvolvidos, onde cada um deles contemplando uma das abordagens de avaliação citadas na introdução do capítulo.

Dois destes projetos estão sendo liderados pelo time de Pesquisa da empresa, e estão relacionados aos processos de avaliação On-line e Off-line. O terceiro projeto está relacionado à determinação das preferências dos usuários e está relacionado ao processo de estudos com usuários descritos anteriormente.

Em outubro de 2011 teve inicio na Chaordic o projeto de avaliação On-line com o objetivo de comprovar o impacto dos produtos da empresa sobre o faturamento dos clientes. Optou-se por um método de avaliação chamado método de teste AB, ou método com grupo de tratamento e grupo de controle. O projeto foi desenvolvido durante os dez meses subsequentes e os esforços relacionados à pesquisa e prototipagem estão praticamente terminados. Recentemente, em abril de 2012, conseguimos a aprovação da publicação de um artigo no evento referência do ramo de pesquisa em sistemas de recomendação, o "*RecSys*", organizado pela "*Association for Computing Machinery*" (ACM) e foi considerado uma grande conquista para a empresa.

### 3.3.3. Propriedades de Sistemas de Recomendação

A avaliação de Sistemas de Recomendação vem amadurecendo nos últimos anos, conforme os pesquisadores e desenvolvedores destes sistemas foram percebendo que nem sempre as medidas relacionadas à assertividade eram traduzidas em valor de negócio. Razões históricas relacionadas aos primeiros sistemas de filtragem colaborativa, que se baseavam em notas dadas por usuários aos itens dos sistemas como sendo as informações usadas para a geração das recomendações, fizeram com que o desempenho ficasse altamente atrelado à medidas de alguns poucos indicadores, tais como Recall, Precisão, Medida F e Raiz Média Quadrática [11].

Destaca-se o fato de Sistemas de Recomendação hoje serem usados em diferentes contextos, e que em vários deles, não é mais suficiente apenas recomendar itens com características de conteúdo obviamente parecidas. Por exemplo, no caso de um sistema de recomendação de destinos turísticos, não é suficiente recomendar cinco pacotes para o mesmo destino, variando apenas o local de hospedagem, mas possivelmente é mais efetivo recomendar cinco destinos diferentes [9].

Com o objetivo de avaliar características diversas de Sistemas de Recomendação, foram listadas 11 propriedades para avaliação de sistemas de recomendação [11]. Cada propriedade deve ser considerada no contexto do sistema de interesse, e pode apresentar maior ou menor importância, dependendo dos objetivos e das restrições. As próximas seções discutem

brevemente o significado de cada uma das propriedades, além de levantar exemplos em que a propriedade em questão apresenta um destaque especial no processo de avaliação. A discussão de cada propriedade terá como foco os sistemas baseados em filtragem-colaborativa, mais especificamente os sistemas centrados no item, pois é o objeto de estudos do projeto.

#### 3.2.1.1. Assertividade

A propriedade de assertividade é a propriedade mais explorada nas pesquisas em Sistemas de Recomendação. Ela tem seus conceitos fortemente baseados nos campos de pesquisa de Recuperação da Informação [9].

Como o objetivo da maior parte dos sistemas é prever a próxima ação do usuário, onde esta ação pode ser desde uma nota atribuída a um filme ou um livro que pode despertar interesse e ser comprado. Esta propriedade visa quantificar a capacidade do sistema em realizar tais acertos.

Uma técnica muito usada para realizar avaliações desta propriedade é a técnica estatística chamada de “Cross-Validation”, que é usada para generalizar as medições feitas por um preditivo. Esta técnica se baseia na divisão do conjunto de dados em um conjunto de treinamento e um conjunto de validação. O conjunto de treinamento é usado para construir o modelo do sistema e o conjunto de validação é usado para avaliar os resultados das predições.

Alguns indicadores relacionados a esta propriedade são os indicadores Recall, também chamado de taxa de acerto, e o indicador Precisão.

Um domínio onde esta propriedade tem grande importância é, por exemplo, o ambiente de comércio eletrônico, onde a medida de desempenho global do sistema é dada pelas taxas de conversão (Porcentagem da apresentação de produtos que efetivamente viram compras), que são analisadas do ponto de vista de sistemas de recomendação estão ligadas diretamente à capacidade do sistema em mostrar itens que promovam no usuário a ação de compra.

#### 3.2.1.2. Cobertura

O conceito de cobertura está relacionado à medição da capacidade do sistema em fazer recomendações dentro de um universo de produtos. No

contexto de sistemas de recomendação centrados no item, cada item possui associado uma lista de itens a serem recomendados. Todos estes itens fazem parte primeiramente do catálogo de produtos do sistema, mas podem também ser agrupados de acordo com outros contextos, como por exemplo, apenas o universo dos produtos comprados em um determinado período, ou dos itens visualizados neste período, ou até mesmo subgrupos destes formados através da aplicação de filtros, como por exemplo, todos os produtos que possuem pelo menos quatro interações de compras, e assim por diante.

A partir da definição do universo de produtos que se deseja analisar, os indicadores relacionados à cobertura dão a informação de quanto deste universo está presente nas possíveis recomendações.

Para analisar esta propriedade pode-se fazer uso de técnicas de teoria de grafos, e com isso é possível analisar as recomendações gerais além de um ponto de vista de “caixa preta”. Com análises baseadas nos grafos de recomendação podemos identificar conglomerados de produtos, identificar como os produtos e recomendações estão distribuídos, além de outras características relacionadas a outras propriedades que também são potencialmente importantes [3].

Um grande problema associado à Sistemas de Recomendação que usam filtragem-colaborativa é chamado de efeito de *Cold-Start*. Alguns itens do catálogo são pouco populares, ou tiveram inclusão recente no sistema, e por isso não possuem dados suficientes de interações para que recomendações possam ser feitas. Por consequência da ausência de recomendações esses produtos são mostrados com menos frequência, o que agrava o efeito de diminuição da similaridade dos mesmos frente aos outros produtos mais populares ou com mais tempo de catálogo. Normalmente estes produtos estão localizados no final da cauda de distribuição de interações, a chamada cauda longa [3].

A propriedade de cobertura é muito importante em ambientes de comércio eletrônico onde as vitrines de produtos desempenham um papel importante na divulgação dos mesmos e no incentivo de compra para os usuários.

#### 3.2.1.3. Convicção

A convicção de um sistema de recomendação está relacionada à confiança do sistema na própria recomendação fornecida. Normalmente esta propriedade é mensurada através do grau de similaridade das recomendações indicadas pelo sistema. Pode-se optar por recomendar apenas quando o grau de convicção estiver acima de certo patamar, para evitar que, mesmo que haja recomendações disponíveis, elas sejam fornecidas com baixa qualidade.

Em termos de comparação de sistemas, é possível que dois sistemas apresentem desempenhos similares de assertividade, porém com capacidades diferentes de prever a confiança em sua recomendação. Provavelmente a melhor opção seria optar pelo sistema que consegue determinar com um intervalo menor de confiança qual é a recomendação com maior probabilidade de acerto [11].

Esta propriedade é importante para sistemas nos quais a consequência de se apresentar um item que não agrade um usuário possa ser negativa. Um exemplo de consequência negativa é fazer com que o usuário deixe de usar o sistema de recomendação e possivelmente até deixe de retornar ao ambiente em questão. Por exemplo, recomendações de itens com alto valor em ambientes de comércio eletrônico podem estar sujeitos a este problema, pois caso o item tenha sido recomendado por um sistema e cause desagrado ao usuário, a sensação de perda pode ser maior e consequentemente pode diminuir a confiança no sistema de recomendação.

#### 3.2.1.4. Confiança

Esta propriedade é semelhante à propriedade de convicção, mas é vista do ponto de vista do usuário. Ela tem o intuito de representar a confiança do usuário sobre as recomendações fornecidas. Normalmente sistemas em que os usuários apresentam grande confiança tem maior número de visitantes que retornam. Este efeito é chamado de fidelização e é um efeito positivo e desejado.

Algumas medidas podem ser feitas nos sistemas de recomendação para melhorar esta característica como, por exemplo, proporcionar uma explicação

sobre a recomendação fornecida. Existem categorias de sistemas de recomendação chamados “Sistemas conversacionais” [8], que não só apresentam as informações sobre a geração da recomendação, mas também dão a possibilidade do usuário controlar tanto o que é recomendado como quais dados serão usados para gerar futuras recomendações. Acredita-se que estas características aumentem a confiança do usuário sobre o sistema e consequentemente promovem maior uso [11].

Alguns sistemas em que esta característica é importante são os sistemas de exploração de novas músicas, disponibilizados por plataformas de fornecimento de media online, como o Last.fm e o Grooveshark[9], onde existe intensa interação do sistema com os usuários.

#### 3.2.1.5. Novidade

Novidade é a propriedade que exprime a capacidade do sistema em recomendar itens não conhecidos pelo usuário [12]. É uma propriedade desejada em cenários específicos, onde o objetivo da recomendação tem o intuito de promover a descoberta. Os sistemas do Last.fm e do Grooveshark citados na seção anterior têm ferramentas específicas que auxiliam o usuário à descobrir novos artistas, sendo assim sistemas que apresentam explicitamente a importância desta propriedade.

Dependendo do contexto, este aspecto das recomendações pode ser preferido em detrimento de outros aparentemente importantes, como a assertividade. Sistemas que apresentam forte caráter de novidade tem potencial para apresentar também baixa assertividade, porém este efeito será positivo se o usuário estiver ciente de que aquilo que irá receber não é foi produzido a partir de uma similaridade óbvia. Mais uma vez, nestes sistemas, o caráter de explicação se mostra importante [2].

[8] Tradução literal do inglês “Conversational Systems”
[9] Grooveshark (www.grooveshark.com)

#### 3.2.1.6. Serendipidade

O conceito de serendipidade[10] está relacionado à capacidade do sistema em fornecer recomendações que causem ao mesmo tempo espanto e surpresa ao usuário, mas que tenha um caráter positivo [11]. Está relacionado à descoberta de novos itens que não seriam descobertos sem a ajuda do sistema, apesar de apresentarem semelhança aos itens já interagidos pelo usuário.

O termo foi adotado pela comunidade científica e serve para sumarizar vários conceitos e propriedades desejadas em apenas uma.

#### 3.2.1.7. Diversidade

A diversidade associada a um Sistema de Recomendação expressa a capacidade de gerar recomendações que ao mesmo tempo apresentem coerência e sejam distintas entre si. Normalmente o conceito é atribuído a uma lista de recomendações, em que o que se deseja medir é o quão diferente os itens desta lista são entre si [14].

Em domínios específicos esta propriedade é bastante apreciada, pois o valor está justamente em apresentar diversas opções para o usuário escolher, ao invés de mostrar diversos itens com similaridade óbvia entre si. Por exemplo, em um site de turismo, apresentar cinco destinos turísticos e variações apenas de estadia na página principal pode ter efeito menos positivo do que apresentar 5 opções diferentes de destinos.

Nos sistemas em que está propriedade é importante, decisões de comprometer a assertividade e outras propriedades, a princípio consideradas positivas, podem ser feitas e podem se refletir em alta comprovação de valor.

#### 3.2.1.8. Risco

Existem sistemas de recomendação nos quais os contextos onde são aplicados apresentam inerentemente potencial risco negativo quando ocorre o

[10] Tradução literal do inglês “serendipity”. A palavra é derivada de um termo usado em um livro de contos Persa, escrito por Horace Walpole, que conta a história de três príncipes do Ceilão que viviam fazendo descobertas inesperadas.

uso da recomendação. Em sistemas de recomendação usados para dar suportes às decisões de negócio, como por exemplo, sistemas usados no gerenciamento de títulos no mercado de ações, ou sistemas usados para dar suporte às decisões de engenharia, os efeitos negativos de uma recomendação podem causar mais do que apenas uma impressão ruim no usuário.

Dependendo do sistema onde este conceito é observado, pode-se usar a avaliação de risco como insumo para o cálculo das recomendações. Por exemplo, no mercado de ações, pode-se desejar uma recomendação de título que ao mesmo tempo em que potencializa o lucro oferece risco baixo, ou um risco até certo patamar determinado [11].

#### 3.2.1.9. Robustez

A robustez de um sistema está associada à capacidade do mesmo em manter a qualidade das recomendações fornecidas mesmo na presença destes dados aviesados. Por serem sistemas preditivos, os sistemas de recomendação apoiam-se fortemente na qualidade dos dados fornecidos como entrada. Em sistemas que funcionam em ambientes de produção real, como por exemplo, os sistemas em ambientes de comércio eletrônico, o conteúdo dos dados está sujeito a sofrer alterações e de usuário, podendo então haver dados corrompidos ou dados tendenciosos.

Alguns sistemas de recomendação se apoiam em ferramentas estatísticas para fazer a identificação do que chamamos de outliers. Outliers em sistemas de recomendação são entidades que apresentam comportamento muito diferente da grande maioria do grupo. Por exemplo, existem alguns usuários nos sistemas de comércio eletrônico usados para fazer testes, e que apresentam um número alto de compras, faturamento acumulado e interações entre itens. Outro tipo de usuário conhecido são os usuários lojas, que apesar de usuários reais, com interações e compras no sistema, não apresentam um padrão de consumo bem definido, e acabam gerando informações de similaridades que não necessariamente agregam informações capazes de ajudar o sistema de recomendação a prover recomendações relevantes. Estes

usuários são então filtrados e desconsiderados na realização dos cálculos de similaridades.

Outro tipo de usuários que corrompem os dados são usuários criados com o objetivo de enaltecer alguma entidade do sistema. Por exemplo, em sistemas de recomendação de filmes baseados em notas de usuários, existem Web Robots criados com o objetivo de simular diferentes usuários e atribuir notas altas a determinados filmes, aumentando assim a popularidade dos mesmo e consequentemente aumentando a probabilidade destes aparecerem em locais de destaque nas vitrines. A identificação de Web Robots é um campo já existente na ciência da computação e as ferramentas produzidas nele são cada vez mais incorporadas aos contextos de sistemas de recomendação [1].

#### 3.2.1.10. Adaptividade

A adaptividade de um Sistema de Recomendação está associada à capacidade do sistema em atualizar as recomendações feitas conforme novos dados vão servindo de alimentação. Sistemas de Recomendação reais precisam operar em um regime de atualização constante, sendo que a frequência de necessidade de atualização depende do tipo de sistema. A frequência deve seguir a dinâmica dos ambientes onde os sistemas estão inseridos, para que novas tendências possam se refletir também nos sistemas de recomendação.

Por exemplo, para um sistema que recomenda artigos jornalísticos, a frequência de atualização de ordem superior a um dia não é suficiente, pois a dinâmica de atualização dos objetos de recomendação, no caso notícias, é mais acelerada do que a dinâmica de atualização. As notícias se tornariam obsoletas mesmo antes de ter a oportunidade de serem apresentadas aos usuários.

#### 3.2.1.11. Escalabilidade

A escalabilidade de um sistema é um termo emprestado da ciência da computação e de outros ramos da tecnologia e está relacionada à capacidade do sistema em crescer e manter seu desempenho conforme novos dados vão sendo acrescentados e novos volumes de requisição são feitos.

Sistemas Reais devem ser capazes de se adaptar à novos registros de usuários fazendo novas requisições sem apresentarem mudanças significativas no desempenho do fornecimento das recomendações. O aumento no tempo de resposta do fornecimento de recomendações online pode causar grande influência no desempenho geral do sistema, mesmo quando da ordem de poucos milissegundos.

# 4. Framework de Avaliação Off-Line

Como contribuição prática do projeto foi proposta a construção de um framework que auxiliasse no processo de experimentação off-line.

Em desenvolvimento de software, um framework é uma abstração que une códigos comuns para a realização de um conjunto de funcionalidades genéricas, e normalmente é usado para dar suporte ao desenvolvimento de outros projetos [8].

O objetivo principal deste projeto foi o desenvolvimento dos conhecimentos em avaliação off-line dos sistemas da Chaordic, e a partir disso o desenvolvimento do framework de experimentação foi necessário para dar suporte inicial ao processo de pesquisa.

Levando em consideração que o processo de avaliação off-line é um projeto novo e em recente desenvolvimento na Chaordic, levantaram-se algumas propriedades desejadas para a ferramenta:

- Garantir **flexibilidade** para ser expandida juntamente com os esforços de pesquisa. Como o trabalho de pesquisa teórica sobre os indicadores de desempenho está no início, foi prevista a necessidade de constante integração de novos indicadores, e eventualmente novas fontes de dados. O projeto do framework considera estas necessidades.
- Prover **agilidade** ao processo de experimentação, para garantir que os ciclos de desenvolvimento da teoria não sejam desacelerados pela necessidade da realização de funções manualmente.
- Projeto considerando **reuso**. A teoria de avaliação desenvolvida neste projeto tem como plano de uso o acompanhamento de certos sistemas da Chaordic, e têm-se planos de integração com outras ferramentas de monitoração e apresentação de resultados. O intuito é que o framework possibilite a realização do processo de análise, mas que outras ferramentas possam ser integradas em apenas algumas etapas e usufruir das funcionalidades específicas sem necessidade de integração total com o framework.

A primeira etapa de desenvolvimento do framework foi o levantamento dos requisitos funcionais e técnicos relacionados. A elucidação destes requisitos é apresentada na seção 4.1.

Definidos os requisitos do sistema sobre quais funcionalidades seriam requeridas, parte-se para a discussão da modularização, sempre tendo em vista os princípios de reusabilidade, flexibilidade e agilidade.

Nesta etapa definiu-se a linguagem de programação a ser usada. Escolheu-se a linguagem de programação Python como linguagem para prototipagem. O Python é uma linguagem de programação de propósito geral, disponibilizada gratuitamente, e que possui suporte para vários sistemas operacionais, suporte a paradigmas de orientação a objeto e programação funcional, além de ser constantemente desenvolvida por uma comunidade ativa. É uma linguagem de alto nível, que possibilita desenvolvimento rápido e fácil integração entre sistemas, caracterizando-se como uma ótima ferramenta para prototipagem graças ao seu código enxuto, altamente legível e à grande facilidade de aprendizado[11].

A escolha foi pautada em duas principais razões. A primeira razão está relacionada ao fato da linguagem ser uma das usadas para desenvolvimento de software dentro da empresa, o que iria garantir então a possibilidade do framework ser amparado e desenvolvido em cooperação. A segunda razão é a apresentação de rápida curva de aprendizado.

A seção 4.2 discute as decisões tomadas quanto à arquitetura e apresenta a proposta implementada.

Com a estrutura do sistema discutida e acordada, o próximo passo foi a definição dos primeiros indicadores a serem incorporados no protótipo. Levou-se em consideração o que seria necessário para as primeiras experimentações. A seção 4.3 apresenta as discussões e decisões sobre as propriedades e indicadores considerados prioritários e a seção 4.4 apresenta os detalhes de implementação e comportamento de cada um dos indicadores escolhidos.

[11] Descrição da linguagem tirada do site oficial. www.python.org.

## 4.1. Definição dos Requisitos

A partir dos três princípios básicos definidos para o framework (reuso, flexibilidade e agilidade) foram definidos os requisitos de funcionalidades e características desejadas. Estes requisitos foram divididos em sete aspectos principais:

1. Realizar processos completos de análise.
   a. Por análise entende-se a programação do cálculo de um conjunto de indicadores sobre um conjunto de dados especificados.
   b. O processo completo consiste no ciclo de aquisição dos dados, adequação do formato dos mesmos, cálculo dos indicadores e armazenamento dos resultados;
   c. A configuração das análises registradas e executadas deverá persistir no sistema. Deverá ser possível efetuar consultas sobre quais análises foram feitas e os parâmetros usados, além de permitir nova execução aproveitando os parâmetros de configuração.

2. Flexibilidade de inclusão de novos indicadores.
   a. Durante o processo de pesquisa e desenvolvimento novos indicadores ou possíveis variações precisarão ser incluídos na ferramenta. Esta deve prover um sistema de inclusão a partir de um padrão definido.

3. Gerenciamento dos dados.
   a. Este requisito cobre tanto a questão de inclusão de novos tipos de dados quanto a possibilidade de modificar a fonte de coleta dos mesmos.
   b. Este requisito é necessário para garantir que a inclusão de novos indicadores não seja limitada pelos dados disponíveis no sistema.
   c. Os dados armazenados pela Chaordic estão distribuídos entre vários repositórios, localizados em diferentes máquinas, e por isso é necessário identificar as diversas fontes possíveis para o mesmo tipo de dado.
   d. As fontes de dados podem ser construídas usando tecnologias diferentes. O framework deve abstrair estas diferenças, fazendo com que o

usuário possa prover mecanismos distintos de acesso aos meios e repassar os dados ao processo de análise com o mesmo formato.

4. Filtragem dos dados de entrada.
   a. O sistema deverá prover capacidade de filtragem dos conjuntos de dados usados para o cálculo de indicadores[12].
   b. A configuração dos filtros dependerá de cada tipo de dado e da análise sendo feita. Essa configuração deverá fazer parte da configuração global do processo da análise.

5. Gerenciamento do formato dos dados.
   a. Cada indicador será desenvolvido de maneira independente e poderá ter requisitos diferentes quanto ao formato dos dados de entrada. O sistema deverá ser capaz de prover meios de alterar o formato dos dados.

6. Agilidade para re-execução de análises.
   a. Como cada análise está atrelada a configurações, tanto dos filtros e fontes de dados usados quanto dos indicadores considerados, previu-se a necessidade de re-execução destas análises usando diferentes parâmetros, e pretende-se que isso seja possível com agilidade, uma vez que a coleta dos dados é demorada dado o volume processado.

7. Armazenamento das informações relacionadas às análises feitas.
   a. Os resultados finais das análises, bem como possíveis resultados intermediários associados ao processamento de cada indicador deverão ser armazenados para posterior consulta.

A especificação dos requisitos detalhou as funcionalidades mínimas necessárias para que a ferramenta gerada pudesse auxiliar na realização dos primeiros estudos de caso. Uma vez que o framework se torne produto, novas funcionalidades serão agregadas.

[12] O mecanismo de filtragem é discutido melhor na seção 4.2.

## 4.2. Arquitetura e Modularização

A partir dos princípios definidos para o framework e dos requisitos funcionais definidas até então, definiu-se uma modelagem para o sistema pautada em divisão por módulos. Uma característica do Python é a facilidade de alta modularização. A figura 1 apresenta a arquitetura modular e o relacionamento entre os módulos. Previu-se uma interface por linha de comando, pensando na fácil automatização de tarefas.

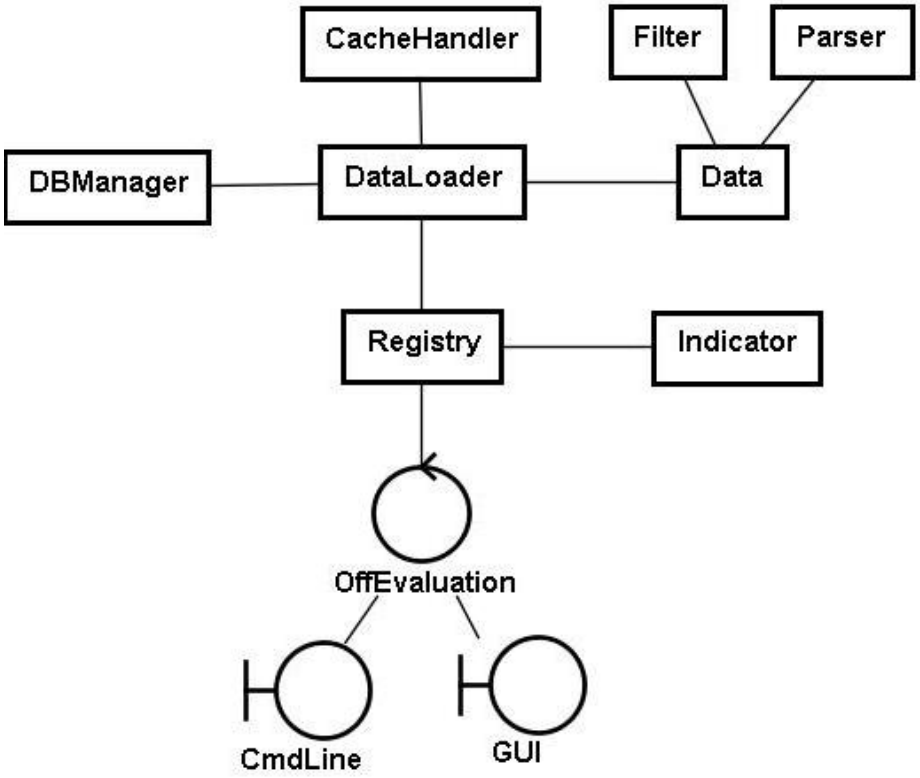


Figura 1 Diagrama de Módulos - Framework

O módulo de registro, chamado “Registry”, é responsável por gerenciar os indicadores, tipos de dados e fontes de dados cadastradas no sistema. Esse módulo é responsável pela manutenção dessas informações, e disponibilização das mesmas quando necessário.

Para a inclusão de novos indicadores foi desenvolvido o módulo chamado “Indicator”, que define o padrão como novos indicadores devem ser implementados para serem reconhecidos pelo framework. Cada indicador pode ser concretizado em módulos distintos feitos em Python, que se registrariam

usando o módulo “Registry” para serem reconhecidos pela ferramenta. Outro aspecto da independência dos módulos de indicadores é a possiblidade do uso dos mesmos por outros sistemas, sendo esta integração já prevista para etapas posteriores do projeto.

De maneira similar à integração dos indicadores tem-se a integração de novos dados. Cada tipo de dado integrado ao sistema está atrelado a um conjunto de funções de filtragem e formatação, identificadas no diagrama pelos módulos “Filter” e “Parser”. O registro de um dado novo no sistema na prática seria o fornecimento de novos módulos “Data”, “Filter” e “Parser”.

O gerenciamento dos dados é feito pelo módulo “DataLoader”, responsável por reconhecer quais tipos de dados estão disponíveis e por fornecer a abstração destes dados ao processo de cálculo dos indicadores. Este módulo é responsável por gerenciar a coleta dos dados, seja ela feita de uma fonte local ou distribuída, e por gerenciar os formatadores e filtros a serem aplicados sobre os dados brutos. O conjunto de dados fornecido para os módulos de indicadores já apresenta então o conteúdo e formato esperados.

Para o gerenciamento das fontes de dados temos os módulos DBManager. Cada módulo deste tipo implementa a interface com uma fonte de dado diferente. Neste módulo estarão os detalhes de integração das tecnologias de captura dos dados, fornecendo uma abstração ao módulo “DataLoader” sobre onde e como realizar efetivamente a coleta.

Para o requisito de re-execução das análises propôs-se um módulo responsável pelo gerenciamento de um cache local com os dados usados nas análises. A responsabilidade do módulo “CacheHandler” é evitar que os mesmos dados tenham que ser buscados mais de uma vez das fontes remotas. A re-execução de uma análise faria o carregamento dos dados localmente, uma vez que eles já tenham sido carregados e armazenados anteriormente. A manutenção destes dados também é de responsabilidade do módulo “CacheHandler”.

Todos estes módulos foram integrados na primeira versão da ferramenta de experimentação. O objetivo desta primeira integração foi fornecer meios para que os estudos de caso desenvolvidos para este projeto pudessem ser concretizados. Os resultados obtidos são apresentados e discutidos nas seções 5 e 6.

## 4.3. Levantamento de Indicadores

Na seção 3.3.3 foram apresentadas as principais propriedades para avaliação de Sistemas de Recomendação. A partir da análise destas propriedades foi realizada uma extensiva pesquisa nos principais repositórios de trabalhos científicos, destaca-se o repositório da Association for Computing Machinery (ACM), organização que promove o desenvolvimento científico dos Sistemas de Recomendação, através de iniciativas como o Recsys[13].

A partir da pesquisa em artigos científicos, surveys de pesquisa e outras publicações do gênero listaram-se os principais indicadores de desempenho usados na literatura para avaliação de Sistemas de Recomendação. O principal objetivo desta relação de indicadores foi entender como a comunidade científica vem explorando as diversas métricas de avaliação, e em que estágio de desenvolvimento e compreensão esses estudos se encontram. Ao final deste trabalho de levantamento de indicadores, foram escolhidas quatro propriedades sobre as quais o desenvolvimento da ferramenta de experimentação foi focado. A partir da definição das principais propriedades, o próximo passo foi listar os indicadores relacionados que parecessem mais pertinentes à avaliação dos nossos sistemas, bem como detalhes de implementação e interpretação desses indicadores.

Ao final do período de estudo foi elaborada uma lista de indicadores relacionados às 9 das propriedades citadas na seção 3.3.3. O Anexo A mostra a relação dos indicadores levantados para as 4 propriedades escolhidas como foco do projeto.

As propriedades escolhidas como foco deste projeto foram: Assertividade, Cobertura, Diversidade e Novidade.

A escolha das propriedades de Assertividade e Cobertura foi feita, pois se tratam de propriedades já acompanhadas na Chaordic. Além disso, essas são propriedades de grande importância e recorrentemente estudadas na literatura disponível sobre avaliação de sistemas de recomendação, o que dá uma base conceitual para as primeiras investigações.

[13] Conferência internacional para estudos no tema de Sistemas de Recomendação.

A escolha das outras duas propriedades, Diversidade e Novidade, deram-se principalmente por se tratarem de aspectos desejados para os sistemas. Existem muitas críticas feitas aos sistemas de recomendação em que se apontam a possibilidade de se fechar o ciclo de interações de um usuário dentro “bolhas”, no sentido de que o sistema de recomendação só promove contato com itens que apresentem características já conhecidas. Para avaliar este possível efeito e sua intensidade sobre os nossos sistemas foram escolhidas estas propriedades.

Além disso, existem aspectos nos algoritmos usados pela Chaordic que se propõem exatamente a atenuar ou acentuar efeitos de itens populares e diversidade dentro das listas fornecidas, sendo assim indicadores capazes de representar esses aspectos se mostraram interessantes.

Outra vantagem das quatro propriedades citadas é o fato de que elas apresentam possíveis relações inversas de desempenho entre si, o que dá possibilidades interessantes de análise e exploração no sentido de conhecer características diversas dos produtos da Chaordic.

A lista completa dos indicadores, bem como sua breve descrição e referência é mostrada na tabela do apêndice A. Na seção 4.4 discutiremos quais indicadores de desempenho foram escolhidos para caracterizar as propriedades, além de discutir detalhes de implementação e comportamentos esperados.

## 4.4. Concretização dos Indicadores

Nesta seção discute-se a implementação dos indicadores considerados para o protótipo da ferramenta de experimentação. A discussão será feita a partir de cada propriedade escolhida como foco do projeto.

### 4.4.1. Assertividade

A propriedade de assertividade é a mais explorada no contexto de avaliação de sistemas de recomendação.

No contexto de recomendações com foco no item, para o caso do algoritmo e do sistema explorado neste projeto foram escolhidos dois indicadores: Recall, ou taxa de acerto e Precisão, Estes termos fazem referência à dois indicadores

referenciados na literatura, porém o método de cálculo usado no protótipo foi implementado com ligeiras diferenças, mantendo, entretanto a ideia de avaliação do algoritmo [11].

#### 4.4.1.1. Taxa de Acerto (Recall) e Precisão

O objetivo dos indicadores Recall e Precisão é mensurar a capacidade do sistema em fazer recomendações que serão usadas pelo usuário. Por usadas entendem-se, no contexto de comércio eletrônico, as recomendações que futuramente se tornarão cliques e/ou compras. A ação do sistema de recomendação pode ser classificada em um dos quatro casos descritos a seguir:

- Verdadeiro Positivo (VP): O item recomendado foi consumido;
- Verdadeiro Negativo (VN): A omissão da recomendação foi acertada;
- Falso Positivo (FP): O item recomendado não foi consumido;
- Falso Negativo (FN): Um item consumido não foi recomendado.

Estas quatro possibilidades podem são sumarizadas na matriz de confusão mostrada na tabela 1.

| | **RECOMENDAÇÃO** | | |
|---|---|---|---|
| **INTERAÇÃO** | | SIM | NÃO |
| | SIM | VP | FN |
| | NÃO | FP | VN |

Tabela 1 Matriz de Confusão – Recall e Precisão

Os indicadores de Recall e Precisão são calculados a partir das relações entre as frequências com que cada caso ocorre. As equações 5 e 6 apresentam as relações usadas para o cálculo dos indicadores.

$$Recall = \frac{VP}{VP+FN} \quad \text{(Eq. 5)}$$

$$Precisão = \frac{VP}{VP+FP} \quad \text{(Eq. 6)}$$

O que diferencia os indicadores apresentados na literatura [11] e os considerados para o framework é a definição do que é considerado caso positivo ou negativo de recomendação. A ideia básica é, para cada um dos usuários com interações dentro de um período, são listadas todas as interações, e para cada interação é verificado se existem recomendações disponíveis. O algoritmo que calcula as frequências usadas para o cálculo dos algoritmos é mostrado no quadro 1.

```
for each user in user_interaction_map
        interacted_items <- user_interaction_map.get_items(user)
        for each item in interacted_items
                hid_item <- item
                for each other_item in (interacted_items – hid_item)
                        if other_item has recommendation
                                if hid_item is in recommendations.get_recs(other_item)
                                        true_positive ++
                                else
                                        false_positive ++
                        else
                                false_negative ++

recall = true_positive / (true_positive + false_negative)
precisao = true_positive / (true_positive + false_positive)
```

Quadro 1 Pseudo-código para cálculo do Recall e Precisão

Esses dois indicadores são influenciados significativamente pelos valores de cobertura do sistema. Isso acontece, pois, em casos com maior cobertura tamanho médio das listas de recomendação aumenta. Com listas maiores, mais recomendações são feitas e, portanto existe maior chance de um item

recomendado estar entre as interações do usuário. Uma opção para amenizar este efeito é normalizar a pontuação de item avaliado pelo tamanho da lista de recomendações referente ao item, porém esta normalização não foi incluída nos cálculos do protótipo, pois alterava significativamente a sensibilidade dos indicadores, e como a resolução dos mesmos já é naturalmente pequena optou-se por evitar a normalização com o intuito de identificar diferenças mais sutis. Em alguns casos a normalização é necessária, porém, para os casos avaliados neste projeto as análises possuem valores de cobertura coincidentes ou extremamente próximos, o que torna possível o uso destes indicadores.

#### 4.4.1.2. Recall com Consideração de Ranking

Para o projeto foi ainda proposta uma variação do algoritmo de cálculo do Recall, onde foi inserido o conceito de posição dentro da lista de recomendações de um item para mudar a intensidade da pontuação de um acerto. No Recall tradicional quando ocorre acerto de recomendação a pontuação atribuída recebe valor 1, ou seja, a contagem direta do número de acertos.

A proposta desta variação com consideração de posição dentro da lista de recomendação (ou o ranking do produto recomendado dentro da lista) é a de pontuar com menor intensidade itens recomendados acertadamente e que aparecem em posições inferiores da lista [3].

Esta análise é interessante, pois o pacote de recomendações atribuídas à um determinado item não é apresentado de uma só vez. As recomendações são apresentadas em bateladas de, em média, quatro itens, e o usuário opta por navegar pelo restante do conteúdo recomendado. Os quatro itens mostrados a princípio são os itens com maior similaridade ao item referência, e a posição do item na lista cai quando ocorre empate na similaridade ou conforme a similaridade decresce. Sendo assim, um item recomendado na ultima posição de uma lista tem menor probabilidade de ser visto do que o primeiro item da lista.

Ao invés de ser constante e igual a 1, a pontuação para esta variação do Recall segue decaimento exponencial, conforme a posição do item se aproxima do final da lista. A pontuação máxima dada a um item é 1, quando ele está no

início da lista, e a pontuação mínima dada depende dos parâmetros da função de decaimento exponencial usada. Para os indicadores usados neste projeto considera-se uma função exponencial que atribui valor 0.5 ao item do final da lista.

Esta consideração de posição dentro da lista de recomendação nos permite medir efeitos de embaralhamento dos itens em determinadas situações.

### 4.4.2. Cobertura

Toda recomendação é composta por item ao qual se deseja fazer recomendações, chamado de item de referência, compondo assim o grupo de referência, e listas de itens propriamente recomendados, compondo assim o grupo de itens recomendados. Os indicadores de cobertura tendem relacionar o quanto do universo de itens estudados está coberto por um destes dois grupos.

Para cada decisão feita a respeito do universo de itens a ser analisado e do grupo de recomendações a ser relacionado temos diferentes possibilidades de análises. De fato, as combinações de conjuntos possibilitam a análise da cobertura de um sistema a partir de diferentes perspectivas.

Para o framework a proposta foi estudar apenas a cobertura relacionada ao grupo de referência das recomendações, e para isto foi proposto um indicador chamada Cobertura de Produtos Referência.

#### 4.4.2.1. Cobertura de Produtos Referência

Este indicador representa a porcentagem de produtos de um dado universo que possuem recomendações. Esta medida dá uma ideia importante do quanto de oportunidade o sistema está dando para que os itens de recomendação sejam vistos.

Para a geração deste indicador, consideram-se vários aspectos. O primeiro deles diz respeito ao universo de itens a serem usados como itens de referência. No caso do framework em questão consideram-se três possiblidades de constituição do universo de itens: todo o catálogo de produtos do cliente, todos os itens comprados dentro de certo período e todos os produtos visualizados dentro de certo período. As análises feitas para cada um

desses universos deu origem a três indicadores com significados distintos, mas que dividem a mesma ideia básica.

O quadro 2 mostra o pseudo-algoritmo usado para o cálculo do indicador.

```
total_products = products.size
for each item in products
        if item has recommendation
                items_with_recommendations ++

oid_coverage = items_with_recommendations / total_products
```

Quadro 2 Pseudo-código para cálculo da Cobertura de Itens de Referência

Para a análise deste grupo de indicadores considera-se ainda a filtragem dos dados de entrada. Normalmente os sistemas de recomendação comerciais apresentam número mínimo de recomendações que um produto deve oferecer para que as recomendações sejam efetivamente mostradas. No caso da Chaordic, que apresenta suas recomendações de “Similar Items” em vitrines com quatro espaços iniciais, se um produto oferecer duas ou menos recomendações então suas recomendações não são mostradas. Acredita-se que o grande espaço vago no layout pode ser um fator negativo para a experiência do usuário. Sendo assim, a análise utilizando filtros tem grande significância.

### 4.4.3. Diversidade

A diversidade em sistemas de recomendação está associada à capacidade do mesmo em compor listas de recomendações com itens que apresentem características distintas entre si. O critério usado para distinguir as características dos itens deve ser determinado como critério de avaliação. A seção 4.4.3.1 apresenta uma proposta de indicador para calcular a diversidade de uma lista de recomendações e nela são discutidos dois critérios para consideração de itens diversos.

#### 4.4.3.1. Diversidade Intra-lista

No caso deste projeto, as listas de recomendação serão geradas usando a abordagem de composição de recomendações usadas em um dos produtos atuais da Chaordic. Para cada usuário temos a lista de interações deste com itens do sistema, onde essas interações podem ser de visualizações ou compras. Cada item interagido apresenta então um conjunto de recomendações próprio, onde cada item tem atrelado a si um grau de similaridade, que é normalizado dentro do conjunto de recomendações de um item de referência. A composição da lista de itens a serem recomendados para um usuário é feita pela ordenação, através do nível de similaridade normalizado, de todos os itens que são recomendações para os itens interagidos. Desta lista, filtram-se os primeiro N itens, que possuem os maiores níveis de similaridade, o que compões o que iremos chamar de top N por usuário [4].

Este indicador tenta representar, para um conjunto de top N, o quão diversas são as recomendações feitas. Para analisar a diversidade dentro de um conjunto top N de recomendações é necessário determinar qual será o critério de diversidade. A aplicação do critério de diversidade é feita através de uma função que analisa os itens da lista de recomendações e retorna uma pontuação relacionada a lista. Essa pontuação é então normalizada pela quantidade de itens recomendados na lista, que pode ser no máximo N e no mínimo um. Para cada usuário teremos então uma lista e uma pontuação associada. Os indicadores finais são compostos pela média destas pontuações em relação ao total de usuário avaliados [14]. O quadro 3 mostra o pseudo-algoritmo usado para calcular o indicador.

```
N <- 10
score <- 0
for each user in user_interactions
        top_N <- user.get_top_recommendations(N)
        score ++ diversity_criterium_score(top_N)

intra_list_diversity <- score / user_interactions.count_users()
```

Quadro 3 **pseudo-algoritmo** para cálculo do indicador Diversidade Intra-lista

A função de diversidade é referenciada no pseudo-algoritmo como “diversity_criterium_score”, e para cada lista de recomendação (top N) atribui uma pontuação, de acordo com o critério escolhido. Para o desenvolvimento do protótipo escolheu-se a aplicação de duas funções de avaliação de diversidade:

- Categoria: contagem do número de categorias diferentes existentes na lista de recomendação.
- Item referência: relacionada ao produto referência do produto recomendado, ou seja, o produto interagido pelo usuário para o qual a recomendação foi feita. O critério é a contagem do número de referências diferentes na lista de recomendação.

### 4.4.4. Novidade

A propriedade novidade está relacionada à capacidade do sistema de recomendação em explorar itens com menor obviedade para o usuário. Existem sistemas de recomendação com objetivo inerente de auxiliar na exploração de algum ambiente, como é o caso de sistemas de exploração de música fornecidos por plataformas como o Grooveshark e o Last.fm.

Além do aspecto positivo que a surpresa da recomendação de um item pode trazer, existe o fato de que sistemas de recomendação não têm como objetivo mostrar itens que já possuem visibilidade grande em campanhas de divulgação ou cuja popularidade seja historicamente expressiva. Os efeitos de recomendar frequentemente itens muito populares pode causar estranhamento

ao usuário e diminuir sua confiança sobre a capacidade do sistema em apresentar personalização.

Esta propriedade ajuda a medir estes aspectos e apresenta informações bem distintas das informações apresentadas por outras propriedades exploradas até então. O indicador desenvolvido para avaliar esta propriedade é a Entropia da Informação, conceito emprestado da estatística [13] e tem como base a probabilidade dos itens serem apresentados ou consumidos em um determinado domínio.

#### 4.4.4.1. Entropia da Informação –Interações

Este indicador baseia-se no conceito da entropia de um sistema. O conceito de entropia em sistemas de informação está relacionado à quantidade de informação contida em um sistema e expressa também como essa informação está espalhada. Este conceito mostra a informação de um sistema como uma medida probabilística [13].

O que queremos medir com este indicador é a capacidade do sistema de recomendação em explorar os produtos ao longo da cauda de distribuição. Como dito na seção anterior, espera-se que um sistema de recomendação de comércio eletrônico ajude a exploração dos produtos distribuídos ao longo da cauda de distribuição, e não somente produtos da parte mais populosa da cauda.

O indicador proposto é calculado da seguinte forma: para cada produto recomendado, calcula-se a relação entre o numero de interações para o produto e o numero de usuários total, gerando assim a probabilidade de um produto ter interação. Esta probabilidade é mostrada pela equação 7, onde I(i) é o número de interações feitas com o produto i, e T é o número total de interações.

$$p_i = \frac{I(i)}{T} \qquad \text{(Eq. 7)}$$

A seguir, para cada produto calcula-se o -log10(p), onde p e a probabilidade associada ao produto. Esses valores são multiplicados pela própria probabilidade, somados e normalizados pelo logaritmo do número de produtos

recomendados. A relação usada para o cálculo do valor da entropia é apresentado na equação 8, onde N é o número total de itens e p são as probabilidades associadas à cada um deles.

$$E(x) = -\sum_{i=0}^{N} p_i \log(p_i) \quad \text{(Eq. 8)}$$

O indicador calculado pela equação 8 ainda passa por uma normalização pelo logaritmo do número de produtos, resultando na expressão final do indicador mostrada na equação 9. Esta normalização é necessária para possibilitar a comparação de casos com número de produtos recomendados diferentes.

$$E(x) = \frac{-\sum_{i=0}^{N} p_i \log(p_i)}{\log N} \quad \text{(Eq. 9)}$$

As figuras 2, 3 e 4 apresentam algumas configurações de distribuição de interações com produtos, visando simular casos extremos e intermediários de distribuições possíveis. Para cada caso apresentado nas figuras considerou-se o mesmo número de interações total, igual a 55. O que se variou foi a distribuição das interações entre os itens. A figura 2 apresenta o caso onde as interações estão concentradas nos produtos do começo da cauda. A figura 4 apresenta uma distribuição mais próxima da uniforme. E por fim, a figura 3 apresenta uma distribuição intermediária.

Espera-se que o sistema de recomendação que promova novidade explore itens do final da cauda, sendo assim, a distribuição de interações dos produtos recomendados vai se aproximar da distribuição mostrada na figura 4 e se distanciar da figura 2. Os valores da entropia calculados para cada uma destas distribuições é apresentado juntamente com o gráfico. Podemos perceber um aumento no valor do indicador conforme a distribuição de interações dos itens recomendados se aproxima de uma distribuição uniforme, que é o caso onde há menos privilégio de itens muito populares.

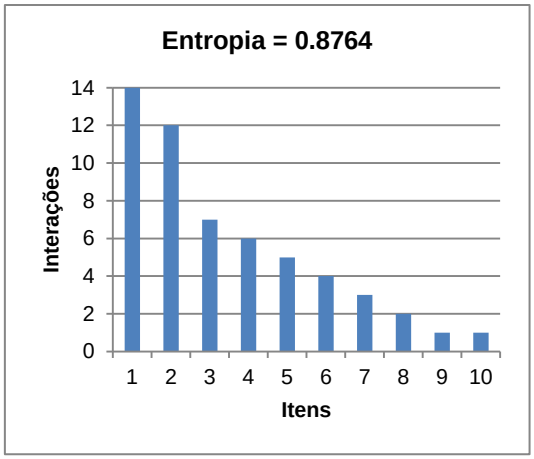


Figura 2 Entropia – Distribuição Cauda Longa

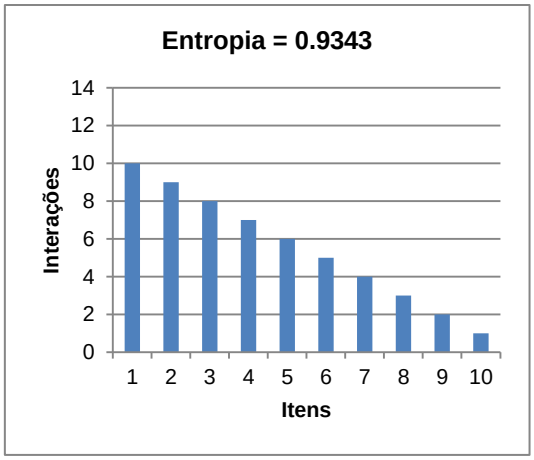


Figura 3 Entropia – Distribuição Intermediária

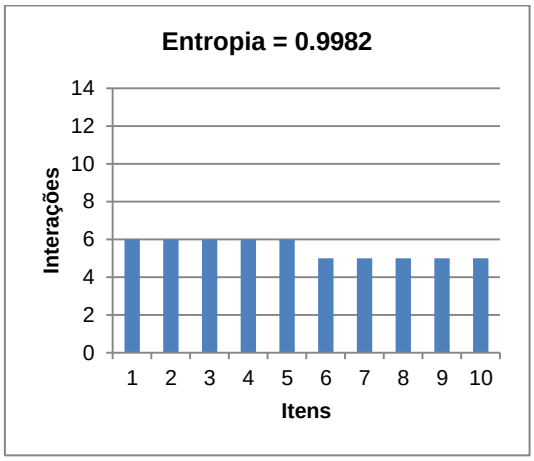


Figura 4 Entropia - Distribuição Uniforme

# 5. Planejamento dos Estudos de Caso

## 5.1. Estudo de caso I: Estimativa de Tempo Necessário para Integração

Durante o período de desenvolvimento do projeto na empresa, o time de pesquisa e desenvolvimento recebeu questionamentos do time comercial sobre estimativas de tempo necessário para que um cliente que esteja entrando no mercado consiga receber boas recomendações. Na época não havia uma resposta para o assunto, mas percebeu-se a importância desta definição, pois é informação importante para auxiliar o processo de vendas. Com o objetivo de responder esta pergunta, ou ao menos traçar uma estimativa, este estudo de caso foi proposto.

Ao entrar no mercado clientes novos não apresentam informações de interações, e consequentemente não é possível prover recomendações aos mesmos com os mecanismos de geração de recomendações atuais da Chaordic, baseados em filtragem colaborativa. Com o passar do tempo de início de funcionamento as informações de interações vão sendo coletadas, porém a qualidade das recomendações geradas só atinge patamar suficiente para serem apresentadas após um tempo determinado. O objetivo deste experimento é estimar aproximadamente qual este período de tempo necessário.

Para responder ao questionamento principal, algumas questões precisam ser respondidas.

A primeira delas está relacionada a definir o que é uma recomendação com qualidade razoável. Durante o processo de integração com novos clientes a Chaordic realiza o processo chamado internamente de “tunagem”, no qual as recomendações geradas passam por avaliações empíricas e qualitativas, realizadas pelos próprios integrantes do time de gerenciamento de recomendações, e tem o intuito de ajudar a determinar os limiares de parâmetros dos algoritmos e para determinar se o conjunto de recomendações fornecidas parece coerente. Este processo apesar de altamente manual vem trazendo bons resultados no que diz respeito à determinação de conjuntos de

recomendações que não apresentem grandes surpresas negativas para os representantes dos e-commerces clientes.

A segunda questão importante é como garantir que a estimativa feita se adeque aos diversos novos clientes os quais o time comercial da Chaordic vai se comunicar. Neste sentido, tomou-se a decisão de limitar o escopo do experimento considerando apenas empresas de comércio eletrônico, com negócios do tipo *Business to Consumer* (B2C)[14] e que vendam bens de consumo duráveis. Além disso, para a escolha dos clientes a serem usados como referência no estudo de caso é realizada a caracterização de alguns dos atuais clientes da empresa. Estes foram escolhidos de maneira a contemplar clientes de médio e grande porte, para garantir assim que as estimativas feitas se aproximem de estimativas válidas para clientes com perfis intermediários.

A caracterização dos clientes é construída pelo levantamento de algumas informações que descrevam questões importantes sobre o negócio, considerando informações relacionadas ao catálogo de produtos, fluxo de interações, usuários e perfis de compra. Na lista abaixo temos a relação das informações a serem levantadas para a caracterização:

- Número de produtos disponíveis em catálogo;
- Número de compras por mês;
- Número de produtos distintos comprados por mês;
- Número de produtos visitados por mês;
- Número de produtos distintos visualizados por mês;
- Faturamento total por mês;
- Total de pedidos realizados: Por pedidos entende-se a compra simultânea de um conjunto de produtos;
- Tamanho médio dos pedidos (TMP): Quantidade média de produtos contidos em todos os pedidos considerados;
- Valor médio dos pedidos (VMP): Valor médio dos produtos contidos em todos os pedidos considerados;

[14] "Business to Consumer" (B2C) é uma modalidade de negócio onde as transações são feitas entre empresa e o consumidor final. Contrapõem-se com outras modalidades como "Business to Business" (B2B) , onde transações são feitas entre empresas, e "Consumer to Consumer" (C2C), onde as transações são feitas entre usuários finais, que revezam nos papéis de ofertante e consumidor.

- Ticket médio dos produtos: Valor médio dos produtos contidos no catálogo de vendas.

Este experimento irá fornecer insumos para o time comercial da Chaordic realizar contato e negócios com potenciais clientes que estejam entrando no ramo do comércio eletrônico. A estimativa feita não tem a ambição de alta exatidão, porém deve ser bem justificada a ponto de servir como argumento comercial.

A partir do que foi apresentado e discutido até o momento, definiu-se a seguinte metodologia de trabalho para a realização do experimento:

1. Caracterização de dois clientes da Chaordic que apresentem portes diferentes. A caracterização foi feita pela determinação das características listadas nesta seção e pelos indicadores off-line, para os clientes escolhidos como referência. Os períodos para cálculo dos indicadores foram os três primeiros meses de 2012.
2. Calcular os mesmos indicadores off-line usados para avaliar a situação de referência, para os mesmos clientes, o mesmo algoritmo e os mesmos parâmetros, mas com recomendações geradas tendo como dados de entrada diferentes conjuntos de dados. Os conjuntos de dados brutos a serem usados para a geração das recomendações serão dados de compras variando de um período de 15 dias até um período de 180 dias.
3. Analisar os resultados dos indicadores off-line calculados na etapa três e confrontá-los com os indicadores de referência calculados na etapa dois. Através desta análise deseja-se determinar um ponto em que os indicadores se aproximam o suficiente para que em uma avaliação manual e qualitativa as recomendações geradas sejam satisfatórias.
4. Por fim, a partir das análises dos indicadores e dos clientes, determinar para cada cliente um número de dias necessários para que possa haver geração de recomendações.

## 5.2. Estudo de caso II: Relação entre Diversidade, Novidade e Penalização de Itens Populares

Como descrito na seção 3, o principal algoritmo usados pela Chaordic hoje é o algoritmo chamado “Similar Items”, usado para gerar recomendações para 6 dos 9 produtos existentes hoje na plataforma web de recomendações. Este algoritmo possui parâmetros que possibilitam a modificação das características dos conjuntos de recomendação gerados, seja através da adaptação do cálculo das similaridades ou da utilização de filtros sobre as entradas e saídas do sistema. Um destes parâmetros é utilizado para a penalização de itens com alta popularidade.

Para algoritmos baseados em filtragem colaborativa, itens com alta popularidade tendem a aparecer com frequência muito grande nas listas de recomendações, justamente por apresentarem um alto número de interações. Graças a este alto número de interações, a probabilidade de um item muito popular ter sido comprado em conjunto com qualquer outro item do sistema é mais elevada que o normal, o que se torna um reforço ao algoritmo de filtragem colaborativa, que o considera muito semelhante a uma gama grande de outros itens. Este efeito é conhecido na literatura de sistemas de recomendação como efeito “Harry Potter” ou efeito “Amelie Poulain”, justamente pelos problemas acarretados pelos mesmos.

Na Chaordic, a determinação dos valores do parâmetro de penalização de itens muito populares é feita tomando como base apenas algumas avaliações de assertividade do sistema, respaldadas pelos indicadores de Recall e Cobertura, apresentados anteriormente. Quando este parâmetro foi incluído no algoritmo, um estudo foi feito e determinou-se uma faixa de valores para os quais as recomendações pareciam apresentar boa qualidade, sendo esta qualidade medida através de avaliações de pessoas sobre algumas poucas recomendações geradas. A partir de então este parâmetro teve sempre seu valor ajustado ao redor desta faixa de valores e muito pouco foi feito para reavaliar seus efeitos e eficiência, e justamente isso confere a motivação e importância da realização deste estudo de caso.

O objetivo deste experimento é prover um estudo quantitativo sobre os efeitos deste parâmetro sobre as recomendações geradas pelo sistema, além

de estimar um valor, ou faixa de valores, do parâmetro de penalização que pareçam indicar o cumprimento de sua responsabilidade, que é penalizar itens com alta popularidade, mas sem comprometer a assertividade do sistema de recomendação.

O método utilizado foi a avaliação das recomendações através de duas propriedades: assertividade e diversidade. Entende-se que o intuito de penalizar recomendações de itens muito populares é evitar que estas recomendações sempre se sobressaiam em relação às recomendações de itens regulares, porém, sem que isso comprometa a habilidade do sistema em fazer recomendações úteis.

Da maneira como o parâmetro de penalização é incluído no algoritmo da Chaordic, tem-se um efeito inversamente proporcional à similaridade. A penalização desloca os itens com alta popularidade para o final da lista de recomendações. Em teoria, se o tamanho das listas de recomendação geradas fosse infinito, a penalização não causaria a exclusão do item. Porém, como listas infinitas não são possíveis, e existem questões práticas de armazenamento e desempenho dos cálculos, o tamanho das listas é fixado, e por este efeito pode ser que a penalização retire um item da lista. O item é retirado caso seja deslocado para uma posição inferior ao limite da lista.

Para a avaliação do efeito de distribuição dos itens mais populares são usadas listas de recomendação construídas para um usuário. Os principais indicadores relacionados à mensuração dos efeitos esperados são os indicadores de assertividade: Recall, Recall com consideração de Ranking e Precisão, apresentados nas seções 4.4.1.1 e 4.4.1.2, e os indicadores de diversidade e novidade: Diversidade Intra-lista e Entropia da Informação.

Merecem destaques maiores os indicadores Recall com consideração de Ranking e o indicador Entropia da Informação. O indicador Recall com consideração de Ranking é sensível ao deslocamento dos itens dentro da lista, e por isso mostra-se um indicador capaz de captar os deslocamentos causados pelo parâmetro de penalização. O indicador Entropia da Informação é calculado com base na probabilidade de um item sofrer interação, e leva em conta o número total de interações do item recomendado e o número total de usuários no sistema para o cálculo desta probabilidade. Este indicador representa quase

em relação direta a presença ou ausência de itens populares nas listas geradas pelo sistema.

Com base na ideia de avaliação de listas, admite-se que a diversidade das listas é diretamente afetada pela existência ou não de penalização. O fato dos itens mais populares aparecerem com mais frequência acaba acarretando em menor diversidade das listas, que vão ter em muitos casos os mesmos itens. Por outro lado, do ponto de vista da assertividade, recomendar itens muito populares garante alta assertividade do sistema, uma vez que a probabilidade de que um item com muitas interações de compra seja comprado por qualquer outro usuário é maior do que a probabilidade de um item com poucas interações ser comprada pelos mesmos usuários.

A avaliação será feita de maneira a entender esta relação inversa entre as propriedades, e a evolução dos indicadores conforme os efeitos de penalização variam. A hipótese é de que a partir de certo valor de penalização a variação da diversidade e similaridade comece a se estabilizar, indicando que a partir dali os efeitos não podem mais ser controlados pela variação do parâmetro. O resultado do experimento serve de base para um método de determinação de valor para o parâmetro de penalização de maneira a respeitar o compromisso entre assertividade do sistema e qualidade da lista de recomendação.

A partir do que foi discutido até então e tendo em vista o objetivo do experimento, os seguintes passos foram realizados:

1. Levantar os valores dos indicadores para faixas crescentes de intervalos de tempo. Mostrar com a evolução dos indicadores o número de dias necessários para que se atinja um regime de estabilidade, ou seja, para que o valor dos principais indicadores mostre que o sistema não varia significativamente conforme o período considerado aumenta. O objetivo desta etapa é determinar um intervalo de tempo suficiente para considerar as recomendações estando no que podemos chamar de regime permanente.
2. Para o valor de intervalo de tempo levantado na etapa anterior, calcular vários conjuntos de recomendação variando o parâmetro de penalização, desde o valor nulo de penalização até um valor onde para os indicadores relacionados à diversidade das recomendações não sofram variações significativas.

3. Encontrando este ponto, calcular e relacionar os indicadores referentes à assertividade, diversidade e novidade, para os diversos conjuntos de recomendação gerados na etapa anterior.

4. Analisar o resultado e propor maneira de determinar o parâmetro de penalização a partir da relação da evolução da diversidade e da assertividade conforme a penalização aumenta.

Espera-se ao final da análise determinar se os valores atuais do parâmetro estão coerentes, se os fatores importantes para sua escolha estão sendo considerados. Como outro ganho relacionado ao experimento, e que é conciliado diretamente com o objetivo do projeto, deseja-se comprovar a utilidade dos conhecimentos adquiridos em avaliação de sistemas e das ferramentas construídas para o auxílio na resolução de questões importantes no contexto da empresa.

# 6. Resultados

## 6.1. Estudo de Caso I: Discussão e Resultados

Para a realização do estudo de caso I considerou-se a caracterização de cinco clientes que potencialmente apresentariam perfis distintos. Por motivos de confidencialidade iremos referenciar cada um dos clientes estudados por um identificador numérico. As informações levantadas foram as previstas na descrição do caso de uso, feita na seção 5.1.

Como previsto no planejamento do experimento, para cada cliente foram levantadas as seguintes informações levando em consideração os meses do primeiro quarto do ano de 2012.

Após o levantamento dos dados de caracterização dos clientes, realizou-se uma análise para determinar quais clientes apresentavam perfis similares e quais apresentavam perfis diversos, lembrando que o objetivo desta etapa de caracterização é fundamentar a escolha dos clientes que serviram como objetos de estudo do experimento.

Por questões de confidencialidade dos dados, todos os gráficos que apresentam informações sobre os clientes foram ajustados para uma escala relativa. Todos têm seus valores apresentados sempre em relação ao cliente número 4.

A figura 5 apresenta o ticket médio dos produtos de catálogo de cada um dos clientes considerados.

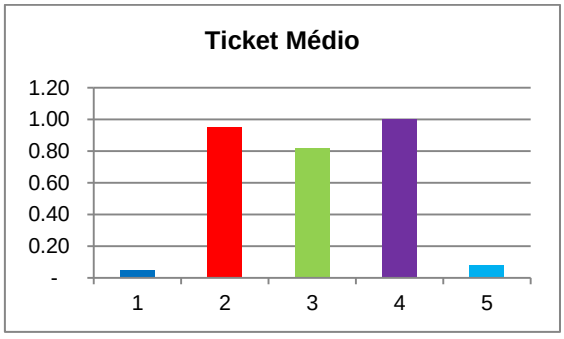


Figura 5 Ticket Médio

As figuras 6, 7 e 8 apresentam os resultados relacionados à navegação e interações explícitas dos usuários com o sistema. Estes dados são os mais representativos no sentido de caracterizar o cliente, pois mostra não só informações de desempenho dos mesmos em indicadores financeiros e de conversão, mas mostra também indícios da popularidade dos mesmos dentro do contexto do comércio eletrônico nacional.

Pela análise dos gráficos podemos ver que os clientes 2, 3 e 4 apresentam sempre comportamentos muito próximos, permutando entre si nas posições mais elevadas, para as três variáveis: visualizações, compras e faturamento. De fato, esse efeito era esperado, pois são clientes com perfis sabidamente muito parecidos, que possuem a maior parte do catálogo comum e que, apesar de assim como os outros clientes também constituírem-se de plataformas de comércio eletrônico, atendem aos mesmos nichos de mercado.

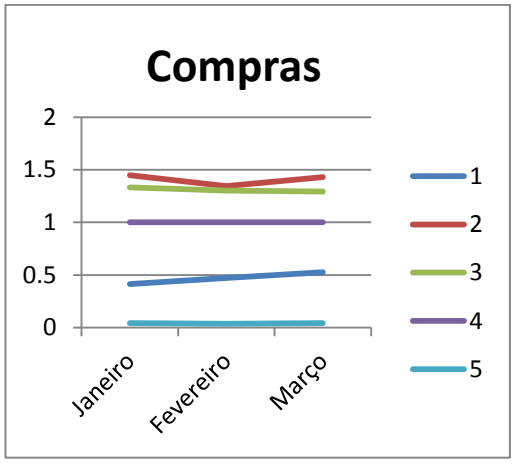


Figura 6 Compras por período

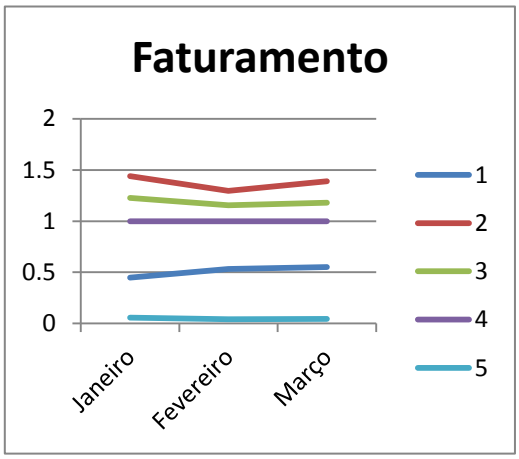


Figura 7 Faturamento por período

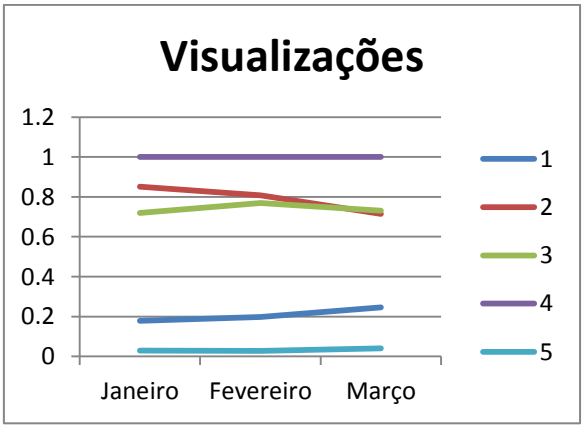


Figura 8 Visualizações por período

Outro efeito óbvio é o posicionamento do cliente 5 nos patamares inferiores, o que é corroborado pela análise de outras variáveis levantadas na caracterização.

Os gráficos 9 e 10 apresentam as informações de tamanho médio dos pedidos e valor médio dos pedidos, onde por pedido entende-se um conjunto de itens comprados no mesmo momento por um usuário. Ao contrário dos indicadores de visualizações, vendas e faturamento, que tendem a mostrar o comportamento geral fluxo de interações ao redor do site, os indicadores relacionados à pedidos fornecem informações a respeito do comportamento da unidade usuário dentro do e-commerce.

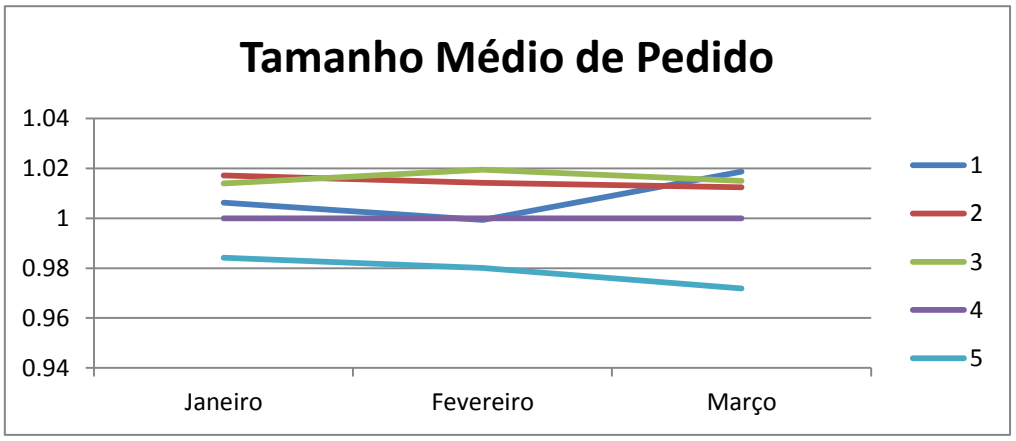


Figura 9 TMP – Tamanho Médio de Pedido

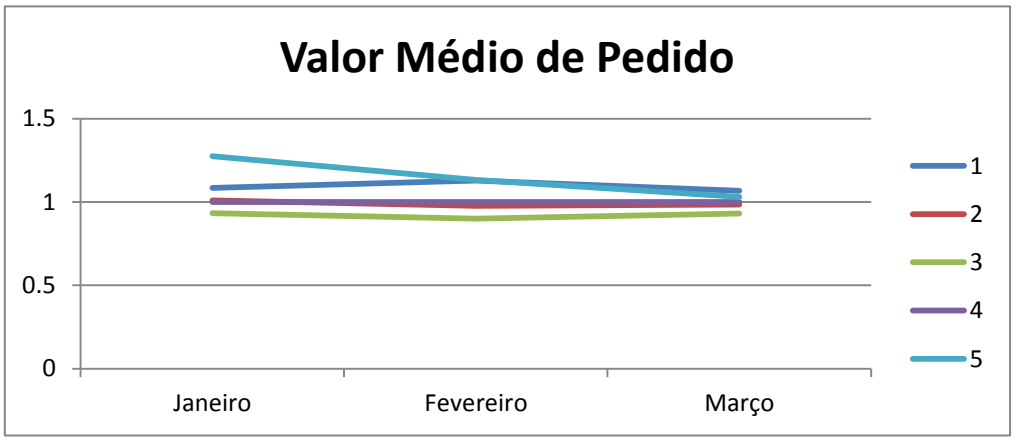


Figura 10 VMP – Valor Médio de Pedido

O cliente 5 se destaca neste contexto pois apresenta um valor médio por pedido em geral maior do que o dos outros clientes, o que indica que possui

um valor médio de produtos maior, visto que o tamanho médio dos pedidos e o fluxo de compras é menor. Apesar desta inversão no desempenho do indicador, o cliente 5 ainda se mantém nas extremidades das análises e por isso é preferido como referência para uma das bordas.

Por fim, temos o gráfico 11 apresentando dados relativos aos catálogos dos clientes, onde obviamente percebe-se a similaridade dos clientes 2, 3 e 4, como mencionado anteriormente, e uma distinta distância do cliente 5 em relação aos demais.

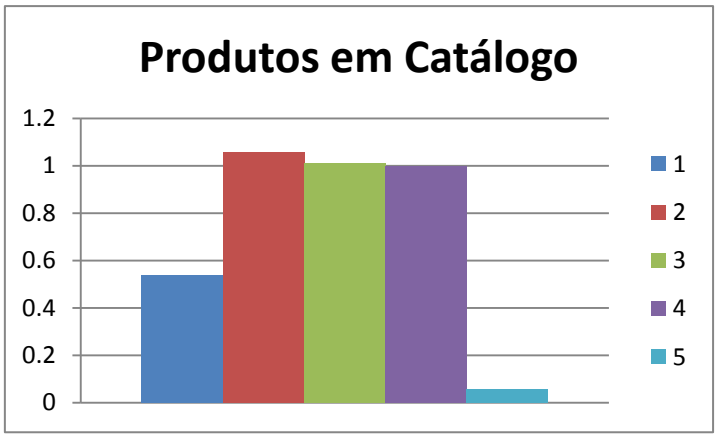


Figura 11 Produtos em Catálogo

Ainda com a intenção de escolher os clientes de referência para o experimento foram calculados os indicadores de desempenho relacionados às quatro principais propriedades escolhidas para este projeto. Foram geradas recomendações para cada um dos clientes usando dados de compras dos primeiros três meses de 2012. Os mesmos dados usados na geração das recomendações foram usados para o cálculo dos indicadores. O objetivo desta etapa era complementar a descrição dos clientes.

As figuras de 12 a 15 apresentam os resultados dos indicadores. Podemos perceber que em geral o desempenho dos indicadores acompanha o porte do cliente. Quanto mais dados usados para a geração das recomendações, mais informações existem para a geração das recomendações, e a tendência é de que isso se reflita positivamente no resultado dos indicadores.

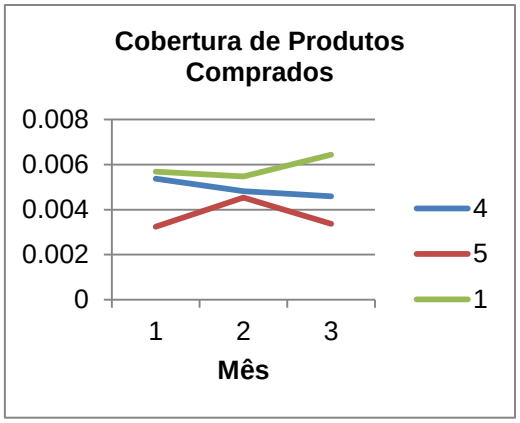


Figura 12 Cobertura Produtos Comprados

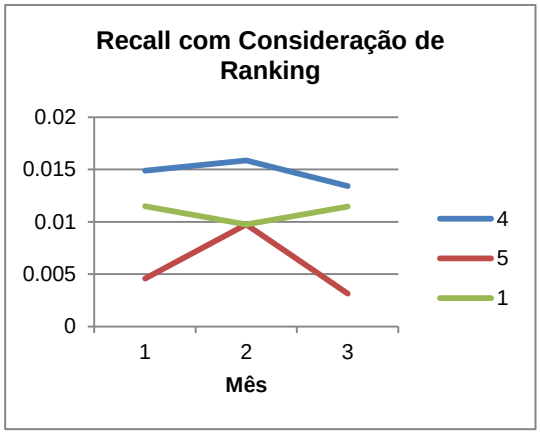


Figura 13 Recall considerando Ranking

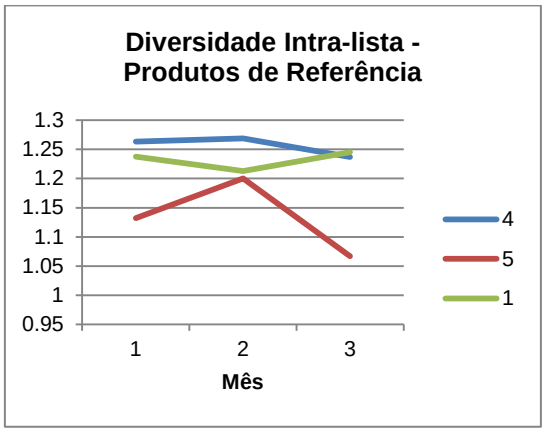


Figura 14 Diversidade Intra-lista por período

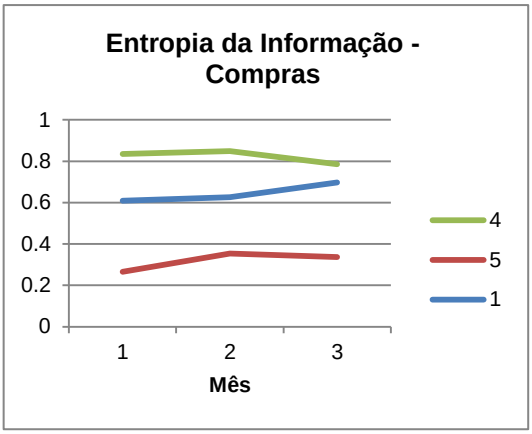


Figura 15 Entropia da Informação por período

Uma vez definidos e justificados os clientes de referência foram calculados os indicadores usando vários períodos diferentes e crescentes de informações de compras. Variou-se o período considerado entre 15 dias e 180 dias, com passos de 15 dias. Para cada período dentro deste limite foram geradas recomendações para cada um dos três clientes, e os dados de recomendações juntamente com os dados de compras usados para gerá-las foram usados para calcular os indicadores de desempenho implementados no protótipo.

Conforme o período de informações cresce, o desempenho indicado pelos indicadores se estabiliza. Esse efeito é mostrado nas figuras de 16 a 19.

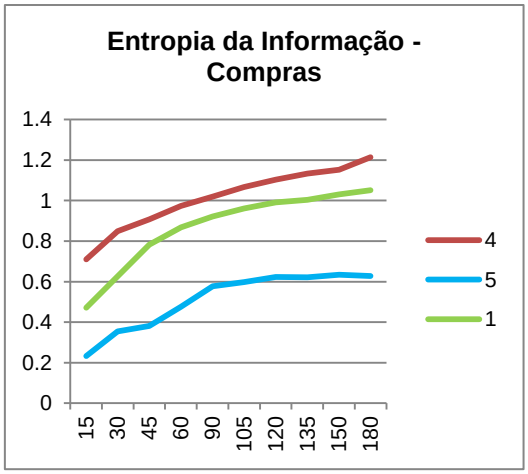


Figura 16 Entropia da Informação – Evolução de intervalo

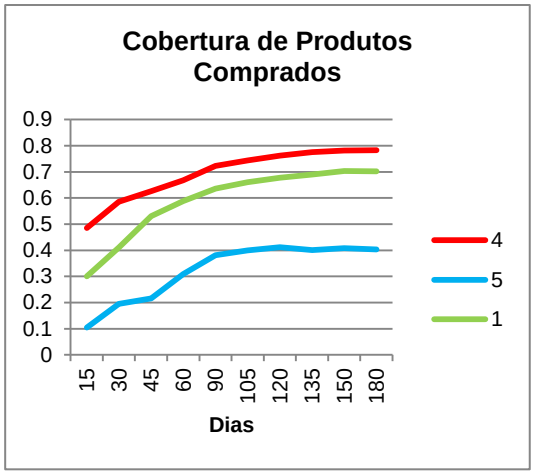


Figura 17 Cobertura de Produtos Comprados – Evolução de intervalo

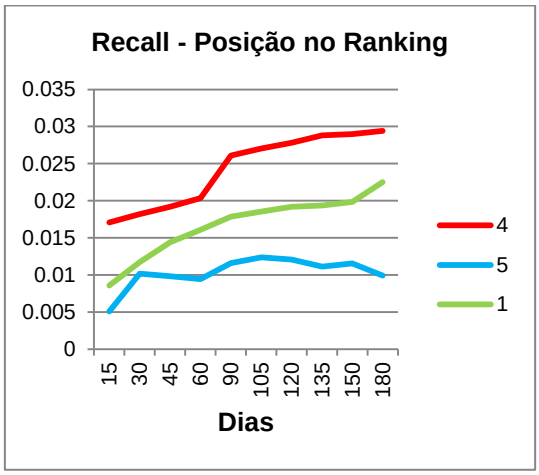


Figura 18 Recall considerando Ranking – Evolução de intervalo

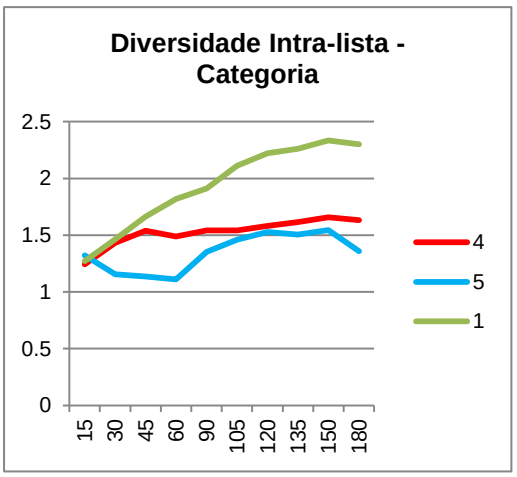


Figura 19 Diversidade Intra-lista – Evolução de intervalo

Analisando o comportamento dos clientes tomados como referências superior e inferior (clientes 4 e 5), percebe-se que o período de estabilidade dos indicadores começa a acontecer por volta da marca de 90 dias. Na faixa entre 90 e 120 dias a variação dos indicadores é consideravelmente menor do que a variação na faixa de 15 a 90 dias. Após 120 dias vemos que a cobertura apresenta pouca variação. Um comportamento similar acontece para o Recall. Apenas o indicador de diversidade apresenta um comportamento distinto.

A partir dos dados levantados chegou-se a conclusão que um período entre 90 e 120 dias parecem suficientes para que as recomendações feitas usando dados de compras atinjam estabilidade, considerando as condições descritas

neste estudo de caso. Verificamos que para o cliente intermediário, identificado por 1, os indicadores seguem uma dinâmica similar.

Sabe-se que esta estimativa é sujeita a falhas. É reconhecido que este estudo de caso foi baseado em clientes que já atuam no mercado há pelo menos 3 anos, para garantir uma estabilidade no número de visitações e de compras diárias, além de serem clientes com um público já cativado e que têm políticas de divulgação bem definidas. Todos esses fatores favorecem a qualidade dos dados de entrada usados para a geração das recomendações, e consequentemente influem na qualidade do produto final.

Em contrapartida, as estimativas feitas se basearam apenas em dados de compras para a geração e avaliação das recomendações. Sabe-se que os dados de compras, apesar de fornecerem informações mais fortes sobre a similaridade dos produtos, são menos volumosos que dados de visualizações. Algumas ações como a captação de dados de visualização e a mistura de dados de compras e visualização podem favorecer o fornecimento de recomendações em menor tempo a esses novos clientes.

Considerando que o objetivo do estudo era realizar uma estimativa mesmo que aproximada, mas pautada em análises quantitativas. E considerando que existem várias possibilidades para complementar as informações de compras de um cliente, entende-se que os valores sugeridos de 90 a 120 dias são razoáveis.

Como próximos passos sugere-se a realização de um experimento qualitativo de avaliação das recomendações geradas com 90 dias de compras. Além disso, este estudo será apresentado ao time comercial a fim de que críticas e questionamentos sejam feitos, de maneira que os argumentos aqui lançados possam vir a ser tornar de fato argumentos comerciais, concretizando assim o preenchimento de uma lacuna no leque de conhecimentos da empresa sobre seu próprio produto. A validação por parte do time comercial vai auxiliar a preencher as lacunas de argumentação que possivelmente existirão, e uma vez que o time considere suficiente a análise para ser usada como argumento comercial, a validação da estimativa estará realizada.

## 6.2. Estudo de caso II: Discussão e Resultados

Conforme descrito na seção 5.2, este experimento visa usar o conhecimento sobre avaliação de sistemas de recomendação e as ferramentas produzidas para auxiliar na determinação do parâmetro de penalização usado no principal algoritmo da Chaordic.

O primeiro passo descrito na metodologia foi a determinação das características dos clientes mais expressivos da empresa, com o objetivo de determinar o número de dias necessário para que as análises feitas atingissem condição próxima à estabilidade, ou seja, pouca variação.

Para esta escolha levantou-se o valor dos principais indicadores de desempenho, para vários intervalos de dados de compras. Os resultados do estudo de caso I foram aproveitados para esta estimativa. Variou-se o intervalo de 15 a 180 dias, com passo de 15 dias, onde para cada intervalo foram geradas recomendações e calculados os indicadores. As figuras de 16 a 19 (referencia à seção anterior) mostram o comportamento de quatro dos indicadores, sendo cada um associado a diferentes propriedades de avaliação.

A evolução dos indicadores indica que as características do sistema passam a se estabilizar para um período de 120 dias. No caso do indicador de diversidade, Diversidade Intra-lista, não houve mudanças significativas.

A partir destas análises escolheu-se o cliente identificado pelo número 1, por ser o cliente com maior fluxo de interações e maior número de informações de produtos. O período de 120 dias foi considerado com início em primeiro de janeiro de 2012 e com fim em primeiro de maio de 2012.

A seguir observou-se a evolução dos indicadores para diferentes valores de penalização. Historicamente tem-se usado o valor de 0.2 para o parâmetro de penalização. Este valor foi determinado através de experimentos com base a avaliação da assertividade do sistema pelo indicador de Recall e por avaliações qualitativas feitas pelos próprios envolvidos no experimento. A partir deste valor, decidiu-se experimentar neste estudo valores de penalização variando de 0 a 1, onde o valor 0 indica não penalização.

Para este intervalo, indicadores como a Precisão, Diversidade Intra-lista para Itens de Referência e indicadores de Cobertura em geral não foram significativamente alterados. Identificou-se pequena variação nos indicadores

de Diversidade Intra-lista para critério de categoria e para os indicadores de Recall. Estes efeitos de alterações podem ser vistos nas figuras 20 e 21, onde os valores percentuais indicam variação em relação ao ponto onde não há penalização. A alta sensibilidade destes indicadores para variações na penalização já eram esperadas, uma vez que os indicadores de Diversidade Intra-lista e o Recall com consideração de Ranking captam diferenças especificamente entre listas de recomendação.

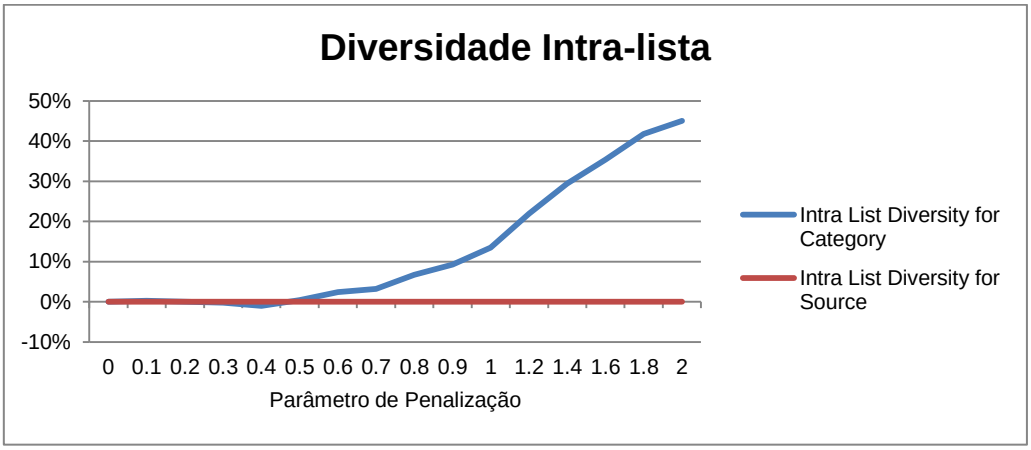


Figura 20 Diversidade Intra-lista – Evolução penalização

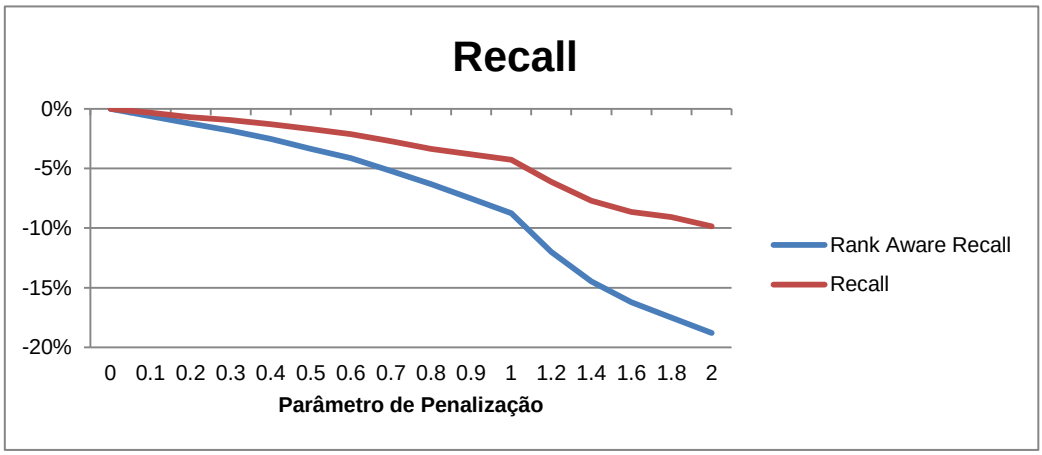


Figura 21 Recall – Evolução penalização

O efeito de baixa sensibilidade em alguns indicadores fez com que novas hipóteses fossem levantadas para explicá-los.

A primeira tentativa foi estender o intervalo de penalizações, indo de 0 até 2. Mesmo com o aumento do intervalo não foram sentidas diferenças significativas nos mesmos indicadores citados anteriormente. Os efeitos no

indicador de Recall com consideração de Ranking tiveram comportamentos próximos a uma curva exponencial com curvatura pouco acentuada, onde para penalização igual a 2 atingiu-se diferença próxima de 20% em relação a situação de não penalização. Para o indicador de Diversidade Intra-lista com critério de categoria a diferença foi bem mais significativa e a taxa de crescimento aproximou-se de um comportamento exponencial acentuado, atingindo para penalização igual a dois uma diferença próxima de 50% em relação a não penalização. O comportamento destes dois indicadores frente ao restante mostrou sua grande sensibilidade e, apesar de não serem os principais indicadores para mensurar penalização de mais populares, se mostraram bons indicadores para analisar diferenças dentro considerando listas de recomendação.

Outra característica investigada para justificar a não alteração do comportamento geral do sistema foram os filtros de frequência também presentes na composição das recomendações. Os filtros de frequência foram explicados na seção 3.2.1 e tratam-se basicamente de filtros que retiram itens da lista de recomendações de um item quando a frequência de compras mútuas é menor do que um determinado valor. Estes filtros são bem importantes, pois, como levantado anteriormente, o processo de inferência do sistema de recomendação usa todas as informações fornecidas, e em alguns casos o processo de inferência dá significado para informações que não são consideradas relevantes para o cálculo de similaridades.

Usando dados de 120 dias de compras para o cliente referência deste experimento temos que 88% das compras mútuas apresentam frequência igual a 1. Como a maior parte dos pares de itens com compras conjuntas apresenta esta frequência, o não uso de algum filtro torna praticamente impraticável a geração das recomendações.

A partir desta discussão decidiu-se explorar mais os efeitos dos filtros de frequência sobre variações do critério de penalização a fim de entender a ausência de sensibilidade de alguns dos indicadores para os quais se esperava perceber mudanças.

A primeira intenção foi avaliar o efeito da assertividade do sistema com diferentes filtros de frequência, porém isso não foi possível graças a um efeito já conhecido e discutido na seção 4.5.2, que consiste na influência da

cobertura sobre os indicadores de Recall e Precisão. Com o aumento do filtro de frequência a cobertura do sistema decresce rapidamente, principalmente quando ocorre a passagem da situação sem filtro e do uso de filtro mínimo, igual a um. Esta mudança drástica na cobertura faz com que os valores destes indicadores tanto menos comparáveis quanto maior a diferença entre as coberturas entre os casos de comparação.

A influência da cobertura também aparece indiretamente quando analisamos o valor do filtro de frequência e a alteração dos valores do critério de penalização. Para a situação sem filtro existe muito pouca variação para os indicadores conforme a penalização aumenta e este fator pode ser percebido nas figuras 22 e 23.

Com filtros crescentes os efeitos das penalizações ficam mais visíveis, uma vez que o número de itens sendo avaliados é menor e a mudança nas listas graças à penalização se torna mais explícita.

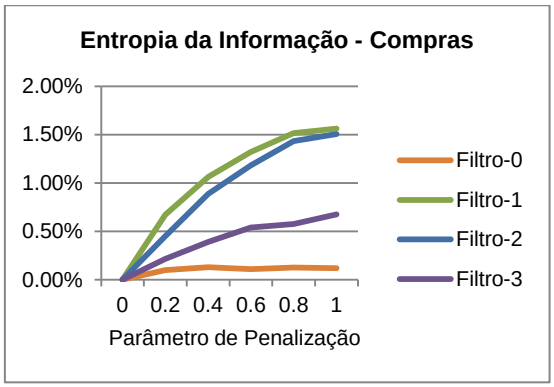


Figura 22 Entropia da Informação – Variação Filtro Frequência

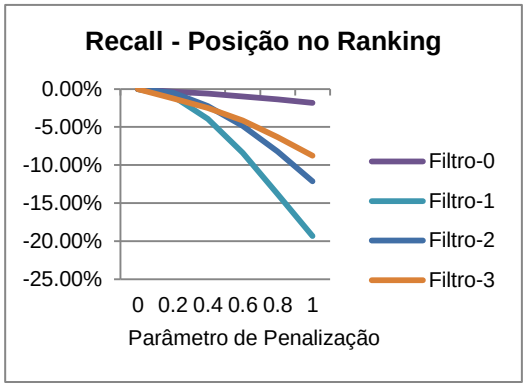


Figura 23 Recall considerando Ranking – Variação Filtro Frequência

A diminuição do efeito da penalização com a presença dos filtros pode ser considerada tanto positiva quanto negativa. É positiva, pois garante menos variação dos indicadores de assertividade conforme o efeito de penalização aumenta, ou seja, garante que mesmo que os itens mais populares sejam deslocados para o final da lista, ainda assim itens relevantes vão ser mostrados. É negativa, pois a diversificação da lista também é diminuída, oposto ao efeito de diversificação esperado.

Estes dois efeitos podem ser vistos nas figuras 22 a 24, pelo comportamento do indicador de Diversidade Intra-lista e Recall. Pela existência

desta relação inversa, existem evidências de que para obtenção de efeitos similares os valores de penalização devem ser aumentar conforme os valores dos filtros de frequência aumentam.

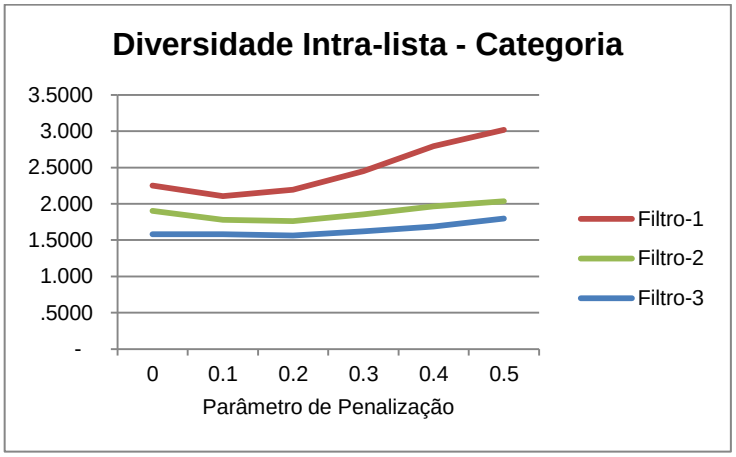


Figura 24 Diversidade Intra-lista – Variação da penalização

Uma vez que o uso dos filtros de frequência é indispensável e que a influência do filtro sobre a análise da evolução da penalização é determinante, escolheu-se limitar a análise para dois valores específicos de filtro, que são os valores usados hoje para os clientes da Chaordic. Podemos perceber que um valor filtro pequeno já promove grandes mudanças nas recomendações, como pode ser visto também nas figuras 22 e 23.

Os valores de filtro usados foram as frequências de compra mútua igual a 2 e 3, ou seja, só serão considerados pares de itens que tiverem sido comprados no mínimo 2 ou 3 vezes em conjunto. Esses valores foram escolhidos após análises de usuários sobre as recomendações geradas e vem se mostrando adequados ao funcionamento. Como o objetivo deste experimento não é a discussão do valor do filtro vamos simplesmente assumir estes valores.

A seguir levantou-se o tamanho médio das listas de recomendação por item. Os valores obtidos foram próximos de 17,5 itens para filtro mínimo igual a 2 e de 20,6 itens para filtro mínimo igual a 3. Intuitivamente, os valores do tamanho médio da lista deveria ser menor para filtros menores, porém como os filtros alteram o número de itens com recomendação, a média reflete o contrário.

Além disso, para filtro igual a 2, 94% dos pares de itens comprados juntos são filtrados, e para o filtro igual a 3, 96% dos pares são filtrados.

Estes valores justificam as análises de Assertividade mesmo com variações na cobertura. A pequena diferença entre os casos torna possível o uso dos indicadores de assertividade para a comparação entre os efeitos dos de filtros. Outro fato importante é que os indicadores de cobertura são muito fracamente influenciados pela variação do parâmetro de penalização, o que também já era esperado, uma vez que o parâmetro não diminui o tamanho das listas de recomendação, e só exclui itens recomendados em casos pouco frequentes, o que permite a comparação sem ressalvas dos valores dos indicadores quando usado o mesmo valor de filtro de frequência.

A última análise interessante antes de retomar o estudo da penalização é quanto à sensibilidade do indicador Entropia da Informação. Para a ausência de filtro de frequência este indicador praticamente não apresenta variação. Para os parâmetros de filtros usados, considerando toda a faixa de penalizações, sua variação máxima foi de 1,5% em relação à não penalização. De fato a influência da penalização altera as recomendações ligadas ao final da cauda de distribuições, e como o espectro completo de distribuição foi considerado o efeito percebido pelo indicador fica diluído. Apesar da baixa sensibilidade do indicador considera-se útil a realização de análises com ele. Como possibilidade futura fica a repetição da análise de influência da penalização neste indicador considerando apenas o final do espectro de distribuição de compras.

Voltando à análise da penalização, para os filtros determinados foram realizados estudos para variações do parâmetro de penalização variando de 0 até 2.

Para penalização igual a 0,5 o indicador Precisão apresenta decréscimo de apenas 0.5% para ambos os filtros. Para todo o intervalo de penalização, ocorre apenas uma pequena variação no indicador. Os efeitos descritos podem ser verificados nas figuras 25 e 26.

Para penalização igual a 0,5, o indicador Recall apresenta um decréscimo de aproximadamente 3,5% para ambos os filtros. A partir da penalização igual a 0,5 o decréscimo no valor do Recall segue um comportamento linear, atingindo decréscimo entre 18% e 25%.

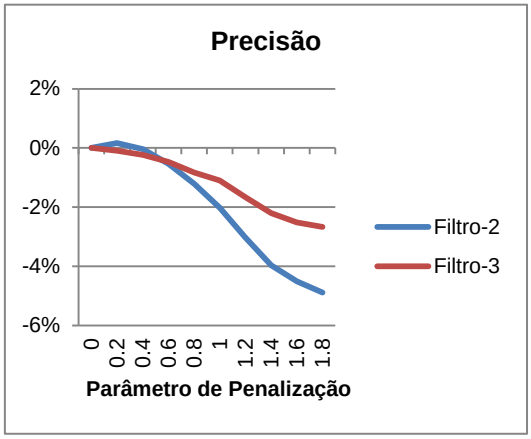


Figura 25 Precisão – Penalização e Filtros

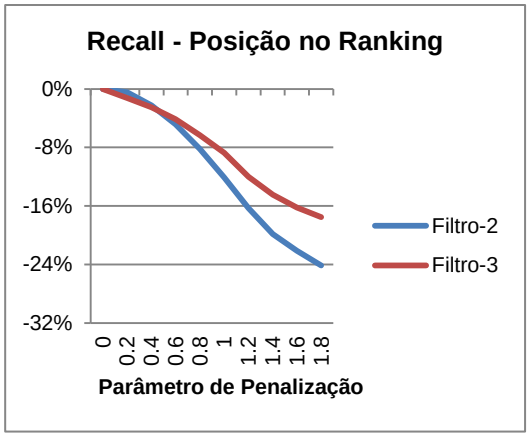


Figura 26 Recall considerando Ranking – Penalização e Filtros

Pela análise dos indicadores de novidade, vemos que a partir do valor de penalização igual a 0,5 os efeitos no indicador de Entropia da Informação tem sua variação suavizada, principalmente para o filtro igual a 2, como mostrado na figura 27. O uso de penalização além desta fronteira passa então a causar apenas a diminuição da assertividade do sistema com poucos ganhos em novidade.

Provavelmente, após este valor de penalização os itens restantes na lista de recomendação possuem frequências de compras muito parecidas, o que implica em probabilidades de compras também semelhantes, sendo que este é o valor base para o cálculo do indicador, o que explica sua estabilização.

Já para a análise do indicador de diversidade, Diversidade Intra-lista, mostrado na figura 28, percebe-se um leve decréscimo para pequenos valores de penalização e um acréscimo quase constante a partir da penalização igual a 0,5, para o filtro de frequência igual a dois. Para o filtro de frequência igual a 3 temos um crescimento constante a partir da penalização igual a 0,4, sem muitos ganhos antes disso.

O efeito comum aos filtros é que a partir de um valor de penalização a diversidade de categorias dentro dos tops dez itens por usuário aumenta quase constantemente. Como a lista considerada tem um tamanho relativamente pequeno, dez itens, variações na ordem de 10% significam grandes impactos na diversidade dos tops dez.

Como diversidade é uma propriedade importante para listas espera-se que o desempenho deste indicador não seja alterado negativamente, e por este

motivo valores de penalização próximos a 0,5 parecem adequados. Mesmo havendo aumento significativo da diversidade a partir desta fronteira, temos que considerar os decréscimos acentuados causados nos indicadores de assertividade para valores de penalização maiores que 0,5.

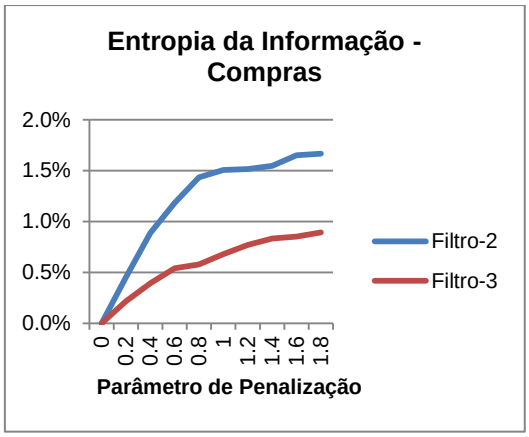


Figura 27 Entropia da Informação – Penalização e Filtros

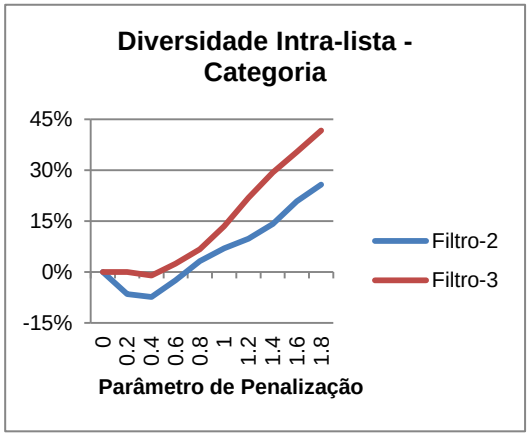


Figura 28 Diversidade Intra-lista – Penalização e Filtros

Conclui-se então que, para recomendações usando o algoritmo “Similar Items”, com dados de entrada sendo interações de compras de usuário e usando filtros de frequência na faixa de valores entre 2 e 3, valores de penalização ao redor de 0,5 se mostram os mais adequados.

Este valor de penalização difere dos atuais valores usados pela Chaordic, que ficam em torno de 0,2. Como explicado na seção 5.2, o valor de 0.2 foi determinado através de experimentos desenvolvidos quando o parâmetro de penalização foi incluído no algoritmo. Este valor vem se mantendo como referência, porém alterações nos valores de filtros são constantemente incluídas, pois o entendimento da influência dos filtros é mais intuitivo e existe mais liberdade em alterá-lo. Acredita-se que essas variações no decorrer do tempo sejam a justificativa da diferença significativa entre o valor atualmente usado e o valor estimado neste estudo de caso.

Como continuação deste experimento, para garantir que as conclusões sejam acertadas sugere-se a avaliação do mesmo efeito com diferentes métodos. Propõem-se uma avaliação qualitativa feita por usuários de teste e uma avaliação on-line, usando a ferramenta de testes AB que está sendo desenvolvida na Chaordic.

# 7. Conclusões e Perspectivas

Este documento descreve os trabalhos desenvolvidos durante 6 meses de estágio como requisito para a conclusão do curso de Engenharia de Controle e Automação na Universidade Federal de Santa Catarina.

Como objetivos do projeto pretendia-se expandir o conhecimento da empresa onde o projeto foi desenvolvido, a Chaordic Systems, na área de avaliação de Sistemas de Recomendação. Além da expansão do conhecimento pretendia-se a concepção de um projeto e implementação de um protótipo de ferramenta que auxiliasse o time de pesquisa da empresa na realização de experimentos comparativos e na caracterização de seus sistemas.

Esperava-se com isso que o time de pesquisa se munisse com análises quantitativas para argumentar a tomada de decisão sem precisar se basear apenas em observações qualitativas e altamente arbitrárias.

Como entrega concreta do projeto pode citar em um primeiro momento uma lista de indicadores de desempenho adequados aos contextos de produto da Chaordic. Esta lista sumarizou grande parte da pesquisa feita e deu diretrizes para o início da prática de avaliação off-line na rotina.

Como maneira de tornar práticos os estudos realizados, uma primeira proposta de ferramenta de experimentação foi concebida no formato de um framework de software. A ferramenta possibilitou o desenvolvimento de dois estudos de caso nos quais foi possível constatar a utilidade da avaliação off-line no sentido de embasar respostas do time de pesquisa, e promover compreensão sobre comportamentos desconhecidos dos atuais produtos da empresa.

Considerando o objetivo de incluir avaliações quantitativas no processo de avaliação dos sistemas da empresa considera-se que o objetivo foi atingido. O processo de avaliação off-line forneceu insumos para futuras experimentações on-line a partir dos dois estudos de caso apresentados. Além disso, as características definidas para a ferramenta de experimentação foram implementadas e um protótipo foi construído, e os aprendizados desta primeira versão servirão de insumo para o desenvolvimento de uma plataforma de experimentação, planejada como projeto que terá início em agosto de 2012.

Como trabalhos futuros propõem-se a realização das pendências propostas ao final das seções 6.1 e 6.2, dos estudos de caso feitos. Essas pendências

correspondem a dúvidas levantadas durante os estudos de caso e a resposta delas permitirá tanto o melhor entendimento dos resultados obtidos quanto à validação das estimativas.

Além disso, propõem-se a validação dos estudos de caso apresentados através de avaliações qualitativas e avaliações on-line. As avaliações qualitativas serão feitas por usuários de teste e as avaliações on-line serão feitas com a ferramenta de testes AB sendo desenvolvida pelo time de pesquisa.

Como plano em médio prazo tem-se a concretização de uma plataforma de experimentação a ser usada não só pelo time de pesquisa, mas também por outras equipes interessadas em trabalhar com os módulos de recomendação da empresa, distribuindo assim a possibilidade de promover inovação para além das fronteiras do time de pesquisa e desenvolvimento. Esta ferramenta de experimentação terá como base as avaliações por indicadores de desempenho, frutos da pesquisa realizada neste projeto.

No futuro espera-se construir na empresa um ciclo de avaliação maduro e consistente, que auxilie no ajuste automático dos nossos sistemas, baseados tanto em avaliações off-line como avaliações on-line. Esse processo automático de configuração e adaptação constante dos sistemas será essencial uma vez que a empresa pretende continuar crescendo e novos clientes irão ser constantemente incluídos.

Como último plano proposto espera-se que a pesquisa em avaliação off-line continue e que seja possível a confecção de um artigo científico a ser submetido ao evento de sistemas de recomendação Recsys em 2013.

## 8. Referências Bibliográficas

[1] Chirita, P., Nejdl, W., Zamfir, C., Preventing Shilling Attacks in Online Recommender Systems. In Proceedings of WIDM, Bremen, Germany (2005)

[2] Celma, O., Herrera, P., A New Approach to Evaluating Novel Recommendations. In Proceedings of RecSys 2008, Lausanne, Switzerland. p. 179 -186 (2008)

[3] Celma, O.; Herrera, P.; Music recommendation tutorial. In Proceedings of 8th International Conference on Music Information Retrieval, Vienna, Austria, Chapters 5, 6 (2007)

[4] Deshpande, M., Karypis, G.: Item-Based Top-N Recommendation Algorithms, ACM Transactions on Information Systems, vol. 22, no. 1, p. 143-177. (2004)

[5] Ekstrand, M., Riedl, J. T., Konstan, J. A.: Collaborative Filtering Recommender Systems. Foundations and Trends in Human-Computer Interaction. Minneapolis, USA (2011)

[6] Kincaid , J.: The Netflix Prize Comes To A Buzzer-Beater, Nailbitting Finish. Tech Crunch, 26 de Julho 2009. Disponível em <http://techcrunch.com/2009/07/26/the-netflix-prize-comes-to-a-buzzer-beater-nailbiting-finish/>. Acesso em 07 de Julho 2012.

[7] Kohavi, R., Longbotham, R., Sommerfield, D., Henne, R. M.:Controlled Experiments on the Web: survey and practical guide. Data Mining and Knowledge Discovery Journal. Springer. p 140-181 (2008)

[8] Krohn-Grimberghe, A., Nanopoulos, A., Schmidt-Thieme. L.: A Novel Multidimensional Framework for Evaluating Recommender Systems, *in* Martin Atzmüller; Dominik Benz; Andreas Hotho & Gerd Stumme, ed., 'Proceedings of LWA2010 - Workshop-Woche: Lernen, Wissen & Adaptivitaet' (2010)

[9] McNee, S. M., Riedl, J., Konstan, J. A.: Being Accurate is Not Enough: How Assertividade Metrics Have Hurt Recommender Systems. In Proceedings of Conference on Human Factors in Computing Systems. Quebec, Canada. (2006)

[10] Ricci, F., Rokach, L, Shapira, B.: Introduction to Recommender Systems Handbook. In: Ricci, F., Rokach, L., Shapira, B., Kantor, P. B.. Recommender Systems Handbook. New York. Springer; p. 1-35 (2011)

[11] Shani, G., Gunawardana, A,: Evaluating Recommendation Systems. In: Ricci, F., Rokach, L., Shapira, B., Kantor, P. B.. Recommender Systems Handbook. New York. Springer; p. 257-297 (2011)

[12] Vargas S. and Castells P., Rank and Relevance in Novelty and Diversity Metrics for Recommender Systems. In Proceedings of RecSys, Chicago, USA. P. 109 – 116 (2011)

[13] Zhang, M., Hurley, N., Avoiding Monotony: Improving the Diversity of Recommendation Lists. RecSyS, Lausanne, Switzerland. P. 123-130 (2008)

[14] Ziegler, C-N., McNee, S., Konstan, J., Lausen, G.: Improving Recommendation Lists Through Topic Diversification. In Proceedings of the International World Wide Web Conference Committee. Chiba, Japan (2005)

## Anexo A: Lista de Indicadores de Avaliação

| Propriedade | Indicador | Descrição |
|---|---|---|
| Assertividade | Recall ou Taxa de Acerto | A razão entre as previsões corretas (recomendadas e usadas) e o total de itens usados. |
| Assertividade | Precisão | Relação entre número de recomendações usadas e todas as recomendações providas. |
| Assertividade | Medida-F | Combina as medidas de Precisão e Recall, usando media harmônica. |
| Assertividade | Área sob a curva ROC | Curva ROC computada comparando a frequência de verdadeiros positivos e falsos positivos mensurados no cálculo do Recall e Precisão. O indicador é a area sob a curva. |
| Assertividade | Recall com consideração de Ranking | Mesmo cálculo do Recall, mas pontua o acerto com um valor dependente da posição no Ranking de recomendações. |
| Assertividade | Utilidade de Meia Vida | Tenta avaliar a utilidade da lista recomendada. A utilidade é definida como o desvio entre a nota dada pelo usuário e a nota média do item. |
| Assertividade | Kendall-tau | Medida de similaridade usada para comparar a lista de recomendações para um usuário com as interações feitas pelo mesmo. |
| Assertividade | Desempenho baseado em Distância – Normalizado | Medida da distância entre a escolha do usuário e a escolha do sistema. Considera usuário e sistema como vetores e mede a distância euclidiana. |
| Assertividade | Rho de Spearman | Mesma ideia do Kendall-tau, porém leva em consideração a posição do ranking de recomendações. |

| | | |
|---|---|---|
| Cobertura | Média do Menor Caminho | Teoria de grafos. Distância entre dois vértices i e j. Os vértices estão conectados se é possível ir de i até j. Também chamada de Média Geodésica. Informa sobre a característica global de navegação no grafo. |
| Cobertura | Clusterização | Teoria de Grafos. Medida de formação de clusters, que informa a porção de arestas de um vértice I com potencial relação com as arestas totais. |
| Cobertura | Grau de Distribuição Acumulativo | Teoria de Grafos. O número de vértices ligados à um vértice. Informa se algum nodo age como hubs. |
| Cobertura | Força da Maior Componente | Teoria de Grafos. O número de vértices conectados por uma geodésica e que estão desconectados de outros vértices. |
| Diversidade | Média da Dissimilaridade de Par | Média das diferenças de similaridade entre pares de itens dentro de um conjunto de recomendações. |
| Diversidade | Distância intra-lista esperada | Distância da similaridade entre itens dentro de uma lista, considerando também posição no ranking. |
| Diversidade | Diversidade Temporal Interna do Sistema | Definida como a razão entre itens recomendados não incluídos nas recomendações feitas anteriormente. Mede evolução do sistema. |
| Novidade | Popularidade Complementar Esperada | Mede a habilidade do sistema em recomendar itens relevantes da cauda longa. Informa o número de recomendações relevantes vistas que não foram previamente vistas. |
| Novidade | Distância de Perfil | Distância da similaridade entre itens recomendados e itens do perfil do usuário. |

## Anexo B: Caracterização dos Clientes

| Id | Mês | *Visualizações | P. D. V. | *Compras | P. D. P. | Pedidos | *Faturamento | *T. M. P. | *V. M. P. | *Catálogo | C. Catálogo |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Janeiro | 0.18 | 29037 | 0.415 | 7176 | 77284 | 0.447 | 1.006 | 1.084 | 0.539 | 9.58% |
| 2 | Janeiro | 0.85 | 76161 | 1.447 | 23445 | 266568 | 1.438 | 1.017 | 1.010 | 1.058 | 15.94% |
| 3 | Janeiro | 0.72 | 79267 | 1.332 | 25190 | 246116 | 1.226 | 1.014 | 0.933 | 1.009 | 17.97% |
| 4 | Janeiro | 1.00 | 75324 | 1.000 | 14980 | 187324 | 1.000 | 1.000 | 1.000 | 1.000 | 10.78% |
| 5 | Janeiro | 0.03 | 4842 | 0.044 | 1967 | 8465 | 0.058 | 0.984 | 1.274 | 0.055 | 25.70% |
| 1 | Fevereiro | 0.20 | 24524 | 0.473 | 7356 | 76572 | 0.534 | 0.999 | 1.129 | 0.539 | 9.82% |
| 2 | Fevereiro | 0.81 | 83106 | 1.345 | 24283 | 214546 | 1.294 | 1.014 | 0.976 | 1.058 | 16.51% |
| 3 | Fevereiro | 0.77 | 82341 | 1.303 | 24726 | 207711 | 1.156 | 1.015 | 0.901 | 1.009 | 17.64% |
| 4 | Fevereiro | 1.00 | 73173 | 1.000 | 14358 | 161842 | 1.000 | 1.000 | 1.000 | 1.000 | 10.33% |
| 5 | Fevereiro | 0.03 | 4648 | 0.037 | 1803 | 6115 | 0.043 | 0.980 | 1.132 | 0.055 | 23.56% |
| 1 | Março | 0.25 | 24190 | 0.526 | 7335 | 75320 | 0.551 | 1.019 | 1.067 | 0.539 | 9.80% |
| 2 | Março | 0.72 | 79622 | 1.429 | 25179 | 205821 | 1.389 | 1.012 | 0.984 | 1.058 | 17.12% |
| 3 | Março | 0.73 | 80294 | 1.293 | 23635 | 184833 | 1.180 | 1.019 | 0.931 | 1.009 | 16.86% |
| 4 | Março | 1.00 | 72080 | 1.000 | 14605 | 145777 | 1.000 | 1.000 | 1.000 | 1.000 | 10.51% |
| 5 | Março | 0.04 | 4608 | 0.043 | 1796 | 6381 | 0.045 | 0.972 | 1.031 | 0.055 | 23.47% |

* Os valores destes campos são sempre relativos, tomando como referência o cliente de Id igual a 4, o mesmo usado como referência nos experimentos.

Legenda:

P.D.V.: Número de produtos distintos visualizados

P.D.P.: Número de produtos distintos comprados

T.M.P.: Tamanho Médio de Pedido

V.M.P.: Valor Médio de Pedido

C. Catálogo: Cobertura Recomendações sobre Catálogo