UNIVERSIDADE DE CAXIAS DO SUL
PROGRAMA DE PÓS-GRADUAÇÃO EM ADMINISTRAÇÃO – PPGA
CURSO DE MESTRADO

OS SISTEMAS DE RECOMENDAÇÃO COMO INSTRUMENTO PARA ATINGIR MERCADOS DE NICHO

ANTÔNIO RÉGIS NODARI

Orientador: Prof. Dr. Antony Mueller

Caxias do Sul, 2008.

ANTÔNIO RÉGIS NODARI

OS SISTEMAS DE RECOMENDAÇÃO COMO INSTRUMENTO PARA ATINGIR MERCADOS DE NICHO

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Administração do Mestrado em Administração da Universidade de Caxias do Sul, como requisito parcial à obtenção do título de Mestre em Administração.

Orientador: Prof. Dr. Antony Mueller

Caxias do Sul, 2008.

ANTÔNIO RÉGIS NODARI

OS SISTEMAS DE RECOMENDAÇÃO COMO INSTRUMENTO PARA ATINGIR MERCADOS DE NICHO

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Administração do Mestrado em Administração da Universidade de Caxias do Sul, como requisito parcial à obtenção do título de Mestre em Administração

Caxias do Sul, 9 de maio de 2008

Banca Examinadora:

Prof. Dr. Antony Peter Mueller (Orientador)
Universidade de Caxias do Sul

Prof. Dr. Odacir Deonísio Graciolli
Universidade de Caxias do Sul

Prof. Dr. Rolando Vargas Vallejos
Universidade de Caxias do Sul

Prof. Dr. Jaime Evaldo Fensterseifer
Universidade Federal do Rio Grande do Sul

Dados Internacionais de Catalogação na Publicação (CIP)
Universidade de Caxias do Sul
UCS - BICE - Processamento Técnico

N761s Nodari, Antônio Régis
Os sistemas de recomendação como instrumento para atingir mercados de nicho / Antônio Régis Nodari, 2008.
86 f. : il. ; 30 cm.

Dissertação (Mestrado) – Universidade de Caxias do do Sul, Programa de Pós-Graduação em Administração, 2008.
"Orientação: Prof. Dr. Antony Mueller"

1. Comércio eletrônico. 2. Administração de vendas. 3. Empresa – Vinhos. 4. Venda por distribuição. 5. Internet. I. Título.

CDU: 658.849:004

Índice para o catálogo sistemático:

1. Comércio eletrônico	658.849:004
2. Administração de vendas	658.811
3. Empresa – Vinhos	658:663.2
4. Venda por distribuição	658.86
5. Internet	004.738.5

Catalogação na fonte elaborada pela bibliotecária
Márcia Servi Gonçalves – CRB 10/1500

RESUMO

O objetivo deste trabalho é estudar o efeito dos sistemas de recomendação em um *site* de vinhos, verificando se os resultados estão de acordo com a teoria *long tail*. Esta proposição prevê que em mercados *online*, os produtos de nicho podem representar uma parcela significativa do resultado de uma empresa. Uma das formas de explorar estas fontes de receitas é pelo uso adequado de sistemas de recomendação que auxiliem o consumidor a encontrar o que deseja. Neste trabalho são efetuados dois estudos de caso, o primeiro utiliza o coeficiente Gini para comparar a distribuição das vendas de duas empresas, sendo uma delas de comércio eletrônico, o segundo estudo de caso seleciona quatro tipos de sistemas de recomendação e compara seus desempenhos na sugestão de vinhos. Os resultados indicam que ocorre um comportamento do tipo *long tail* nas vendas da loja virtual e que os sistemas de recomendação baseados nos gostos de outras pessoas são os preferidos.

Palavras-chave: Comércio eletrônico, sistemas de recomendação e *long tail*.

ABSTRACT

The goal of this work is to study the effect of recommender systems in a wine website, verifying if the results are in conformity with the long tail theory. This proposition supposes that, in online markets, the niche products can represent a significant portion of a firm's profit. One method that can be used to explore these revenues is acommplished through an adequate use of recommender systems, which can help the custommers to find what they wish. In this work, two case studies are performed: The first one, uses the Gini coefficient to compare the sales distributions from two firms, where one of them is an ecommerce firm. The second one, selects four recommender systems types and compares their performance on wine recommendations. The results suggest that there is a long tail behavior in the ecommerce firm and that the recommender systems based on tastes from others are the favorite ones.

Keywords: e-commerce, recommender systems and long tail.

SUMÁRIO

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

INTRODUÇÃO

O objetivo deste trabalho é estudar os sistemas de recomendação em dois *sites* de vinhos e verificar se os resultados estão de acordo com a teoria *long tail*.

Long tail é o termo criado por Chris Anderson em seu artigo *The Long Tail* publicado na revista *Wired* em 2004 (ANDERSON, 2004) após ter pesquisado a distribuição de música, livros e filmes na internet. Posteriormente expande o conceito em forma de um livro *The Long Tail: Why the Future of Business is Selling Less of More*, traduzido no Brasil como "A Longa Cauda: Do Mercado de Massa para o Mercado de Nicho" (ANDERSON, 2006).

A principal idéia do *long tail* é que nas lojas virtuais, uma grande quantidade de itens que individualmente possuem pequeno volume de vendas, podem em conjunto representar um faturamento igual ou maior que os poucos produtos mais vendidos. Para que isso aconteça, um conjunto de ferramentas de *ranking* e de recomendação direcionam a demanda para estes nichos.

Os sistemas de filtragem de informações e recomendações da internet auxiliam o consumidor a encontrar produtos que de outra forma não teriam acesso. A internet trouxe ao alcance das pessoas uma maior variedade de opções mas também trouxe ferramentas para filtrar esta diversidade. Os mecanismos de busca se aprimoram para fornecer respostas mais relevantes e específicas. O resultado é que o consumidor praticamente não possui limitação de opções de escolha. Os comerciantes, por sua vez, devem aproveitar estes recursos tecnológicos para melhorar a experiência do seu cliente de forma que ele seja bem atendido e saia satisfeito.

1.1 OBJETIVO

Analisar o efeito de sistemas de recomendação em *sites* de vinhos para verificar se apresentam resultados de acordo com a teoria *long tail*.

1.1.1 Objetivos específicos

a) Pesquisar, analisar e comparar os métodos utilizados para gerar as listas de recomendações;
b) Testar métodos de recomendação e avaliar seu desempenho.
c) Verificar se os sistemas de recomendação levam os consumidores para o *long tail*.

1.1.2 Questão de pesquisa

Os sistemas de recomendação podem auxiliar o comércio eletrônico de vinho a alcançar o *long tail*?

1.2 CONCEITOS CENTRAIS

1.2.1 The Long Tail

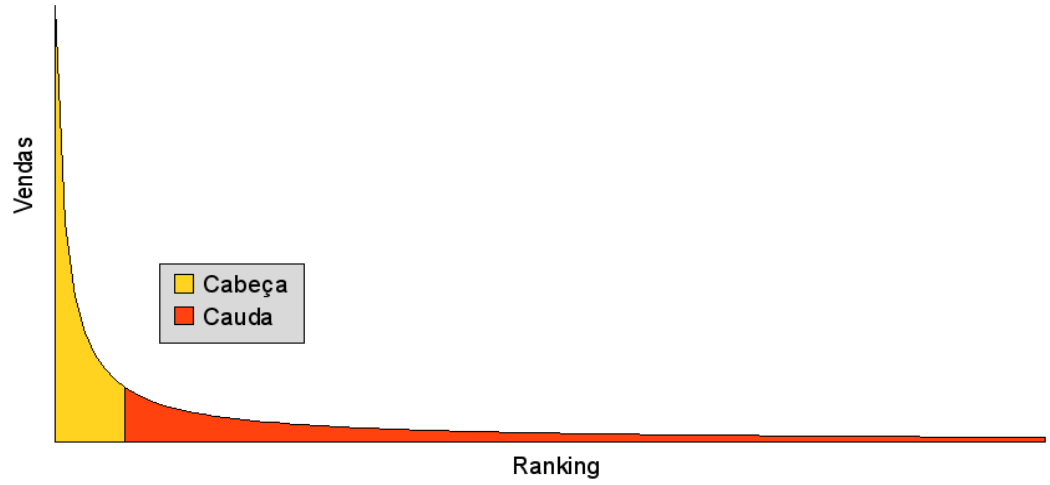
O termo *long tail* ou cauda longa refere-se a uma distribuição estatística, também chamada de lei de potência, na qual a população com freqüência elevada ocorre no início, a cabeça, e decai bruscamente formando uma cauda que pode conter um volume total maior que a cabeça, ao ser desenhado forma uma curva conforme o exemplo da figura 1.

A idéia do *The Long Tail* de Chris Anserson, é descrever uma mudança na cultura do consumo. Ele mostra que uma grande quantidade de itens que individualmente possuem pouca demanda ou baixo volume de vendas podem, em conjunto, obter uma fatia de mercado tão grande ou maior que os poucos *best sellers* e *blockbusters* existentes. O exemplo tradicional é o caso da Amazon.com, no qual uma grande proporção dos livros vendidos não é originada pelos dez maiores, mas sim pelos títulos obscuros que não são encontrados nas lojas físicas. A internet possibilita às pessoas encontrarem aqueles títulos, abrindo as portas de um gigantesco mercado antes inacessível.

Os seis temas que trata o *The Long Tail* são: (1) nos mercados virtuais, há muito mais

itens de nicho do que de *hits;* (2) *o*s custos para atingir estes nichos estão em queda; (3) um conjunto de ferramentas de *ranking* e de recomendação, está direcionando a demanda para estes nichos; (4) como há variedade, meios para filtrá-la e organizá-la, a curva de demanda torna-se achatada, continua havendo *hits* e nichos, mas os *hits* tornam-se menores e os nichos relativamente maiores; (5) os nichos se acumulam, mesmo que nenhum deles venda grandes quantidades, juntos eles compreendem um mercado que rivaliza com os produtos mais vendidos; (6) uma vez que isso ocorre, surge a forma verdadeira e natural da curva de demanda, sem as distorções provocadas pelas limitações de espaço de prateleira, escassez de informação ou gargalos de distribuição. A demanda é muito mais diversificada que o tradicionalmente aceito (ANDERSON, 2006).

Figura 1: Exemplo de gráfico para distribuição do tipo *long tail*.



Fonte: Elaborado pelo autor.

A maior parte do crescimento do consumo nos últimos anos é atribuído ao aumento da variedade de produtos (BRESNAHAN E GORDON, 1997 apud BRYNJOLFSSON, HU E SIMESTER, 2007). Mesmo que a internet tenha baixado os preços no mercado de livros, o benefício ao consumidor de ter a possibilidade de buscar dentro de uma grande variedade de títulos, é de 5 a 7 vezes mais importante que os preços menores (BRYNJOLFSSON, HU, SMITH, 2003). Enquanto a tecnologia da informação continua a baixar o custo de buscar os produtos, a variedade e a distribuição das vendas entre os produtos também muda de forma correspondente. Muitos mercados são dominados por poucos produtos mais vendidos, como por exemplo os livros, em que as vendas estão concentradas em alguns poucos títulos de alguns autores *best sellers*. (GRECCO, 1997 *apud* BRYNJOLFSSON, HU E SIMESTER, 2007).

Os economistas e administradores geralmente utilizam o princípio de Pareto, também chamado de 80/20 para descrever este fenômeno de concentração de vendas. Esta regra expressa que uma pequena proporção dos produtos em um mercado (20%), produzem uma grande proporção no valor total das vendas (80%). Contudo, a internet possui o potencial de mudar este balanço. Por reduzir os custos de busca, a tecnologia da informação pode aumentar radicalmente o compartilhamento coletivo de produtos de nicho achatando a distribuição das vendas, de forma que estes produtos representem uma fatia significativa do faturamento da empresa.

É possível que na internet, o Princípio de Pareto tenha dado lugar ao *long tail*. As evidências sugerem que os mercados de internet tenham alterado o balanço de poder entre os produtos mais vendidos e os anteriormente obscuros produtos de nicho. Brynjolfsson, Hu e Smith (2003) mostram que títulos desconhecidos, que não são estocados por livrarias convencionais por causa de seu baixo volume de vendas, representaram 40% do faturamento em livros da Amazon.com no ano 2000. Resultados semelhantes foram obtidos em mercados *online* de música, DVDs e eletrônicos.

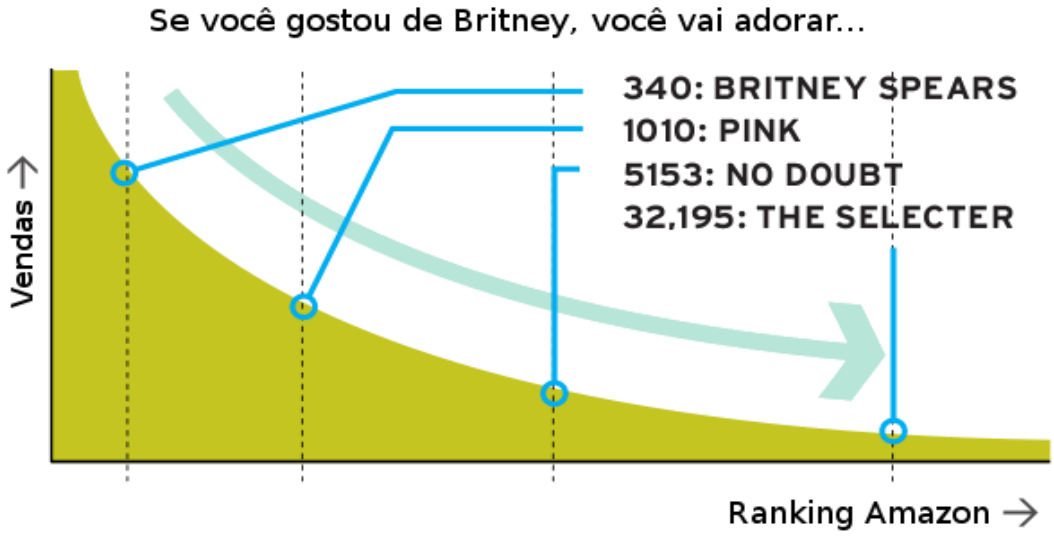
1.2.2 Sistemas de recomendação

Os sistemas de recomendação são ferramentas que podem sugerir algum produto de forma automatizada. São utilizados pelos *sites* que fornecem indicações aos seus clientes. Os itens podem ser recomendados usando como base os mais vendidos, as características demográficas do cliente ou produto, as suas compras anteriores, por indicação direta de outros usuários ou ainda através do uso de tecnologias de *data-mining*. Desta maneira o *site* se adapta ao usuário personalizando a experiência do consumidor.

Da mesma forma que os preços mais baixos, os sistemas de recomendação também podem encaminhar os clientes para resultados obscuros que não seriam encontrados de outra forma (figura 2). Os sistemas de recomendação podem melhorar as vendas de três formas: a primeira é transformando os visitantes em compradores, já que ajudam os visitantes a encontrar o que procuram, a segunda maneira é aumentando as oportunidades de vendas cruzadas, pois podem sugerir produtos complementares ao selecionado pelo cliente e a terceira forma é aumentando a lealdade do cliente, já que este irá retornar ao *site* quando

perceber que suas necessidades estão sendo bem atendidas (SCHAFER, 1999).

Figura 2: Como os sistemas de recomendação conduzem o consumidor ao *long tail*.



Fonte: Anderson (2006).

1.2.3 Comércio eletrônico

O comércio eletrônico é definido como a compra e venda de produtos, serviços e informações através de redes de computadores (KALAKOTA e ROBINSON, 2002). Também são consideradas como comércio eletrônico as atividades de suporte para qualquer tipo de transação de negócios que ocorram através de infra-estrutura digital (BLOCH, PIGNEUR e SEGEV, 1996).

O comércio eletrônico também pode ser visto como um canal de distribuição que enfatiza o meio no qual os produtos ou serviços são distribuídos ou um como mercado onde os compradores e vendedores se encontram para negociar (JANSEN; STEENBAKKERS; JÄGERS, 1999). Permite atingir o consumidor de forma direta e torna possível automatizar transações entre empresas. O modelo se adapta particularmente bem para produtos digitais que são entregues ao cliente através da rede e bens físicos de pequeno volume ou valor agregado, em que a venda dependa pouco do contato físico com o produto e que o envio ao destino possa ser efetuado através de serviços de entrega tradicionais como os correios por

exemplo, é o caso dos livros, CDs e DVDs.

As mudanças trazidas pela utilização da internet como canal de vendas são tão significativas que se faz necessário tentar analisá-las e procurar entendê-las, especialmente quando se pensa nos desenvolvimentos tecnológicos que ainda estão por vir. Os sistemas de transmissão de dados cada vez mais velozes, o aumento da capacidade de processamento e armazenamento dos computadores pessoais, a integração das telecomunicações aos computadores têm sido identificados como exemplos da revolução que se aproxima. Mas todos os setores da economia estão sofrendo, ou virão a sofrer, em maior ou menor escala, os impactos desta nova forma de fazer negócios (LOHSE & SPILLER, 1998).

Não existe uma linha clara que divida o comércio eletrônico e o comércio tradicional, trata-se de um gradiente que vai do puramente tradicional até o totalmente eletrônico, muitas empresas trabalham com catálogos eletrônicos ou possuem vitrines virtuais e outras que iniciaram no comércio eletrônico abrem lojas físicas em pontos estratégicos.

1.3 METODOLOGIA

Como o trabalho está dividido em dois estudos de caso independentes, cada um conta com uma metodologia diferente. O estudo analisa os dados obtidos em empresas e confronta os resultados com a teoria. Os dados são os registros das vendas e os registros de acessos efetuados pelos visitantes dos *sites*. Esta metodologia tem por objetivo testar a teoria, aumentando a compreensão preditiva do fenômeno (NEVES, 1996).

O primeiro estudo de caso tem por objetivo comparar as distribuições das vendas de vinhos de duas empresas em termos de *long tail*. As empresas comparadas são a VinhosNet, especializada em vendas *online* de vinhos e afins e a outra é um supermercado local, o Apolo. Para esta, somente foram utilizados dados das vendas de vinhos. A comparação utiliza o coeficiente Gini como um índice que mede a desigualdade nas vendas. O *long tail* prevê que em mercados *online*, a desigualdade entre os mais vendidos e os menos vendidos é menor.

Para o estudo dos sistemas de recomendação, foram utilizados dados de registros de acessos de um outro *site* especializado em recomendação de vinhos. Neste caso são implementados sistemas de recomendação e os resultados de seu uso são acompanhados. Neste caso o *long tail* prevê que ocorra uma redistribuição das visitas de forma que diminua a

desigualdade entre os produtos mais visitados e os menos visitados.

1.4 JUSTIFICATIVA

Na *internet*, as opções de escolhas multiplicaram-se, entretanto ela também fornece mecanismos para filtrar estas opções. Em uma situação hipotética de um consumidor que planeja comprar um aparelho eletrônico, ele primeiramente busca em diversos *sites* pesquisando preços e as ofertas, lê as opiniões de outros compradores dos mesmos produtos que deseja, vai ao *site* do fabricante para conhecer as características técnicas do aparelho e verificar se existe assistência técnica em sua cidade, pesquisa em fóruns e *sites* de reclamações sobre a marca e o modelo e finalmente se em sua avaliação de valor este produto realmente apresentar uma boa relação custo/benefício ele opta pela compra. Esta compra poderá ser em uma loja virtual ou física. O sistema de recomendação da empresa pode ser o diferencial que faça o consumidor optar pela compra.

Um dos objetivos do comércio eletrônico é tornar os visitantes em compradores e os sistemas de recomendação podem auxiliar a atingir esta meta. Desta forma, um sistema de recomendação bem estruturado poderia se tornar uma fonte de vantagem competitiva para a empresa.

O lojista virtual deve conhecer e identificar as oportunidades de atender melhor aos seus clientes, as tecnologias adequadas podem trazer melhores resultados para seu negócio, tanto em termos absolutos quanto em termos indiretos de forma a aumentar a satisfação de seus clientes.

A escolha do produto vinho como item do estudo também é justificada pela localização geográfica, já que a região da serra gaúcha representa 90% da produção de vinhos no Brasil e o mercado é constituído em sua maioria por pequenas empresas.

1.5 RELEVÂNCIA

A internet é importante no cotidiano dos indivíduos e das empresas, não apenas o uso

trivial como meio de comunicação bidirecional de texto, imagem e som, mas também como plataforma para prestação de serviços, alguns deles, nem existiriam de outra forma, como a Wikipedia por exemplo. Serviços voltados para a informação, como busca, catalogação, compartilhamento e difusão de conhecimento são os mais adaptados.

Além disso, de acordo com alguns estudos (E-BIT, 2006), houve um aumento de 72% no volume de vendas com comércio eletrônico e este número vem chamando a atenção, não apenas de empresários, mas também de empreendedores e entidades públicas e privadas, como o Sebrae que lançou um programa para estimular empreendimentos *online*.

A questão de inovação também é relevante, pois o assunto trata de uma área que é relativamente recente e encontra-se em evolução. Esta evolução faz uso intensivo de tecnologia da informação e telecomunicação, desta forma, o comércio eletrônico exige que seus participantes invistam e criem oportunidades em desenvolvimento tecnológico.

A variedade de produtos oferecidos no varejo que antes era limitada por fatores econômicos e financeiros como o espaço físico, hoje praticamente não existe, isto faz com que a forma de pensar do empresário do comércio deva ser revista. Os consumidores mais exigentes já não se limitam a pesquisar produtos e preços em poucos locais, eles sabem que o mundo tornou-se menor e tudo está ao seu alcance. Neste aspecto, entender o comportamento deste novo consumidor é importante para saber como atendê-lo e mantê-lo.

2 CARACTERÍSTICAS DO COMÉRCIO ELETRÔNICO

2.1 O QUE É A ECONOMIA DIGITAL?

O comércio eletrônico é resultado da evolução da economia digital, para grande parte das pessoas, a expressão economia digital refere-se simplesmente a economia da internet, mas ela é muito mais ampla. Ela representa o uso pervasivo da Tecnologia da Informação e Comunicação – TIC em todos os aspectos da economia, isto inclui as operações internas das organizações, as transações entre elas e entre indivíduos. A tecnologia da informação trouxe ferramentas para criar, manipular, organizar, transmitir, armazenar e agir com a informação digital. (ATKINSON e MCKAY, 2007)

As tecnologias que fundamentam a economia digital, vão muito além da internet e computadores pessoais, a TI está inserida em uma vasta gama de produtos além dos computadores e celulares, como os automóveis, máquinas de lavar, cartões de crédito, aviões, equipamentos médicos e máquinas industriais. Em 2006, 70% dos microprocessadores produzidos não eram usados em computadores, mas eram embarcados em equipamentos e cada vez mais conectados e as empresas continuam inventando novos usos para a TI a cada dia (ATKINSON e MCKAY, 2007).

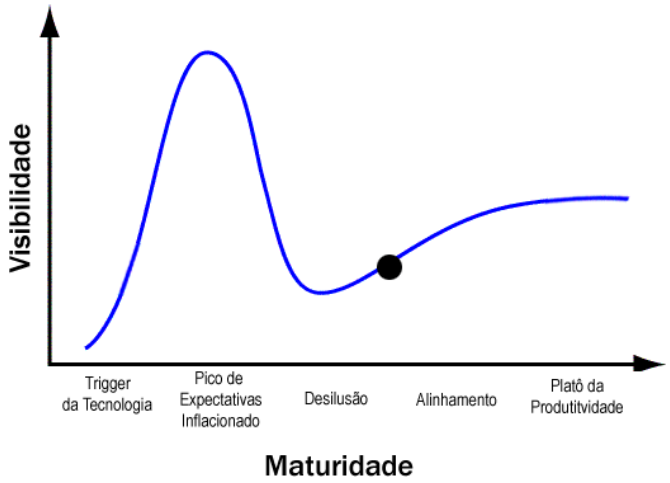
A economia digital é apoiada pelo fato da TI ser uma tecnologia de uso geral. As tecnologias de uso geral possuem 3 características: primeiro, elas são usadas em muitos setores; segundo, sua performance e preço melhoram com o tempo e terceiro; elas tornam mais fácil criar novos produtos, processos e modelos de negócios (BRESNAHAN e TRAJTENGERG, 1995). O motor a combustão, a máquina a vapor e a eletricidade são exemplos de tecnologias de uso geral que guiaram as transformações e o crescimento econômicos no passado. A TI possui as características de tecnologia de uso geral, e hoje está em virtualmente todos os setores, os preços de produtos de TI continuam em declínio, a sua capacidade continua aumentando seguindo a lei de Moore e novos modelos de negócio,

produtos e serviços apoiados em TI são criados a todo momento.

2.2 O IMPACTO ECONÔMICO DO COMÉRCIO ELETRÔNICO

O comércio eletrônico no Brasil já ultrapassou os 10 anos de idade, mesmo assim existem muitos aspectos que não estão estabelecidos na cultura empresarial. Observa-se, que a maioria dos grandes magazines possuem vitrines na internet mas que as pequenas lojas ainda não possuem, ou pelo menos não as divulgam. Mas, ao acessar a internet, percebe-se uma enxurrada de ofertas de produtos em centenas de propagandas nos principais portais que são oferecidos por lojas muitas vezes pouco conhecidas.

Figura 3: Curva de maturidade da tecnologia



Fonte: Gartner Group adaptado pelo autor (GARTNER, 2007).

A tecnologia do comércio eletrônico já alcançou maturidade e pode atingir o platô da produtividade dentro de alguns anos (figura 3). É um dos resultados visíveis da economia digital, sua importância econômica deve-se a cinco fatores principais (WYCKOFF e COLECCHIA, 1999):

a) O comércio eletrônico transforma o mercado. Ele muda a maneira pela qual os negócios são conduzidos. As funções tradicionais do intermediário são alteradas, novos produtos e mercados são desenvolvidos, os relacionamentos entre empresas e

consumidores se tornam muito mais próximos, altera a organização do trabalho, cria novos canais para a difusão de conhecimento e cria novas funções exigindo novas habilidades dos profissionais.

b) O comércio eletrônico age como um catalisador. Ele serve para acelar e difundir de forma mais ampla as mudanças que estão em curso na economia, estabelecendo ligações eletrônicas entre empresas, globalizando as atividades econômicas e a demanda por trabalhadores altamente qualificados. Setores como bancos, viagens e marketing são acelerados pelo comércio eletrônico.

c) O comércio eletrônico incrementa a interatividade da economia. O aumento das possibilidades de comunicação e transações instantâneas em qualquer hora e local erodem as fronteiras econômicas e geográficas.

d) A abertura é o fundamento do comércio eletrônico. A internet nasceu como uma plataforma aberta e está suportada por padrões abertos, permite a comunicação e transparência entre empresas que dão acesso ao seus bancos de dados e pessoal. Esta abertura torna possível que o consumidor se torne um parceiro na criação e desenvolvimento de produtos. Ela pode tanto trazer benefícios como a transparência e competição ou problemas como a diminuição da privacidade.

e) O comércio eletrônico altera a importância relativa do tempo. muitas das rotinas que auxiliam a definir a aparência da economia são dependentes do tempo. A produção em massa é a maneira mais rápida de reduzir os custos. O comércio eletrônico reduz a importância do tempo por incrementar a velocidade dos ciclos de produção, permitindo às empresas operarem em coordenação e aos consumidores, conduzirem suas transações com menos perda de tempo.

2.3 O PAPEL DO COMÉRCIO ELETRÔNICO NA EFICIÊNCIA ECONÔMICA

Ainda de acordo com Wyckoff e Colecchia (1999), a razão fundamental que determinou o crescimento rápido do comércio eletrônico, especialmente entre empresas, foi o seu impacto significativo sobre os custos fixos e a produtividade. Um fator chave para redução dos custos de estoques é a adoção de sistemas *just-in-time* e uma melhora habilidade

de previsão de demanda e os dois podem ser obtidos pela adoção do comércio eletrônico, que aproxima os fornecedores e clientes.

Ao efetuar o pedido eletronicamente, em um formato acessível, as empresas ampliam a eficiência do processo de vendas. (WYCKOFF e COLECCHIA, 1999) Mesmo quando o consumidor efetua uma transação ao vivo ou pelo telefone, ele chega sabendo quais produtos quer e já está pronto para comprá-los, permitindo um aumento de produtividade dos vendedores em um fator de 10. A interface eletrônica também reduz o retrabalho no processamento dos pedidos, pois somente são efetivados após terem passado por uma pré-conferência automatizada pelo sistema.

Para economias baseadas em conhecimento e dominadas por produtos sofisticados, o pós-vendas representa até 10% do custo operacional. (WYCKOFF e COLECCHIA, 1999) Nestes casos, as empresas podem fornecer suporte online, manuais interativos e dinâmicos cortando custos e melhorando a qualidade do serviço.

2.4 ESTATÍSTICAS DO COMÉRCIO ELETRÔNICO

A pesquisa sobre o uso da TIC no Brasil 2006, conduzida pelo Comitê Gestor da Internet no Brasil, (CETIC.BR, 2007) apresenta um mapa da internet brasileira. Ele mostra que a grande maioria das empresas que tem computador utiliza a internet (94,8%) e 38,8% de seus funcionários têm acesso à rede.

Realizado entre os meses de julho e novembro de 2006 em todo o território nacional, o estudo apresenta os números levantados em empresas com 10 funcionários ou mais pertencentes ao setor organizado da economia no Brasil, listadas na RAIS (Relação Anual de Informações Sociais), e pertencentes a sete segmentos da Classificação Nacional de Atividades Econômicas.

2.4.1 Uso da internet

O acesso à Internet já é quase onipresente entre as empresas brasileiras, no entanto,

apenas 48,8% das empresas pesquisadas possuem *website*, e pouco mais da metade já utilizou a rede para compras (52,1%) e para venda de produtos e serviços (50,2%) e os resultados estão apresentados quadro 1 (CETIC.BR, 2007) mostrando que ainda há muito a ser feito para que o potencial das tecnologias da informação e da comunicação seja absorvido pelo setor produtivo brasileiro.

Quadro 1: Proporção de indivíduos que já compraram produtos e serviços pela internet.

Percentual sobre o total de pessoas que já acessaram a internet[1]				
Percentual (%)		SIM	NÃO	Não sabe
Total		14,00	85,55	0,45
REGIÕES DO PAÍS	SUDESTE	13,70	85,72	0,58
	NORDESTE	10,97	88,68	0,35
	SUL	13,89	85,85	0,26
	NORTE	14,03	85,81	0,15
	CENTRO-OESTE	16,83	82,55	0,62
SEXO	Masculino	15,53	84,12	0,34
	Feminino	12,38	87,05	0,57
GRAU DE INSTRUÇÃO	Analfabeto/ Educação infantil	2,74	96,39	0,87
	Fundamental	5,67	93,66	0,68
	Médio	11,74	87,77	0,49
	Superior	27,03	72,85	0,12
FAIXA ETÁRIA	De 10 a 15 anos	3,35	95,58	1,06
	De 16 a 24 anos	12,39	87,37	0,24
	De 25 a 34 anos	19,67	79,95	0,39
	De 35 a 44 anos	18,06	81,78	0,16
	De 45 a 59 anos	23,19	76,52	0,29
	De 60 anos ou mais	15,70	81,27	3,03
RENDA FAMILIAR	ATÉ R$300	1,17	94,76	4,07
	R$301-R$500	6,09	93,50	0,41
	R$501-R$1000	8,18	91,19	0,63
	R$1001-R$1800	15,66	84,09	0,25
	R$1801 OU MAIS	26,70	72,88	0,43
CLASSE SOCIAL[3]	A	40,54	59,46	-
	B	20,02	79,64	0,34
	C	10,81	88,63	0,57
	DE	4,72	94,82	0,46
SITUAÇÃO DE EMPREGO	Trabalhador	17,60	82,14	0,26
	Desempregado	9,13	90,87	-
	Não integra a população ativa[2]	6,92	92,17	0,91

1 Base: 3.502 entrevistados que já usaram a internet.
1Projeção populacional: 51 milhões de pessoas, com 10 anos ou mais, segundo estimativa realizada com base na PNAD 2005.
2Na categoria não integra população ativa estão contabilizados os estudantes, aposentados e as donas de casa.
3O critério utilizado para classificação leva em consideração a educação do chefe de família e a posse de uma serie de utensílios domésticos, relacionando-os a um sistema de pontuação. A soma dos pontos alcançada por domicílio é associada a uma Classe Sócio-Econômica específica (A, B, C, D, E).
Fonte: (CETIC.BR, 2007).

Dados relacionados ao tipo de acesso mostram que 88,8% possuem conexão por meio de banda larga e a atividade mais freqüente, com 97,7% é enviar e receber *emails*. A busca de informações (92,4%) e os serviços bancários e financeiros (80%) aparecem logo em seguida. 83% das empresas utilizam a Internet para interagir com órgãos públicos.

A pesquisa apresenta dados sobre a situação do comércio eletrônico segmentada por região, sexo, grau de instrução, faixa etária, renda familiar, classe social e situação de emprego. No resultado geral, somente 14% dos que já acessaram a internet compraram produtos e serviços pela rede. A faixa que apresenta maior índice é representada pela classe A com 40,54%.

Outra dimensão do setor é mostrada pela pesquisa FGV-EAESP de Comércio Eletrônico no Mercado Eletrônico, edição 2007, (FGV-EAESP, 2007) que considerou 402 empresas, dos vários setores econômicos, ramos de atividades e portes. As empresas, tanto nacionais como multinacionais, operam no mercado brasileiro e atuam em diversos níveis no ambiente digital.

Segundo os dados obtidos, o percentual do comércio eletrônico nas transações de negócio-a-negócio representa 36,45% do total do mercado, e 12,71% para negócio-a-consumidor. A utilização das aplicações de comércio eletrônico para a integração das empresas com seus fornecedores é aproximadamente 70% no setor de comércio, sendo que esta integração inclui a troca eletrônica de dados com plataforma proprietária. Por outro lado, cerca de 78% das empresas pesquisadas, também considerando o setor de comércio, já utilizam estas aplicações na integração com clientes.

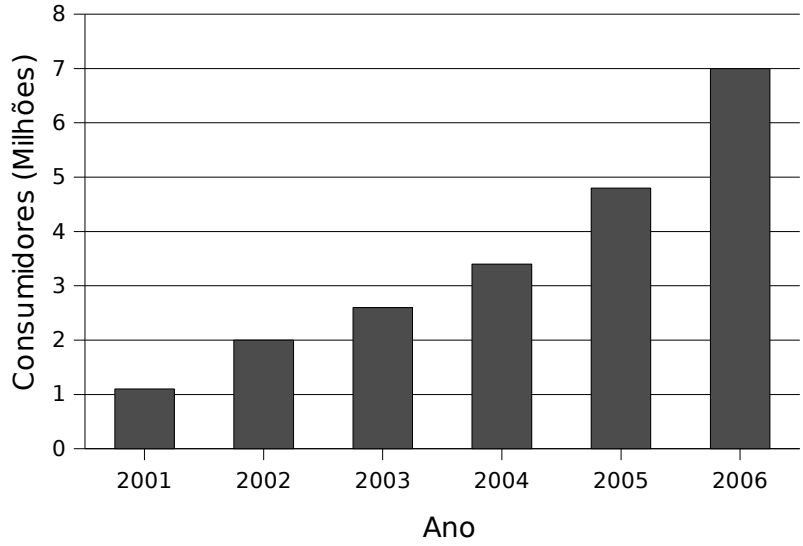
Nos Estados Unidos, as vendas no comércio eletrônico para consumidor atingiram US$93 bilhões no ano de 2005 demonstrando um crescimento de 22.2% em relação a 2004. O crescimento rápido tem sido a regra, de 2000 a 2005, as vendas tiveram um crescimento médio anual de 27.3% e pode ser comparado com os 4.3% do comércio em geral. No entanto, no mesmo período representou apenas 2.5% do total das transações. (US CENSUS BUREAU, 2007).

2.4.2 O Relatório *WebShoppers*

A empresa de pesquisa e marketing *online* e-bit vem coletando informações sobre o comportamento do comércio *online* no Brasil desde o ano 2000. Para isso utiliza um sistema de coleta de dados automatizado que obtém diariamente informações detalhadas sobre as vendas geradas pelo próprio consumidor após a compra *online* em mais de 700 lojas virtuais brasileiras agregando mensalmente ao seu banco de dados 100 mil questionários. Desta forma

já coletou mais de 4 milhões de questionários desde janeiro de 2000. Estas informações são compiladas e geram relatórios de inteligência de mercado, o perfil sócio demográfico do e-consumidor, produtos mais vendidos, meios de pagamento mais utilizados (E-BIT, 2006).

Figura 4: Evolução do Número de consumidores *online* no Brasil 2001-2006.



Fonte: Grupo de Pesquisas e-bit (E-BIT, 2006).

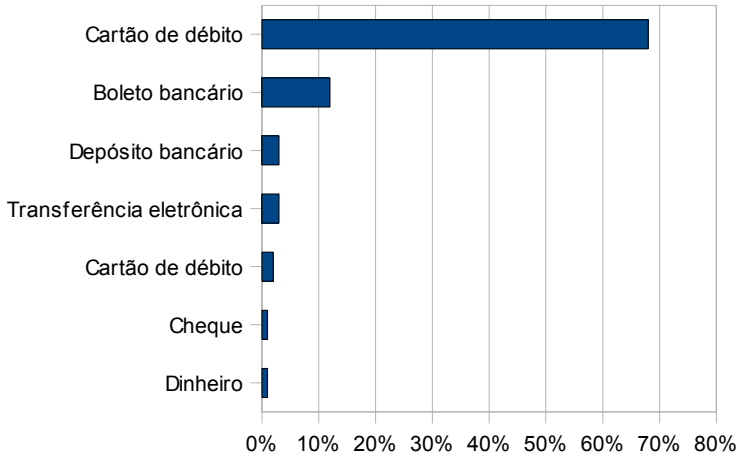
O estudo calculou que no fechamento de 2006, o comércio eletrônico obteve um faturamento de R$4,4 bilhões significando um aumento nominal de 76% em relação a 2005, (Figura 4) representando o mesmo valor que a soma dos anos 2001 a 2004. O valor médio gasto em cada compra foi de R$296,00, mas supõe que o valor deva aumentar devido à maior quantidade de aparelhos eletrônicos sendo comercializados com valor agregado maior.

Não apenas a base de consumidores aumentou (Figura 4), mas aumentou também o uso do meio, desta forma o setor fechou com 14,8 milhões de compras ao longo do ano de 2006. Os produtos mais vendidos foram livros, revistas e jornais, com 17%, em segundo lugar ficaram os títulos de CD's, DVD's e vídeos com 16% e em terceiro lugar ficaram os eletrônicos como TV's, aparelhos de som, vídeo ou DVD. A tendência é que os eletrônicos subam para a primeira posição devido ao fato de as primeiras categorias serem substituídos por meios eletrônicos e a medida que os internautas adquiram maior confiança no canal passem a comprar produtos de maior valor agregado. O público feminino representa 43% e 70% encontra-se na faixa etária entre 25 e 49 anos. O estado de São Paulo possui 65% dos e-consumidores brasileiros.

O estudo também cita o índice de satisfação do consumidor *online* criado pelo grupo em que são usados 10 quesitos de satisfação que englobam de aspectos técnicos como a facilidade de encontrar os itens no sistema, pontualidade, política de segurança e informações sobre os produtos. Neste índice, a satisfação do consumidor atingiu 86%.

O meio de pagamento preferido na internet continua sendo o cartão de crédito com 68%, apesar de ser o mais inseguro, seguido pelo boleto bancário com 12%, percebe-se o aparecimento de TEF (transferência eletrônica de fundos) com 3% e cartões de débito com 2% (Figura 5).

Figura 5: Meios de pagamento na compra de bens de consumo na internet em 2006



Fonte: Grupo de Pesquisas e-bit (E-BIT, 2006).

2.5 O COMÉRCIO ELETRÔNICO E SUAS CLASSIFICAÇÕES

De acordo com Andam (2003) o comércio eletrônico pode ser classificado em cinco categorias principais, são elas: o comércio eletrônico entre empresas ou B2B, entre empresa e consumidor ou B2C, entre empresas e governos B2G, entre pessoas C2C, entre consumidores e empresas C2B. O *M-commerce*, que utiliza dispositivos móveis, não seria uma categoria mas uma forma de acesso.

O comércio eletrônico entre empresas, B2B abreviação de *business to business*,

representa cerca de 80% do total de operações efetuadas no comércio eletrônico. A maioria das utilizações consistem em gerenciamento da cadeia de suprimentos, especialmente processamento de ordens de compras, gerenciamento de estoques, de distribuição, gerenciamento de canais e gerenciamento de pagamentos.

O segundo tipo, *business to consumer* (B2C), envolve o processo de obter informações e a compra de produtos por consumidores, é a segunda maior forma de comércio eletrônico e alguns exemplos são a Amazon.com e a Submarino.com. Também é considerado nesta classificação investimentos efetuados eletronicamente, como ações e aplicações usando ferramentas de *e-banking*.

O comércio B2C reduz os custos de transação, especialmente os custos com pesquisa de preços por permitir aos consumidores encontrar o melhor preço para um produto ou serviço rapidamente. Também reduz as barreiras de entrada no mercado pois o custo de instalar uma loja virtual é muito menor que o de uma loja física.

A terceira categoria em importância, é o comércio eletrônico entre empresas e governos ou setor público (B2G). Refere-se ao uso da internet para licitações, licenciamentos e outras atividades governamentais. A vantagem desta forma, é que ela permitiria processos de compra menos burocráticos, mais transparentes e rápidos.

Comércio eletrônico consumidor a consumidor, é simplesmente o comércio entre indivíduos, enquadram-se os mercados eletrônicos e sistemas de leilão online. Como exemplos pode-se citar os portais de leilão, como o *eBay*, os portais de classificados, como o *Craig Lists* e sistemas de troca de arquivos entre indivíduos chamados de *peer to peer*.

O M-commerce é o uso de dispositivos móveis para efetuar transações eletrônicas, esta tecnologia está sendo utilizada principalmente por bancos e empresas de telecomunicações.

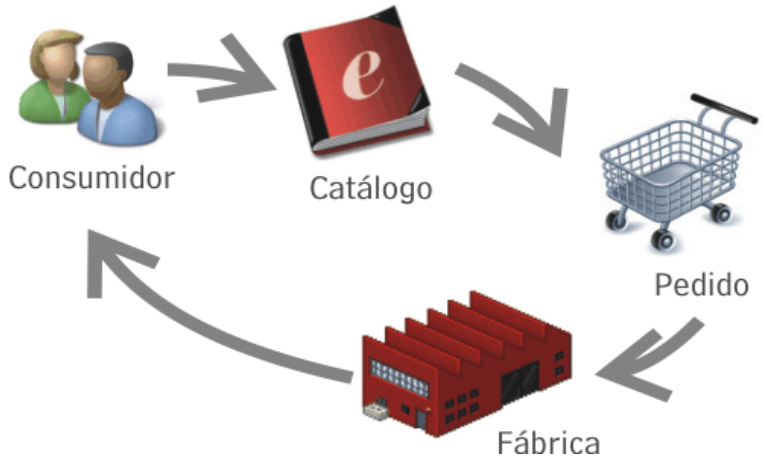
2.6 A ESTRUTURA DE UM COMÉRCIO ELETRÔNICO

Esta seção trata dos aspectos relativos à organização e estruturação de uma empresa de comércio eletrônico. Apresenta os elementos que um empreendedor deverá se preocupar ao

decidir investir em um negócio de comércio virtual. Alguns destes elementos são comuns tanto para o varejo *online* quanto para o eletrônico enquanto outros são específicos. No varejo virtual, o ponto que deve receber atenção inicial é o *software* que representa a estrutura que vai apoiar todo o negócio e o *website*. Depois de definido este item, a atenção do empreendedor deverá se voltar para a logística, divulgação. Os meios de pagamento que serão oferecidos e a segurança com que o cliente pode comprar também serão tratados.

Uma empresa tradicional de comércio para consumidores necessita um espaço físico para armazenar os estoques, expor seus produtos e receber os seus clientes. A localização da empresa é importante e varia de acordo com o tipo de produto. No caso de uma empresa de comércio eletrônico, a localização não chega a ser crucial, sendo que em alguns casos, é possível que nem haja a necessidade da existência deste local.

Figura 6: Processo de compra em loja virtual.



Fonte: Elaborado pelo autor.

A localização “real” da loja virtual é no seu endereço internet ou o endereço www. Por trás deste endereço existe um computador conectado à internet que pode estar localizado em qualquer ponto do planeta. Este servidor constitui a verdadeira localização física da empresa. Já o seu endereço internet poderia ser considerado o equivalente à vitrine e ao balcão da loja tradicional.

A estrutura de software para uma loja virtual, na sua forma mais simples, é de um catálogo *web* de produtos e um formulário de pedido. O cliente acessa este catálogo para pesquisar os itens disponíveis e se tiver interesse, pode adicionar os produtos ao pedido para comprá-los.

As formas de implementação variam pouco de uma empresa para outra, pode-se dizer que exista quase um padrão para lojas virtuais, isso é bom pois o consumidor acaba ficando familiarizado com o processo. A grande maioria das lojas utiliza a metáfora do carrinho de compras que é um sistema de armazenamento temporário das informações do catálogo que serão posteriormente transferidas para o pedido. O processo pode ser visualizado na figura 6, o visitante e possível cliente, navega pelo catálogo de produtos da empresa e se resolve comprar, efetua um pedido.

Do lado da empresa, ocorre o processamento deste pedido. Este trabalho envolve alguns passos como a captura do crédito na operadora de cartões ou a verificação do pagamento na conta bancária no caso de outras formas de pagamento e finalmente o envio da mercadoria.

3 REVISÃO DA BIBLIOGRAFIA

O século XXI pode ser chamado de era da abundância de informação. Durante os anos 90, ocorreu uma explosão de tecnologias para entretenimento e também das escolhas que podem ser feitas. Uma pessoa pode escolher entre centenas de canais de televisão, milhares de filmes, músicas e livros. A internet, através do comércio eletrônico, facilitou o acesso a produtos e serviços aumentando ainda mais a possibilidade de escolhas. Ainda ofereceu uma variedade de novos serviços e formas de entretenimento, como redes sociais, chats e jogos *online*. Apesar do número de opções ter aumentado muito, a capacidade de uma pessoa em escolher permaneceu a mesma e o trabalho para selecionar um produto aumentou, assim as pessoas não conseguirão avaliar todas as possibilidades a menos que restrinjam bem o que desejam.

Em seu artigo A Tirania da Escolha, Barry Schwartz (2004) demonstra que o consumidor, ao deparar-se com uma quantidade muito grande de opções, tende a ficar insatisfeito com a opção escolhida. O motivo, segundo o autor, seria uma frustração ao imaginar que outros produtos poderiam ser melhores que a escolha efetuada. Um estudo em um supermercado mostrava que as vendas de extrato de tomate diminuíam quando a variedade era maior que quatro ou cinco opções. Os sistemas de recomendação podem ser uma ferramenta útil para os consumidores, se estes conseguirem auxiliar a reduzir o estresse da escolha.

Este capítulo apresenta as técnicas e tecnologias utilizadas para a geração e apresentação de recomendações em *sites* de internet. Na seção mineração de dados, são discutidos os conceitos em que se baseiam as técnicas de extração de informação em grandes bases de dados e nas quais estão apoiados os sistemas de recomendação.

A seção seguinte, investiga os sistemas de recomendação identificando os problemas, desafios e oportunidades. São apresentados exemplos e identificados os conceitos básicos e questões de projeto. Também são estudadas as diferentes formas de implementação e as características principais de cada uma delas.

3.1 MINERAÇÃO DE DADOS

Mineração de dados ou *data mining* é definido como "descobrir conhecimento novo escondido em grandes massas de dados" por Carvalho (2001), "um método automático para descobrir padrões em dados, sem a tendenciosidade e a limitação de uma análise baseada meramente na intuição humana" por Braga (2005) ou ainda a "extração de informação preditiva oculta de grandes bases de dados" por Berson, Smith e Thearling (1999), ainda pode ser "o processo de exploração e análise automática e semi-automática de uma base de dados com o objetivo de descobrir modelos e regras siginificativas" por Berry (2004).

As técnicas de *data mining* são o resultado de um processo de pesquisa e evolução de algoritmos[1] que somente puderam ser utilizadas comercialmente devido a existência de computadores e bancos de dados capazes de armazenar e processar volumes massivos de dados. O poder computacional dos processadores permitem que seja possível, não apenas armazenar os dados gerados por milhares de transações *online*, mas também tornam viável analisar estes dados e extrair informações.

Como o termo mineração indica, significa escavar em uma montanha de dados tentando extrair algo valioso para o negócio. Este valor pode estar na descoberta de relações ocultas entre variáveis, como produtos que são comumente comprados em conjunto ou detectar situações comuns em tentativas de fraudes, no caso de empresas de cartão de crédito que detectam padrões fraudulentos, ou ainda a predição automatizada de comportamentos ou tendências, como a oferta de produtos financeiros mais adequados para o cliente com o perfil mais adequado.

A mineração de dados está inserida em um contexto maior chamado descoberta de conhecimento em bancos de dados, sigla KDD em inglês, ela estaria restrita apenas às fases de obtenção de modelos. Autores de metodologias e fabricantes de ferramentas descrevem diversos processos e metodologias (RAPIDMINER, 2007) para descoberta de conhecimento. Estes processos definem os passos necessários para atingir os objetivos. Os processos mais aceitos pela indústria são o CRISP-DM[2] e o SEMMA[3] (BRAGA, 2005). Ambos possuem basicamente as mesmas etapas: coleta de dados, depuração e análise, resultando em um

1 Um algoritmo é uma seqüência de passos necessários para solucionar um determinado problema.
2 Acrônimo de Cross Industry Standard Process for Data Mining.
3 Sample, Explore, Modify, Model, Access.

modelo descritivo e caso se deseje, os resultados podem ser utilizados na construção de um modelo preditivo. De uma forma genérica, os dois processos contém as fases de definição do problema, aquisição e avaliação dos dados, extração de características e realce, prototipagem e desenvolvimento de modelos, avaliação do modelo, implementação e avaliação do investimento.

As técnicas estatísticas tradicionais também são utilizadas na mineração de dados. Estatísticos e mineradores de dados resolvem problemas semelhantes, contudo, devido a diferenças históricas e na natureza do problema, também há diferenças nos métodos. Os mineradores de dados, geralmente possuem quantidades enormes de dados com poucos erros de medições, estes dados variam conforme o tempo e os valores muitas vezes são incompletos. Métodos estatísticos tradicionais são muito úteis na análise de dados, ao olhar para os dados é importante estudar histogramas para saber quais valores são os mais comuns e também olhar para os valores dentro de uma escala de tempo. Uma outra utilização da estatística é verificar se os valores são os esperados ou não, para isso o desvio padrão da média pode ser usado para calcular a probabilidade do valor ser casual resultando em hipóteses nulas. O teste chi-quadrado é um método estatístico muito útil, este método calcula diretamente os valores estimados para dados em linhas e colunas e basendo-se nestes cálculos, é possível saber se os resultados são plausíveis ou não.

Técnicas de mineração de dados

Além da estatísticas, técnicas podem ser implementadas por algoritmos de inteligência artificial como as redes neurais artificiais e os algoritmos genéticos. Exitem muitos algoritmos para implementação da mineração de dados, mas é possível classificá-los em cinco técnicas gerais: classificação, estimativa, previsão, análise de afinidade e análise de agrupamentos. (CARVALHO, 2001).

A classificação é uma das técnicas mais utilizadas da mineração de dados simplesmente porque é uma das tarefas humanas mais comuns. O ser humano está sempre classificando tudo à sua volta, criando classes e relações entre elas, criando estereótipos e grupos. Assim, ao receber um estímulo é capaz de rapidamente classificar dentro de alguma pré-classificação. Na mineração de dados, são comuns as tarefas de classificação como por exemplo clientes com perfil de alto, médio ou baixo risco. As redes neurais artificiais,

estatística e algoritmos genéticos são algumas das ferramentas muito utilizadas para classificar dados.

Estimar significa determinar um valor aproximado ou mais provável para um indicador utilizando-se dados passados ou indicadores semelhantes. Um exemplo é estimar a nota de avaliação que um consumidor daria para um vinho baseando-se em notas dadas por outros consumidores. Certamente as estimativas podem conduzir a erros, mas também podem fornecer aproximações que auxiliem a tomada de decisões.

A técnica de previsão é semelhante a estimativa, mas diferencia-se pois esta é dependente do tempo, por exemplo, pode-se prever que a temperatura média do planeta deva aumentar nos próximos anos devido as mudanças climáticas. Para isso, pode-se utilizar dados estatísticos ou redes neurais.

A análise de afinidade determina fatos que ocorrem conjuntamente ou quais itens em uma massa de dados estejam presentes juntos. O exemplo mais comum é o carrinho de compras de um supermercado, onde se pode inferir quais produtos serão comprados agrupados, possibilitando a oferta de um próximo ao outro ou em forma de kits.

A técnica de análise de agrupamentos, permite descobrir quantas e quais classes existem em uma massa de dados. Ao contrário da técnica de classificação, onde as classes já encontram-se definidas, na análise de agrupamentos a dificuldade reside no fato de que em uma base de dados, pode não haver classe alguma. Os grupos são construídos com base na semelhança entre os elementos e cabe ao analista determinar se eles têm algum significado ou utilidade (CARVALHO, 2001).

3.2 OBTENÇÃO DAS RECOMENDAÇÕES

Quando uma pessoa precisa escolher sem ter conhecimento das alternativas, a forma natural de agir é depender das experiências e opiniões de outros. As pessoas buscam recomendações de outras que já experimentaram ou de outros que sejam reconhecidos como especialistas no assunto. Geralmente obtém estes conselhos com parentes, amigos, colegas de trabalho, o atendente da locadora, críticas em jornais, revistas, *sites* e associações de defesa do consumidor. Desta forma, o processo social de compartilhamento de interesses é tão

importante quanto a troca de recomendações.

Neste contexto, os sistemas de recomendação informatizados estão tendo um papel importante. Da mesma forma que a tecnologia da informação e a internet trouxe mais opções, também trouxe a possibilidade de auxiliar o ser humano a escolher. Os sistemas de recomendação buscam mediar, dar suporte ou automatizar o processo de compartilhar recomendação. (TERVEN e HILL, 2001)

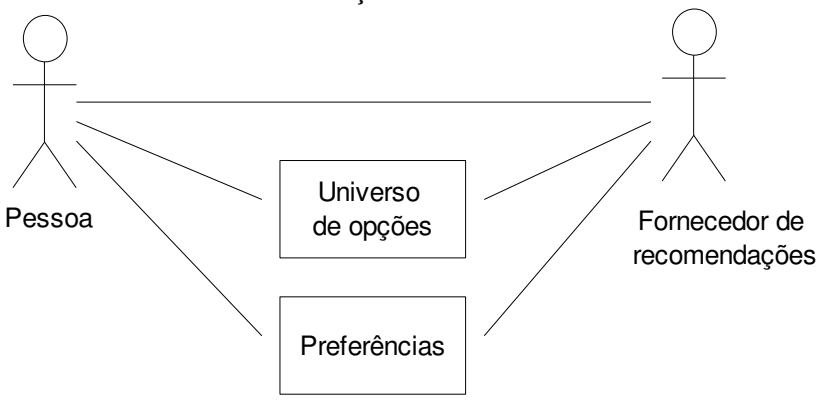
3.3 O PROCESSO DE RECOMENDAÇÕES

Quando uma pessoa precisa escolher entre um universo de alternativas, ela deve tomar uma decisão. Pode ser que ela nem conheça todas as opções possíveis pois este universo de opções pode ser grande. Se esta pessoa não possuir experiência anterior ou conhecimento suficiente para tomar esta decisão, ela pode ter que ir buscar recomendações de outros. Desta forma, a recomendação é um ato de comunicação.

A recomendação é baseada nas preferências do recomendante e também nas preferências daqueles que buscam as recomendações. (TERVEN e HILL, 2001) Neste caso, uma preferência é um estado mental individual considerando-se um subconjunto de itens de um universo de alternativas. As pessoas formam suas preferências basando-se em suas experiências anteriores relativas ao mesmo assunto, como por exemplo, gosto musical, filme, comida, etc. Uma recomendação pode ser direcionada para um indivíduo ou difundida para qualquer um que tiver interesse. A recomendação recebida torna-se um recurso que auxilia a filtrar todo um universo de opções.

A figura 7 representa um modelo geral para o processo de recomendações. Quando uma pessoa busca por alguma recomendação, ela pode optar por buscar no universo de opções ou pode buscar ajuda em algum fornecedor de recomendações. O fornecedor de recomendações conhece o universo de opções e as preferências associadas. Pode conhecê-las por sua experiência pessoal no assunto ou pelo relacionamento com outras pessoas que tiveram interesses semelhantes, e poderá auxiliar a pessoa a encontrar alguma opção que agrade.

Figura 7: Processo de recomendações.



Fonte: Elaborado pelo autor.

3.4 SISTEMAS DE RECOMENDAÇÃO INFORMATIZADOS

Um sistema de recomendação informatizado ou computacional automatiza ou fornece apoio ao processo de recomendação. O sistema automatizado assume o papel do fornecedor de recomendações pois ele oferece sugestões baseando-se nas preferências dos usuários e talvez de outras pessoas também. Um sistema de suporte de recomendações simplesmente auxilia as pessoas a compartilharem suas recomendações. Devem ser observadas quatro características ao projetar um sistema de recomendações automatizado: as preferências, os papéis e a forma de comunicação, os algoritmos e a forma de interação entre usuário e computador e serão explicadas a seguir.

As recomendações são baseadas em preferências, portanto o sistema de recomendações automatizado deve coletar as preferências das pessoas relativamente ao domínio do assunto tratado. Para isso devem ser tratadas as questões sobre quais as preferências que serão utilizadas. Podem ser as da própria pessoa e também de outras que já utilizaram o sistema. A forma que as preferências serão obtidas: explicitamente, quando o usuário informa seus gostos, por exemplo “ficção científica”, ou implicitamente fornecendo exemplos, como “Star Wars” e o sistema busca as características do exemplo. Também pode ser levantadas as questões sobre os incentivos que as pessoas terão para deixar suas preferências para o sistema e a forma que isso deve ser feito e armazanado.

As definições de papéis e a forma de comunicação representam o segundo ponto no projeto. Devem ser definidas as questões de quem deve fazer o papel de fornecedor de

recomendações, pode ser o próprio sistema ou pessoas e se for o sistema, de que forma esta comunicação será efetuada, o usuário deve pedir a recomendação ou o sistema deverá apresentá-las por conta própria. Quais informações sobre quem recomendou serão divulgadas para quem recebe e as formas de garantir a privacidade dos usuários.

O terceiro ponto é a seleção dos algoritmos para calcular as recomendações. Como as recomendações serão calculadas e como o sistema vai determinar quais preferências utilizar nestes cálculos. Finalmente o último item é a forma de interação entre computador a pessoa, que é a interface de usuário. Deve ser definido como o sistema vai exibir estas recomendações.

3.4.1 Apresentação das recomendações

Para Schafer (1999) as recomendações podem ser apresentadas aos usuários em sete formas principais. Estas interfaces de apresentação podem variar conforme o objetivo. Por exemplo, para as empresa de comércio eletrônico, o principal interesse é tentar vender mais dos seus produtos e o sistema pode exibir as recomendações em diversos momentos, já nos sistemas de busca, os resultados somente serão mostrados quando o usuário os solicitar. Os sete tipos de interfaces para recomendações relacionados são:

a) **Navegação:** Ao acessar o *site*, o usuário encontra os itens categorizados e pode ir diretamente ao grupo que interessa, como por exemplo em uma locadora, o usuário poderia buscar um filme policial dos anos 60, neste caso ele navega até a categoria filmes policiais e filtra o período.

b) **Itens semelhantes**: Esta técnica é herdada do comércio tradicional, onde o balconista, ao perceber o interesse do cliente em algum produto, pode apresentar outros itens parecidos, como variações de cores ou tamanhos.

c) **Email**: As recomendações podem ser enviadas para o cliente por email, neste caso, o sistema pode oferecer aos clientes novos itens que o sistema saiba que atinja seu perfil ou novas sugestões, a medida que o sistema aprende através das preferências de outros clientes.

d) **Comentários de texto**: Os *sites* permitem que os visitantes escrevam um

depoimento sobre a sua experiência com o produto, este depoimento geralmente é menos parcial que a descrição fornecida pelo fabricante, assim os outros usuários podem contar com uma segunda opinião aumentando a confiança na escolha.

e) **Média das avaliações**: Os *sites* podem pedir simplesmente um voto com uma nota, utilizando uma escala do tipo Likert e apresentar a média das notas. Como é mais simples e rápido do que escrever um depoimento, pode-se obter uma quantidade maior de avaliações.

f) **Os N mais vendidos**: Vários sites exibem a lista dos produtos mais vendidos, esta lista pode ser personalizada de acordo com a seção em que o cliente se encontra ou de acordo com o seu gosto aprendido em outras compras.

g) **Resultados de busca ordenados**: Ao informar os resultados de uma consulta por algum produto, o sistema pode retornar os itens ordenados pela relevância dos itens em relação às preferências do usuário.

3.5 PRINCIPAIS ABORDAGENS

Muitos dos sistemas de recomendação foram criados durante os anos 90 e a terminologia no assunto proliferou. Termos como filtragem colaborativa, filtragem social e navegação social foram utilizadas para descrever vários trabalhos desenvolvidos.

Schafer (1999) descreve uma taxonomia para classificar os sistemas de recomendação segundo duas dimensões principais: o grau de automação e a persistência da recomendação. As recomendações completamente automáticas aparecem para o usuário sem que ele precise pedir, as manuais exigem que o usuário procure pelas recomendações. O grau de persistência divide as recomendações entre aquelas que são geradas na hora, sem que o sistema utilize dados de visitas ou compras anteriores do cliente, que são chamadas de efêmeras e as persistentes, para as recomendações que são geradas baseadas na história de visitas ou compras do usuário.

Para Schafer (1999), as técnicas de sistemas de recomendação estão divididas em quatro grupos de alto nível: recomendações não personalizadas, baseadas em atributos,

correlações item-a-item, e correlações pessoa-a-pessoa. Os sistemas de recomendações não personalizadas, simplesmente apresentam avaliações feitas por outras pessoas, elas são invariantes e todos que acessarem, receberão as mesmas recomendações. Recomendações baseadas em atributos utilizam as características do produto para gerar as recomendações e sugerem produtos semelhantes. Os sistemas de correlação item-a-item, geram recomendações baseando-se em um pequeno conjunto de itens que o usuário tenha selecionado previamente ou produtos que sejam complementares ou estejam relacionados por algum motivo. Correlação pessoa-a-pessoa também chamada de filtragem colaborativa gera as recomendações baseando-se em avaliações de outras pessoas com gostos semelhantes.

Terven e Hill, (2001) classificam as abordagens em quatro grupos segundo a forma de coleta de preferências, os papéis representados pelos usuários, os algoritmos utilizados e a interface de usuário. Os quatro grupos identificados são os sistemas baseados em conteúdo, os sistemas de suporte a recomendações, a mineração de dados sociais e a filtragem colaborativa.

a) **Sistemas baseados em conteúdo** utilizam somente as preferências de quem busca a recomendação; eles tentam recomendar itens que sejam similares àqueles que o usuário tenha gostado no passado. Seu foco são os algoritmos que aprendam as preferências do usuário e filtrem dentro de um conjunto de novos itens aqueles que fiquem mais próximos deste gosto.

b) **Sistemas de suporte a recomendações** não automatizam o processo de recomendações, portanto eles não necessitam representar as preferências ou o calcular as recomendações. Eles servem apenas como ferramentas para ajudar as pessoas a compartilharem suas recomendações com outras, auxiliando tanto aquelas que buscam quanto as que produzem recomendações.

c) **Mineração de dados sociais** é uma técnica computacional que busca registros de preferências implícitas em atividades sociais do usuário como os fórums, blogs e redes sociais. Estes sistemas também focam as questões envolvidas no projeto de interfaces de usuários necessárias para exibição dos resultados. Estas visualizações tem sido utilizadas para auxiliar na navegação de espaços de informação como a internet.

d) **Sistemas de filtragem colaborativa** exigem que as pessoas que estão buscando recomendações, avaliem uma dúzia de itens antes de fornecerem os resultados, desta forma o usuário assume tanto o papel de fornecedor de recomendações quanto de

buscador. Estes sistemas focam em algoritmos que casam as preferências dos usuários com as de outros com gostos semelhantes.

O quadro 2 apresenta de forma resumida estas quatro formas de implementação e as características de projeto relacionadas. Cada forma de implementação será melhor detalhada nas próximas seções.

Quadro 2: Sistemas de recomendação, características do projeto e formas de implementação

Características de projeto	Formas de implementação			
	Baseados em conteúdo	**Suporte a recomendações**	**Mineração de dados sociais**	**Filtragem colaborativa**
Preferências	Somente as preferências de quem busca		Coleta em sites sociais	Quem busca deve fornecer suas preferências
Papéis e comunicação	Comunicação automatizada e papéis assimétricos	O sistema apenas fornece o suporte para humanos.	Automatizado pelo sistema, potencial para comunidades. Papéis assimétricos	Automatizado pelo sistema, potencial para comunidades. papéis simétricos
Algoritmos	Aprendizagem de máquina, recuperação de informação.		Mineração de dados	Ponderação e agrupamento de preferências.
Interface de usuário	Simples	Simples	Interfaces complexas, mapas e gráficos.	Simples

Fonte: Terven e Hill, (2001)

3.5.1 Sistemas de recomendação baseados em conteúdo

Os sistemas de recomendação baseados em conteúdo são construídos com o intuito de encontrar coisas que o usuário tenha gostado no passado. Eles aprendem as preferências do usuário através do seu *feedback.* Este *feedback* pode ser explícito, por exemplo, os usuários avaliam como bom ou ruim. Ou pode ser implícito, baseando-se em indicadores indiretos, um exemplo poderia ser o tempo que o usuário permaneceu ouvindo uma música. As preferências são representadas como um perfil dos interesses do usuário naquele tipo específico de conteúdo, muitas vezes expresso na forma de um conjunto de palavras-chave balanceadas de acordo com um peso proporcional de cada uma. Técnicas de aprendizagem de máquina e recuperação de informações são aplicadas para aprender e representar as preferências.

3.5.2 Sistemas de suporte a recomendações

Sistemas de suporte a recomendações são ferramentas computacionais para auxiliar as pessoas nas tarefas de compartilhar, produzir e buscar recomendações (TERVEN e HILL, 2001). O sistema Tapestry, desenvolvido pelos pesquisadores do Xerox PARC, é considerado o primeiro sistema de recomendações. Tratava-se de um sistema de mensagens que permitia aos usuários classificarem as mensagens como bom ou ruim e associarem anotações às mesmas. As mensagens armazenadas em bancos de dados poderiam ser recuperadas não apenas com base em seu conteúdo, mas também com base nas opiniões de outros usuários. Assim, por exemplo, alguém poderia buscar mensagens contendo palavras-chave inseridas em anotações feitas por outros.

Sistemas de suporte a recomendações são efetivos quando há pessoas suficientes dispostas a buscar e recomendar informações. Em sua maioria não são pagas por isso, fazem simplesmente por gostarem. Assim as necessidades de ambos os lados são satisfeitas, quem busca uma recomendação encontra alguém que ajude e quem fornece, sente-se reconhecido e gratificado pelo *feedback* positivo recebido. Para os usuários, as recomendações devem ser relevantes e interessantes pois não adianta simplesmente receber dicas de assuntos que não interessem.

Muitos *blogs* encontram-se nesta categoria, seus editores pesquisam temas por motivos profissionais ou mesmo por *hobby* e registram em seus *sites* de forma organizada, permitindo que outros usem estas informações de forma gratuita. Com o tempo, um *blog* pode ganhar notoriedade e desta forma tornar-se uma autoridade sobre algum assunto (TERVEN e HILL, 2001). *Blogs* como engadget.com recebem tanto ou mais tráfego que *sites* de editoras de publicações respeitadas.

3.5.3 Mineração de dados sociais

Na mineração de dados sociais os pesquisadores buscam locais onde grupos de pessoas estão naturalmente produzindo registros digitais como fóruns, listas de emails e redes sociais e a informação potencialmente implícita nestes registros é colhida e agregada utilizando-se técnicas computacionais e ao mesmo tempo são criados sistemas que permitam a visualização destas informações. Sistemas de mineração de dados sociais não necessitam que

os usuários alterem suas formas de trabalho ou façam qualquer nova atividade pois eles tentam encontrar as preferências das pessoas implícitas dentro das suas atividades normais.

A Word Wide Web, com seus conteúdos, links e registros de uso, é um terreno fértil para obtenção de dados para a pesquisa em mineração de dados sociais. Um link entre um site e outro já indica que há um relacionamento entre os dois. Diversos algoritmos de agrupamento e classificação foram projetados para formalizar esta informação. O Google utiliza um algoritmo de análise de similaridade entre os links para agrupar e ordenar os sites. Outros trabalhos focam em extrair informações de fóruns. O sistema PHOAKS (TERVEEN et al, 1997) busca informações sobre *sites* dentro das mensagens dos grupos de emails da Usenet e as categoriza e agrupa fornecendo informações sobre os sites mais populares dentro do grupo. Donath e seu grupo (VIEGAS e DONATH, 1999) garimparam em *chats* e em *newsgroups* e criaram visualizações destas conversas com o objetivo de extrair alguma informação que possa ser desejada.

Da mesma forma que os sistemas de suporte a recomendações, eles trabalham em situações onde as pessoas naturalmente possuem seus próprios papéis, alguns poucos produzem e compartilham suas opiniões e gostos enquanto a maioria somente sai a procura quando tem necessidade. Os sistemas podem preservar e transmitir a informação sobre o contexto da atividade sobre a qual as preferências foram extraídas, isto pode levar a recomendações mais úteis e informativas. Também cria oportunidades para construção de comunidades pois permite que os usuários possam ser colocados em contato com outros que compartilhem suas preferências. Contudo, ao contrário dos sistemas de suporte a recomendações, que auxiliam as pessoas que intencionam compartilhar as recomendações, sistemas de mineração de dados sociais extraem preferências de contextos onde os provedores possam não ter esta intenção. Assim, as oportunidades devem ser balanceadas em relação às questões de privacidade.

3.5.4 Filtragem colaborativa

A filtragem colaborativa é a tecnologia de sistema recomendações de maior sucesso na *web*. Ela explora técnicas estatísticas que possibilitam encontrar pessoas com interesses similares e então efetuar as recomendações. Este método baseia-se em três pontos: deve ter

muitas pessoas participando para que os usuários possam encontrar outros com gostos semelhantes; deve ser fácil para os participantes representarem seus interesses no sistema; e os algoritmos devem ser capazes de encontrar as pessoas com interesses parecidos. (SARWAR et al., 2000)

A tarefa do usuário torna-se muito simples, basta que ele expresse suas preferências avaliando alguns itens que o sistema apresenta. Estas avaliações servirão como uma aproximação de seu gosto neste campo. O sistema então busca dentro das avaliações fornecidas pelos demais usuários e calcula o resultado que será o conjunto dos "vizinhos mais próximos", que são as pessoas com gostos semelhantes. Finalmente o sistema recomenda itens melhores avaliados pelos vizinhos mais próximos que ainda não tenham sido avaliados pelo usuário e que provavelmente ele nem conheça. O usuário pode ainda avaliar estas recomendações recebidas e assim, o sistema vai adquirindo um melhor conhecimento das preferências do usuário e conseqüentemente, melhorando sua precisão.

Os primeiros sistemas de filtragem colaborativa foram o GroupLens (RESNICK et al, 1994), o Bellcore Video Recommender (HILL et al., 1995), e o Firefly (SHARDANAND e MAES, 1995). Estes sistemas apresentavam diferenças na forma que pesavam as recomendações, mudando as proximidades dos vizinhos e a forma que combinavam as avaliações.

A filtragem colaborativa encontrou muitas aplicações na internet. Sites de comércio eletrônico como a Amazon.com possuem recomendadores, onde além das avaliações de experts, os usuários também podem deixar suas avaliações e receber recomendações calculadas por um sistema de filtragem colaborativa. As preferências do usuário também são inferidas pelo uso do site, por exemplo, ao comprar um livro, o sistema pode coletar interesses sobre o assunto em outras áreas.

A principal força da filtragem colaborativa é que as recomendações são personalizadas. Se os vizinhos mais próximos realmente possuírem gostos semelhantes, o usuário pode obter recomendações que ache interessantes e que sozinho não encontraria. O segundo ponto forte deste método é que o usuário não precisa buscar as recomendações, elas são apresentadas pelo sistema quando ele informa o que gosta. Finalmente, sob o ponto de vista computacional, a implementação deste tipo de sistema é sob a forma de uma matriz de usuários e itens, com células contendo as avaliações o que a torna útil para várias manipulações computacionais.

A filtragem colaborativa exige que os papéis de avaliadores e usuários de recomendações sejam simétricos, ou que haja uma uniformidade de papéis. Pois para receber uma recomendação, um usuário deve antes fornecer seus gostos. Isso faz com que o conhecimento do sistema cresça e portanto, consiga fornecer melhores recomendações com o tempo. A uniformidade de papéis tem tanto um lado bom quanto ruim. Por um lado, na prática observa-se que a maioria das pessoas não quer dar recomendações, elas preferem apenas usá-las. Por outro lado, avaliar um item não é uma tarefa difícil e quem quer receber, precisa primeiro fornecer.

Finalmente a filtragem colaborativa separa o contato pessoal do processo de recomendação (HILL et al., 1995) pois não é necessário que o cliente de recomendações conheça o fornecedor daquela. No entanto, se o projetista do sistema desejar, os usuários podem conhecer os seus vizinhos mais próximos e isso pode ser uma boa técnica para a formação de comunidades com interesses comuns.

Há vários desafios na utilização da filtragem colaborativa, como no caso do primeiro usuário a efetuar uma avaliação não deve receber recomendações e quando houverem poucas avaliações, é difícil encontrar similaridade entre os usuários significando que os vizinhos não estão próximos e as recomendações, podem não ser tão boas. O problema pode-se tornar pior quando aumenta o número de itens.

Uma tática para resolver estes problemas é combinar a filtragem colaborativa com as recomendações baseadas em conteúdo extraindo informações implícitas das avaliações. Por exemplo, o sistema FAB (BALABANOVIC e SHOHAM, 1997) analisa o conteúdo das avaliações favoráveis para criar perfis dos interesses do usuário e então aplica a filtragem colaborativa para identificar outros usuários com interesses semelhantes.

Billsus e Pazzani (1998), observaram que a tarefa de prever um item que um usuário gostaria, baseando-se em avaliações de outros usuários, pode ser conceituada como uma classificação, que é uma técnica já bastante estudada pela comunidade que pesquisa sobre aprendizagem de máquina, seus algoritmos conseguiram reduzir a necessidade de usuários precisassem avaliar os mesmos itens antes que um possa servir como gerador de previsões para outro.

Aggarwal et al (1999), conseguiu evitar algumas limitações dos algoritmos tradicionais através de uma abordagem que utiliza a teoria dos grafos para a filtragem colaborativa. Esta abordagem consegue computar recomendações mais precisas com dados

esparsos.

Como qualquer sistema que oferece resultados com base em quantidades expressivas de processamento de dados, um sistema de filtragem colaborativa enfrenta o problema da explicação. Porque o sistema pensa que o usuário gostará daquele item? Herlocker (2000) propõe técnicas para explicar as recomendações calculadas por um sistema de filtragem colaborativa e também as experiências para avaliar a eficácia destas explicações.

Finalmente, uma questão importante é a noção de serendipidade que é definida como a capacidade de recomendar itens que o usuário ainda não saiba ou recomendações para itens que não possuam o mesmo conteúdo que o usuário estava procurando (HERLOCKER, 2000). Um sistema que simplesmente indique itens semelhantes mas que sejam relativamente óbvios pode não ter grande utilidade (TERVEN e HILL, 2001). Um exemplo seria um usuário que avaliou bem o livro Código da Vinci de Dan Brown e o sistema lhe apresenta recomendações somente os outros livros do mesmo autor .

3.5.5 Metodologia da filtragem colaborativa

A metodologia para a filtragem colaborativa é dividida em três tarefas: a representação das informações, a criação da vizinhança e a geração das recomendações (SARWAR et al., 2000). A representação trata da estrutura que será utilizada para armazenar os itens que já foram votados ou comprados pelo usuário, no caso do comércio eletrônico. A fase da formação da vizinhança foca o problema de como identificar os outros clientes próximos. Finalmente, a geração das recomendações é onde são calculados os *n* produtos mais adequados para aquele cliente. Esta seção descreve técnicas de efetuar estas tarefas.

3.5.5.1 Representação

Em um sistema de recomendações baseado em filtragem colaborativa típico, os dados são uma coleção de registros históricos de votos ou compras de *n* clientes para *m* itens representadas por uma matriz *n x m* (SARWAR et al., 2000). Apesar de uma matriz ser uma

representação simples, ela pode trazer problemas para os sistemas de recomendação baseados em vizinhos mais próximos. O primeiro problema enfrentado são os dados esparsos, muitos sistemas de recomendação possuem uma base de itens muito grande comparada ao número de itens comprados por cliente, tornando difícil estabelecer correlações. O segundo problema para a representação dos dados é a escalabilidade, que é a capacidade de expansão do sistema, muitos algoritmos de vizinhos mais próximos necessitam que o poder computacional aumente conforme aumenta o número de itens e clientes. O último desafio é a capacidade do sistema de reconhecer itens semelhantes, mesmo que tenham nomes ou códigos diferentes .

3.5.5.2 Formação de vizinhança

O cálculo da similaridade entre os clientes de forma a estabelecer a formação da vizinhança entre um cliente e outros com gostos semelhantes é o passo mais importante para os sistemas baseados em filtragem colaborativa. O processo de formação da vizinhança é a construção do modelo ou o processo de aprendizado de um sistema de recomendações. O objetivo principal do processo é encontrar para cada cliente, uma lista ordenada de outros clientes de tal forma que a similaridade entre eles seja máxima e decrescente.

A medida de proximidade entre dois clientes é geralmente dada pela correlação ou pelo cosseno. Pelo método de correlação, a proximidade entre dois usuários *a* e *b* é medida calculando-se a correlação Pearson (COOPER e SCHINDLER, 2003) obtida dividindo-se a covariância de duas variáveis pelo produto de seus desvios padrões (equação 1).

$$corr_{ab} = \frac{cov(a,b)}{\sigma_a \sigma_b} \tag{1}$$

No caso da medida da distância calculada pelo cosseno, os dois clientes devem ser vistos como vetores em um espaço n dimensional e a proximidade é o cosseno do ângulo destes vetores (2).

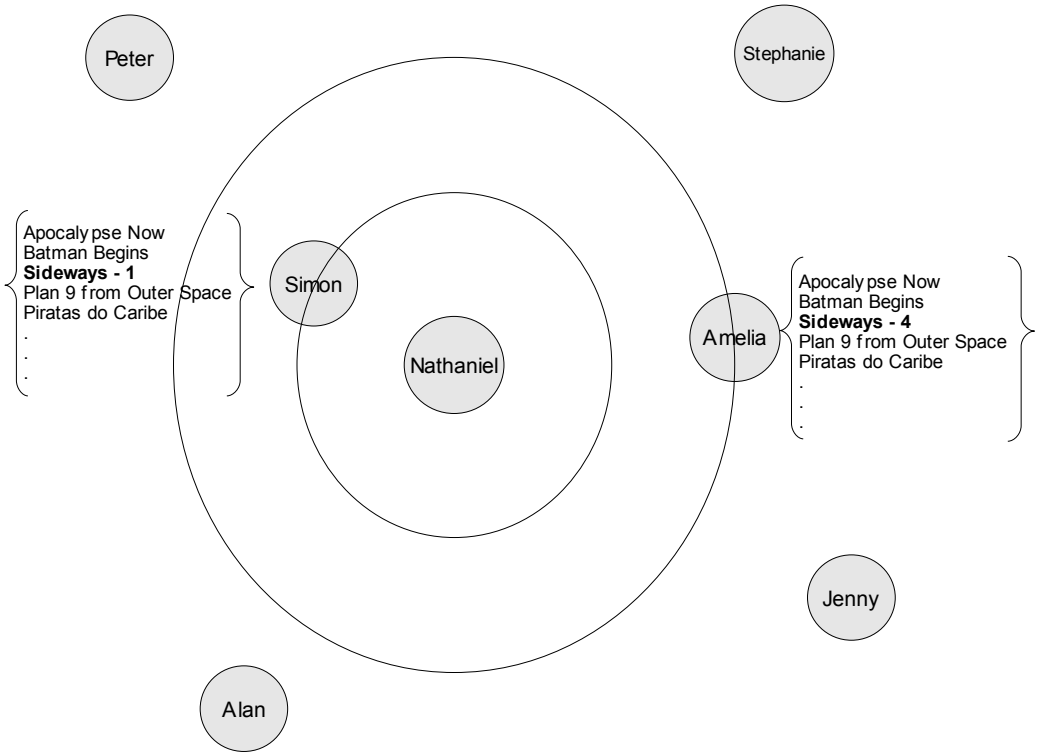
$$\cos(\vec{a},\vec{b}) = \frac{\vec{a}\cdot\vec{b}}{\|\vec{a}\| * \|\vec{b}\|} \tag{2}$$

Após calcular a proximidade entre todos os clientes, a tarefa é formar a vizinhança utilizando-se algum esquema que selecione e agrupe os clientes mais próximos daquele.

3.5.5.3 Geração das recomendações

O passo final é derivar as recomendações obtidas pela geração da vizinhança. Uma técnica é obter os itens mais freqüentes, para isso o sistema busca entre as compras dos vizinhos mais próximos, os produtos mais comprados entre aqueles ainda não comprados pelo cliente. Outra técnica é através de descoberta de regras de associação entre os produtos utilizando mineração de dados.

Figura 8: Efetuando predições através dos vizinhos mais próximos.



Simon está a uma distância de 2 unidades e deu uma avaliação de -1 para o filme Sideways, enquanto Amélia que está a uma distância de 4 unidades, avaliou o mesmo filme em -4.
Fonte: Adaptado de (BERRY e LINOFF, 2004) pelo autor.

A figura 8 ilustra a formação da vizinhança do usuário Nathaniel. Seus vizinhos mais próximos são o Simon e a Amelia, as avaliações destes vizinhos serão usadas para efetuar predições das avaliações do Nathaniel.

No exemplo, Simon está a uma distância de 2 unidades e deu uma avaliação de -1 para o filme Sideways, enquanto Amélia que está a uma distância de 4 unidades, avaliou o mesmo filme em -4. Uma estimativa para a avaliação de Nathaniel poderia ser feita ponderando-se as notas em relação as distâncias .

3.5.6 Algoritmos de filtragem colaborativa

Na seção anterior foram apresentados os conceitos e desafios da metodologia, esta seção estuda alguns dos algoritmos utilizados para a sua implementação. Em termos computacionais, a tarefa da filtragem colaborativa é predizer a utilidade dos itens para um usuário baseando-se em um banco de dados de avaliações de uma amostra ou população de outra base de dados de usuários. Para tanto, são examinadas duas classes genéricas de algoritmos, os baseados em memória, que atuam sobre toda a base de dados de usuários para fazer as predições e os baseados em modelo, que utilizam a base de dados de usuários para estimar ou aprender um modelo que será usado para fazer as predições.

Para a implementação do sistema, deve-se estabelecer a forma de obtenção das avaliações. Podem ser votos (ex: gostei, não gostei), notas (ex: 1 ruim a 5 excelente) de maneira que o usuário expresse sua opção de forma explícita e também de forma implícita pelo seu histórico de compras efetuadas ou pelo fato de seguir ou não um *link* apresentado e até mesmo pelo tempo que ele permanece na página. Independentemente da forma de obtenção dos dados, os algoritmos devem tratar também o problema de dados inexistentes pois muitas vezes não há nenhuma avaliação disponível para um item. Algumas das implementações podem utilizar avaliações implícitas, considerando que a existência de um registro no banco de dados signifique uma preferência positiva.

Os algoritmos são divididos em duas classes principais, os baseados em memória e os baseados em modelos (BREESE, 1998).

3.5.6.1 Algoritmos baseados em memória

Algoritmos baseados em memória foram os primeiros algoritmos colaborativos propostos para o problema de filtrar informações (RESNICK et al, 1994). Estes modelos baseiam suas predições no fato de que as pessoas geralmente acreditam nas recomendações de outras pessoas com idéias semelhantes e aplicam técnicas de vizinhos mais próximos para prever a nota do usuário corrente baseando-se nesta vizinhança.

A tarefa da filtragem colaborativa é prever o voto do usuário corrente utilizando-se um banco de dados de votos de usuários dentro de uma população ou amostra de outros usuários. O banco de dados consiste do conjunto de votos vi,j correspondendo ao voto (v) do usuário i para o item j. Se Ii é o conjunto de itens no qual o usuário i já votou, pode-se definir o voto médio como na equação 3.

$$\bar{v}_i = \frac{1}{|I_i|} \sum_{j \in I_i} v_{i,j} \tag{3}$$

Nos algoritmos baseados em memória, deve-se prever o voto do usuário ativo, indicado pelo subscrito *a*, baseando-se em alguma informação parcial sobre ele e um conjunto de pesos calculado utilizando-se as informações do banco de dados de usuários. É assumido que o voto previsto, $p_{a,j}$ do usuário ativo para o item *j*, é uma soma ponderada dos votos dos outros usuários (equação 4):

$$p_{a,j} = \bar{v}_a + \frac{\sum_{i=1}^{n} (v_{i,j} - \bar{v}_i) * w_{a,i}}{\sum_{i=1}^{n} w_{a,i}} \tag{4}$$

Sendo que, n é o número de usuários no banco de dados com pesos maiores que zero. Os pesos *w(a,i)* podem refletir a distância, a correlação ou a similaridade entre cada usuário *i* e o usuário ativo. Partindo-se deste ponto, a maior diferença entre os algoritmos baseados em memória é a forma de calcular as correlações entre os usuários e são listadas a seguir as mais comuns:

Coeficiente de correlação de Pearson (RESNICK et al, 1994). Mede o grau pelo qual uma relação linear existe entre duas variáveis, pode ser representada pela equação 5.

$$w_{a,i} = \frac{\sum_{i=1}^{m} (v_{a,j} - \bar{v}_a)(v_{i,j} - \bar{v}_i)}{\sqrt{\sum_{j} (v_{a,j} - \bar{v}_a)^2 \sum_{j} (v_{i,j} - \bar{v}_i)^2}} \tag{5}$$

Correlação de Spearman (HERLOCKER et al, 2004) Semelhante ao coeficiente Pearson, mas não necessita que a relação entre as variáveis seja linear. Calcula a medida de correlação entre os *ranks* no lugar dos votos. É representado pela equação 6.

$$w_{a,i}=\frac{\sum_{i=1}^{m}(\rho_{a,j}-\bar{\rho}_a)*(\rho_{i,j}-\bar{\rho}_i)}{\sqrt{\sum_j(v_{a,j}-\bar{v}_a)^2\sum_j(v_{i,j}-\bar{v}_i)^2}} \quad (6)$$

Similaridade de vetores (BREESE, 1998). Neste caso os pesos são calculados de forma semelhante à técnica de recuperação de informações calculando-se o cosseno do ângulo de dois vetores formados pela freqüência dos votos dos usuários de acordo com a equação 7.

$$w_{a,i}=\sum_j \frac{v_{a,j}}{\sum_{k\in I_a} v_{a,k}^2}\frac{v_{i,j}}{\sum_{k\in I_i} v_{i,k}^2} \quad (7)$$

Entropia (HERLOCKER et al, 2004). Este modelo utiliza técnicas de probabilidade condicional para medir a redução da entropia das avaliações do usuário ativo que resulta do conhecimento das avaliações dos outros usuários.

3.5.7 Um exemplo de Filtragem Colaborativa

O quadro 3 ilustra uma matriz de avaliações de itens *In* feitas por usuários *Un*, cada usuário pode ter ou não uma avaliação para os itens, ele será usado para exemplificar a predição da avaliação do Item 6 pelo usuário 4 aplicando-se a filtragem colaborativa.

Quadro 3: Matriz de avaliações de usuário para filtragem colaborativa

	I1	I2	I3	I4	I5	I6
U1	2	4		5	1	
U2	2	4	3			5
U3	4	2		5	1	2
U4	1	5		2	4	?

Fonte: Calderón-Benavides e González-Caro, 2003.

Para começar, é calculado o coeficiente de correlação Pearson entre o usuário corrente, o usuário U4 e os demais (8):

$$w_{4,3}=\sum_i \frac{(v_{4,i}-\bar{v}_4)(v_{3,i}-\bar{v}_3)}{\sqrt{(v_{4,i}-\bar{v}_4)^2(v_{3,i}-\bar{v}_3)^2}}=-0.8 \quad (8)$$

Os cálculos utilizam apenas os itens nos quais os os dois usuários tenham avaliações

comuns. Este cálculo é efetuado para todos os outros usuários da base com avaliações comuns ao usuário corrente. Para U1 o coeficiente resulta em 0 e para U2 o coeficiente resulta +1. Significando que U2 concorda com U4 e U3 cuja correlação é negativa (-0,8) discorda. O próximo passo é prever o voto de U4 para o item I6, para obter o resultado é efetuada a média ponderada das correlações conforme (9):

$$p_{4,6} = \bar{v}_4 + \frac{\sum_{u=1}^{n} (v_{u,6} - \bar{v}_u) * w_{4,u}}{\sum_{u=1}^{n} w_{4,u}} = 4{,}56 \quad (9)$$

O número 4,56 representa a estimativa do voto do usuário U4 para o item I6, ficando próximo ao voto do usuário U2 que é o mesmo que obteve a maior correlação.

3.6 ALGORITMOS BASEADOS EM MODELOS

Para a filtragem colaborativa baseada em modelos, deve ser construído um modelo das preferências do usuário e a partir deste que são inferidas as predições. Para a construção do modelo, cada usuário é representado como um vetor dos seus itens avaliados e neste vetor são aplicadas técnicas de aprendizagem de máquina, tal como a classificação, agrupamento ou métodos baseados em regras (SARWAR et al., 2001). Para apresentar os algoritmos baseados em modelos é interessante exemplificar o funcionamento.

Geralmente os dados são representados em forma de matriz esparsa, pois cada usuário avalia um pequeno subconjunto de itens disponíveis. A tarefa de predição pode ser vista como o preenchimento dos itens que faltam desta matriz. A construção do modelo para o usuário vai permitir predizer os valores da linha que o representa na matriz.

Quadro 4: Exemplo de matriz de avaliações de usuários e itens.

	I1	I2	I3	I4
U1	4		5	
U2		3		2
U3	5	1	5	1
AU	4	2	4	?

Fonte: Calderón-Benavides e González-Caro, 2003.

Para prever as avaliações do usuário ativo (AU), poderia-se treinar um algoritmo de aprendizagem de máquina com os dados de suas outras avaliações. No exemplo do quadro 4, o usuário ativo possui 3 exemplos de treino para os itens I1, I2 e I3. Usando a terminologia de aprendizagem de máquinas a matriz pode ser representada em forma de vetores de características (CALDERÓN-BENAVIDES e GONZÁLEZ-CARO, 2003) no qual os usuários correspondam aos vetores de características e os valores das suas avaliações, correspondem aos valores das suas dimensões.

A filtragem colaborativa pode ser encarada como uma tarefa de classificação. Tomando-se por base um um conjunto de avaliações de um usuário, para um item específico, um algoritmo de classificação ser usado para as predições separando em classes como por exemplo, gostou ou não gostou, estes algoritmos podem ser probabilísticos (CALDERÓN-BENAVIDES e GONZÁLEZ-CARO, 2003) que avaliam a probabilidade de um item pertencer à uma determinada classe, classificadores bayesianos e máquinas de suporte vetorial.

Uma outra abordagem para o desenvolvimento de modelos são os algoritmos de agrupamento, eles têm sido aplicado em filtragem colaborativa de duas maneiras: agrupamento de itens, como uma forma de pré-processamento e o agrupamento de usuários. Os algoritmos K-Means e maximização de expectativa são os algoritmos mais referenciados nesta abordagem.

A terceira abordagem também considerada para a geração dos modelos é o uso de é o uso de algoritmos de mineração de regras de associação. Estes algoritmos servem para descobrir os relacionamentos entre itens ou usuários baseando-se em padrões de ocorrência entre as transações. (CALDERÓN-BENAVIDES e GONZÁLEZ-CARO, 2003)

3.6.1 Slope One

Os algoritmos do tipo *Slope One*, segundo Lemire e Maclachlan (2005), provêem fácil implementação e manutenção, os dados podem ser atualizados dinamicamente trazendo as previsões imediatamente sem a necessidade de processamento *off line*, permitem consultas rápidas e sem consumo excessivo de memória, fornecem resultados válidos mesmo tendo poucas avaliações e finalmente, apresentam resultados competitivos em relação a outros

esquemas mais precisos.

Os algoritmos *Slope One* trabalham com o conceito de diferencial de popularidade entre itens ou usuários. Uma forma de medir este diferencial é simplesmente subtrair a média de avaliações de dois itens, esta diferença é usada para prever a avaliação de outro usuário para este item. No exemplo do quadro 5, o usuário A, avaliou dois itens I e J dando notas 1 e 1,5 respectivamente, o usuário B avaliou o item I com nota 2. O *Slope One* calcula o diferencial das avaliações do usuário A que é 1,5 – 1=0,5, observando-se que o item J é avaliado em 0,5 a mais que o item I, prediz-se que a avaliação do item J será 2 + 0,5 = 2,5.

Quadro 5: Slope One: As duas avaliações do usuário A para dois itens são usadas para prever a avaliação do Usuário B para o item J.

	Item I	Item J
Usuário A	1	1,5
Usuário B	2	? = 2 + (1,5 – 1) = 2,5

Fonte: Adaptado de Lemire e Maclachlan (2005) pelo autor.

3.7 SISTEMAS DE RECOMENDAÇÃO EM USO NAS EMPRESAS

Apesar da ampla utilização dos sistemas de recomendação nas grandes empresas de comércio eletrônico, não são apenas estas que lucram com eles. Algumas empresas tem seu modelo de negócio baseados em fornecer recomendações, como os buscadores, especialmente o Google, que obtém suas receitas com propaganda do tipo *links* patrocinados. O seu sistema utiliza análise de conteúdo para relacionar os textos dos resultados e das buscas com as palavras chave. Outro modelo é cobrar para recomendar os produtos, baseando-se em sistemas de suporte a recomendações, como o Buscapé[4] e UOL Shopping[5]. Esta seção descreve algumas empresas que fazem uso destas tecnologias.

3.7.1 Amazon.com

4 http://www.buscape.com.br
5 http://shopping.uol.com.br

É o exemplo comumente citado ao se falar de uso comercial sistemas de recomendação. A empresa, na última década, desenvolveu, aprimorou e patenteou seus sistemas de recomendação (LINDEN, SMITH e YORK, 2003), impondo uma barreira significativa aos concorrentes. Não surpreende que o sistema seja tão abrangente, todas as recomendações são baseadas na história das compras ou navegação do indivíduo mais o próprio item e ainda nas preferências de outras pessoas que compraram na Amazon. Tudo estudado para conduzir o comprador ao pedido. A figura 9 ilustra um exemplo de recomendações exibidas ao consultar um produto na Amazon.com.

Para obter as recomendações oferecidas na figura 9, o sistema utiliza um algoritmo para determinar os itens mais semelhantes buscando aqueles que os usuários costumam comprar em conjunto construindo uma matriz de produto x produto e calcula uma medida de similaridade para cada par.

Figura 9: Exemplo de recomendações na Amazon.com.

Fonte: Amazon.com.

3.7.2 Snooth

O Snooth[6] é um sistema de busca de vinhos e plataforma de recomendações que gera recomendações personalizadas de vinhos utilizando os dados das avaliações deixadas pelos usuários sobre os vinhos que tenham provado. O sistema é composto por uma base de usuários, vinhos e suas avaliações que além de registrar as avaliações da comunidade de participantes, fornece recomendações personalizadas sobre os vinhos. Seu banco de dados contém mais de 300 mil vinhos e 2 milhões de avaliações de usuários e críticos profissionais.

6 http://www.snooth.com

Seu algoritmo proprietário leva em conta a avaliação, a quantidade de avaliações e o nível de confiança do avaliador. (SNOOTH, 2008).

3.7.3 Netflix

Depois dos livros, os filmes são uma outra área em que os sistemas de recomendação tem sido utilizados. Locadoras de vídeos usam o recurso para incrementar suas conversões. Uma delas é a locadora de vídeos *online* americana Netflix[7] que faturou próximo a um bilhão de dólares em 2006 (NFLX, 2008). Na Netflix, ao devolver um filme, o usuário recebe outras recomendações de filmes fornecidas por um sistema que utiliza as informações do seu histórico. O objetivo é agradar o cliente de forma que ele repita o ciclo indefinidamente.

A empresa criou um desafio em outubro de 2006: quem conseguir melhorar em 10% o seu sistema de recomendações de filmes receberá um prêmio no valor de um milhão de dólares. O prêmio chamou a atenção de pesquisadores e contava com mais de 30 mil times participantes no início de 2008. A competição continua até outubro de 2011, alguns resultados já apareceram (NFLX, 2006).

O desafio utiliza um sistema de benchmark da própria Netflix chamado de *root mean square error* (RMSE) que é em essência o quanto uma predição erra a avaliação real. Quando iniciou a competição, o seu sistema de recomendações chamado de Cinematch possuía um RMSE de 0.9525, isso significa que as suas predições ficavam quase um ponto distante da avaliação real do usuário, para uma escala de 1 a 5 pontos isso não parece ser muito impressionante. Para ganhar o milhão, o time deve baixar o RMSE para 0.8572. (ELLENBERG, 2008)

Um bom sistema de recomendações pode fazer a diferença para qualquer negócio *online*. Quando um consumidor sabe exatamente o que está procurando, ele busca o produto, serviço ou informação. Quando ele não sabe exatamente o que deseja, ele navega e quando navega é que o sistema de recomendação vai fazer a diferença pois ele vai estar mais suscetível a sugestões. Ao mostrar algo que seja atraente para o visitante, o site vai incrementar as chances de venda ou de atrair e manter sua atenção.

7 http://www.netflix.com

3.8 LEIS DE POTÊNCIA

As leis de potência, no qual estão incluídas as distribuições do tipo long tail, de Pareto e as leis de Zipf (CLAUSET, SHALIZI e NEWMAN, 2007; NEWMAN, 2006) são distribuições que ocorrem tanto em fenômenos naturais, por exemplo, o tamanho das crateras lunares e a potência dos terremotos, em que somente alguns poucos são muito grandes e a grande maioria são muito pequenos, quanto em fenômenos humanos, como a distribuição da riqueza pessoal, em que existem poucos bilionários e muitos pobres. Ao contrário das distribuições do tipo normais ou gaussianas, os eventos raros apresentam um impacto maior que o esperado e poucos produtos, pessoas e sites parecem receber uma grande fatia do mercado, riqueza e atenção.

As leis de potência apresentam uma forma geral de acordo com (10), onde *C* é uma constante, *x* é a freqüência e *k*, é o expoente da lei de potência (NEWMAN, 2006).

$$p(x) = Cx^{-k} \tag{10}$$

Outra propriedade importante das leis de potência, é que elas são sem escala, pois não importa qual região da distribuição, a proporção entre os eventos pequenos e grandes permanece sempre a mesma, isto é, se for plotado um gráfico log-log, de qualquer região, a inclinação da curva será constante.

A distribuição de Pareto, criada por Vilfredo Pareto, economista que estava interessado em estudar a distribuição de renda, sua abordagem foi calcular quantas pessoas possuíam um rendimento maior que *x*. A lei de Pareto é dada em termos da de uma função de distribuição acumulada, em que um certo número de eventos X maiores que uma ocorrência *x* são uma potência inversa desta ocorrência *x* conforme (11) (ADAMIC, 2002).

$$Pr(X > x) = \left(\frac{m}{x}\right)^{k} \tag{11}$$

O seu parâmetro de inclinação *k*, pode ser obtido pelo método dos mínimos quadrados utilizando regressão linear simples ou pela fórmula (12) (NEWMAN, 2006).

$$k = 1 + n\left[\sum_{i=1}^{n} \ln \frac{x_i}{x_{min}}\right] \quad (12)$$

onde $x_i, i = 1..n$ são os valores medidos de x e x_{min} é o menor valor que x pode assumir, e o seu desvio padrão, é obtido pela fórmula (13) (NEWMAN, 2006).

$$\sigma = \frac{k-1}{\sqrt{n}} \quad (13)$$

3.8.1 Curva de Lorenz e Coeficiente Gini

Uma forma de descrever a concentração das transações é utilizar a Curva de Lorenz e o Coeficiente de Gini. Os economistas os têm usado em pesquisas sobre desigualdade econômica e concentração de renda. Brynjolfsson, Hu e Simester (2007) afirmam terem sido os primeiros a utilizá-los como forma de medir a concentração das vendas.

A Curva de Lorenz é desenhada dentro de um quadrado em que o eixo X mede o percentual acumulado de indivíduos e o eixo Y mede o percentual acumulado de riqueza. O Coeficiente de Gini representa a proporção entre a área da Curva de Lorenz e a área sob uma linha de 45 graus e a área total sob esta linha. O cálculo do Coeficiente de Gini pode ser obtido pela equação 14 (ELBERSE, 2007) :

$$Gini = 1 - \frac{2\sum_{i=1}^{n}(n+1-i)popularidade_i}{(n+1)\sum_{i=1}^{n} popularidade_i} \quad (14)$$

onde n é o número de valores medidos e $popularidade_i$ é a quantidade de vendas do item de acordo com sua curva de vendas.

Se uma empresa possuísse n produtos à venda e todos eles vendessem igualmente, a sua curva de vendas coincidiria perfeitamente com esta diagonal e a área entre ela e a curva de vendas seria zero e seu Coeficiente Gini também seria zero. No caso de uma empresa que possuindo muitos itens, vendesse somente um deles, o Coeficiente Gini seria igual a 1, representando a desigualdade perfeita.

3.9 TRABALHOS RELACIONADOS

Em *Frictionless Commerce*, Brynjolfsson e Smith (2000) analisam empiricamente as características da internet como canal de vendas para duas categorias de produtos, os livros e CDs e encontram preços entre 9% e 16% menores e 100 vezes menos remarcações de preços, concluindo que apesar de haver menor atrito econômico em muitas dimensões da competição no comércio tradicional, a marca e a confiança continuam sendo importantes fontes de diferenças entre as lojas virtuais.

Brynjolfsson, Hu e Smith (2003), estimaram o impacto econômico causado pelo aumento da variedade de produtos que os mercados eletrônicos tornaram disponíveis. Enquanto os ganhos de eficiência advindos do aumento da competição levam a menores preços, sua pesquisa mostra que o aumento da variedade de produtos levam a um aumento maior dos excedentes do consumidor.

Ghose e Gu, (2006) estudaram o impacto dos custos de busca na competitividade das empresas usando para isso uma metáfora teórica de que os custos de busca criam uma imperfeição na demanda agregada quando as empresas mudam seus preços. Sua pesquisa conclui que os custos de busca influem na elasticidade dos preços das empresas e que estes diminuem com a passagem do tempo a medida que as alterações de preço se dissipam entre os consumidores, levando a uma maior elasticidade de preços.

Em *Goodbye Pareto, Hello Long Tail*, Brynjolfsson, Hu e Simeste (2007), analisaram o princípio de Pareto também conhecido como regra 80/20 na concentração das vendas sob a ótica da redução dos custos de busca trazidos pela tecnologia e mercados na internet, e encontram evidências de que consumidores com experiência prévia de compra possuem maior tendência de experimentar produtos obscuros do que os que não possuem tal experiência, o contrário do que ocorre em compras por catálogos impressos.

Elberse (2007), analisou 20 milhões de transações de locações *online* avaliando o padrão de consumo para itens obscuros e mais vendidos, encontrou que uma grande parte dos consumidores, especialmente entre os que compram com uma freqüência maior e concentram em um grupo estreito de gêneros, regularmente optam por itens obscuros da cauda que

dificilmente estão em lojas físicas. Contudo, mesmo para estes consumidores, os itens mais vendidos são a maior parte dos seus gastos e são os mais apreciados.

Tucker e Zhang (2007) pesquisaram em sites que recomendam produtos *online* e encontraram evidências de que existe também um efeito *steep tail*, uma concentração de vendas para os produtos mais vendidos especialmente quando os consumidores observam a quantidade de acessos de outros consumidores. Eles questionam se este efeito complementa o *long tail* por atrair consumidores que de outra forma não chegariam até aquela loja e encontram evidências que este efeito dos produtos mais vendidos pode ser utilizado como uma poderosa ferramenta de marketing para atrair os consumidores e aumentar a venda.

Hevas-Drane (2007), apresentou um modelo para explicar o quanto os sistemas de recomendação explicam o *long tail*. E conclui que, como os consumidores buscam opiniões de outros consumidores via boca-a-boca, a introdução de um sistema de recomendações age como um intermediário entre os consumidores e incrementa os lucros e as concentrações nas vendas.

Elsberse e Oberholzer-Gee (2007), observaram que a indústria e a academia discordam muito sobre como a distribuição *online* vai alterar a quantidade e a variedade de produtos que os consumidores compram. Enquanto os proponentes do *long tail* argumentam que um incremento agudo na oferta de produtos através de canais *online* vai dirigir o consumo para fora dos mais vendidos para um grande número de itens de nicho menos vendidos, a teoria dos superastros, prevê o oposto, a medida que os consumidores tenham acesso aos seu conteúdo favorito, os padrões de consumo se tornarão mais uniformes. Em seu estudo, utilizando dados de vídeo locadoras, os resultados mostram que há um incremento entre os títulos menos vendidos ao mesmo tempo que ocorre uma maior concentração entre os mais vendidos.

Fleder e Hosanagar (2007), examinaram o efeito dos sistemas de recomendação na diversidade das vendas utilizando para isso a modelagem e seu teste computacional. Eles obtiveram quatro resultados: 1) os sistemas de recomendação comuns, como os de filtragem colaborativa, reduzem a diversidade das vendas pois não conseguem recomendar produtos com histórico limitado e podem criar o efeito ricos ficam mais ricos para os produtos mais populares e vice-versa para os mais obscuros, 2) a diversidade a nível individual aumenta, ao passo que a diversidade agregada diminui pois eles levam todos os clientes para os mesmos novos produtos, 3) os sistemas de recomendação aumentam a possibilidade de criação de sucessos por acaso e 4) sugerem decisões de projeto que afetam os resultados de forma que os

gerentes possam escolher um recomendador que seja adequado a sua estratégia de diversificação de produtos.

4 ESTUDOS DE CASO

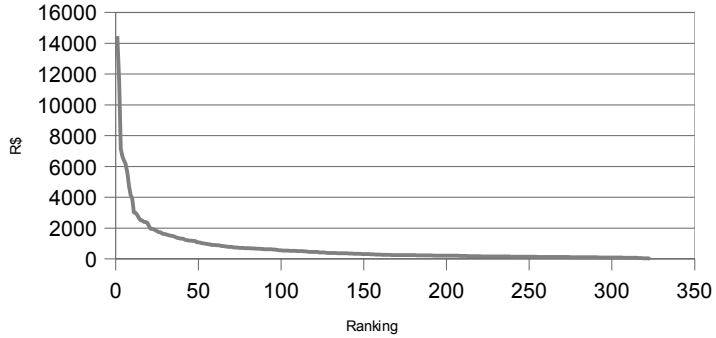
Para o desenvolvimento da pesquisa, foram efetuados dois estudos de caso: o primeiro envolve a análise das vendas da empresa de comércio eletrônico sob a ótica do *long tail*. Esta análise se dá em duas etapas, a primeira estuda se a curva de vendas possui formato de lei de potência, a segunda etapa, compara as curvas de vendas da empresa com dados de produtos semelhantes de outra empresa que não vende pela internet.

No segundo estudo de caso, são testados 4 diferentes sistemas de recomendação. A análise é feita em relação à profundidade de navegação dos usuários que seguem estas recomendações e também quais tipos possuem melhor aceitação pelos visitantes do site. Ao final do capítulo, os resultados são discutidos e as conclusões apresentadas.

4.1 ESTUDO DE CASO 1

A primeiro estudo de caso consiste em avaliar a distribuição das vendas de uma empresa de comércio eletrônico de vinhos para verificar se esta segue uma lei de potência e comparar esta distribuição com uma outra empresa não virtual. A VinhosNet trabalha somente com comércio eletrônico de vinhos da serra gaúcha atendendo todo país desde 1999. Possui em seu catálogo 699 rótulos de vinhos finos, espumantes e sucos e uva e não mantém

Gráfico 1: Vendas VinhosNet em 2007 apresenta forma de lei de potência.

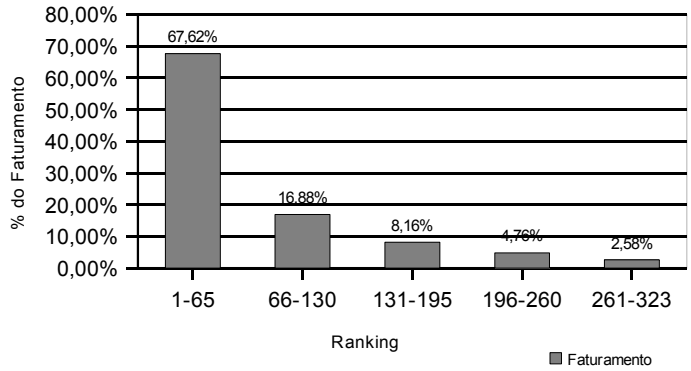


Fonte: Elaborado pelo autor.

praticamente nenhum estoque destes produtos que são adquiridos sob demanda dos seus 50 fornecedores em sua maioria da região. O gráfico 1 mostra a distribuição das vendas da VinhosNet em relação ao ranking dos produtos, que apresenta a forma de lei de potência.

Para as análises, foram obtidos os dados sobre as vendas no ano de 2007. Para efeitos de comparação, foram utilizados dados de vendas de vinhos de outra uma empresa local que não possui venda *online,* o Supermercado Apolo.

Gráfico 2: Ranking x Faturamento VinhosNet, 2007.



Fonte: Elaborado pelo autor.

4.1.1 Obtenção dos Dados

Para a VinhosNet, os dados das vendas de 2007 foram selecionados pelos pedidos do ano de 2007 agrupados pelo código do produto. Foram selecionados apenas o total vendido do item, em reais, a quantidade total vendida e a freqüência das vendas, que é a contagem do número de ocorrências de vendas de cada produto.

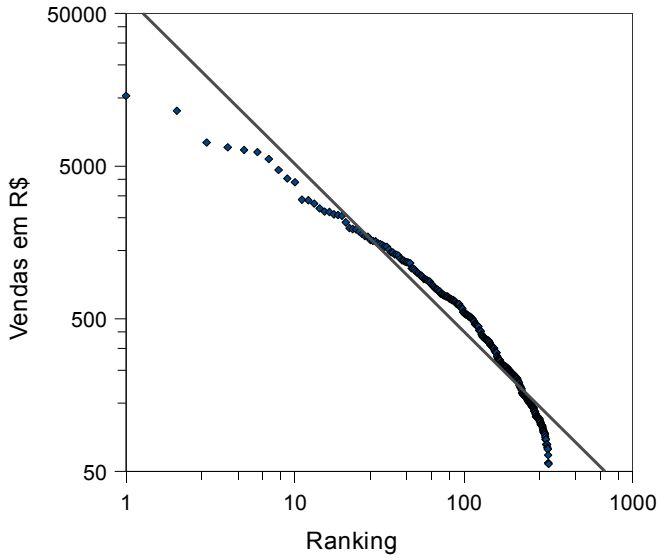
Tabela 1: Ranking x Faturamento em percentual VinhosNet, 2007.

Ranking	% do Faturamento
1- 65	67,6%
66-130	16,9%
131-195	8,2%
196-261	4,8%
262-323	2,6%

Fonte: Elaboração própria a partir de dados fornecidos pela empresa.

Como resultado, obteve-se 323 registros. A primeira observação é o fato de que somente 46% dos produtos do catálogo foram vendidos que levanta a questão sobre os custos de manter um catálogo *online.* Para o caso da empresa em questão eles variam em função do espaço ocupado em disco no seu servidor, o tráfego de dados pelas consultas dos produtos não vendidos, o custo de propaganda, o custo do comunicação e mão de obra para consultar os fornecedores e efetuar a atualização dos dados e imagens dos produtos do seu catálogo. A empresa não possui dados ou estudos sobre estes custos, mas uma pista é que uma única pessoa consegue manter o catálogo atualizado.

Gráfico 3: Diagrama de dispersão e reta de regressão em escala logarítmica para vendas da VinhosNet em 2007.



Fonte: Elaborado pelo autor.

A tabela 1 e o gráfico 2 apresentam o resultado da comparação do total de vendas em relação à quantidade de itens vendidos, observa-se que os 20% dos produtos mais vendidos representam 67,62% do faturamento. Se forem considerados os produtos não vendidos do catálogo, esta proporção muda bastante como pode-se ver na tabela 2 em que os 20% mais vendidos representam 86% do total de vendas e correspondem aos 140 primeiros produtos. A empresa de comércio tradicional apresenta suas vendas em proporções semelhantes a Pareto com 20% dos produtos, os primeiros 95 do ranking, representando 79% das vendas.

Utilizando a fórmula (15), onde $x_i, i = 1..n$ são os valores medidos de x e $x_{min} = ١$, resulta em k=1,41 e o seu desvio padrão pela fórmula (16) resulta em $\sigma = ٠,٠٢٣١$ para n=323.

$$k = 1 + n\left[\sum_{i=1}^{n} \ln \frac{x_i}{x_{min}}\right] \quad (15)$$

$$\sigma = \frac{k-1}{\sqrt{n}} \quad (16)$$

O diagrama de dispersão (3) das vendas e sua respectiva reta de regressão obtida pelo método dos mínimos quadrados (DOWNING e CLARK, 1998) de acordo com as fórmulas (17) e (18). Neste método obteve-se o coeficiente angular $\hat{m} = -1.097$ e o intercepto $\hat{b} = 64.390$ com ajuste r² = 0,931 indicando um ajuste forte.

$$\hat{m} = \frac{\overline{xy} - \bar{x}\,\bar{y}}{\overline{x^2} - \bar{x}^2} \quad (17)$$

e

$$\hat{b} = \bar{y} - \hat{m}\,\bar{x} \quad (18)$$

Tabela 2: Ranking x faturamento VinhosNet considerando total de itens do catálogo em 2007.

Ranking	% sobre o total de itens	% do faturamento
1-140	20%	86,1%
141-699	80%	13,9%

Fonte: Elaboração própria a partir de dados fornecidos pela empresa.

Anderson, (2006a) afirma que não se deve avaliar a concentração das vendas utilizando-se o percentual de vendas em relação ao total dos itens do catálogo, deve-se comparar pelo número absoluto de itens de uma eventual loja equivalente de tijolos. Refere-se ao fato de que a concentração percentual dos itens aparenta ser muito maior nas lojas virtuais, mas eles representam de forma absoluta um número muito menor de itens pois os catálogos de lojas virtuais são muito maiores. Pois o efeito de um incremento maciço na quantidade de títulos sempre fará com que as vendas aparentem maior concentração em termos percentuais.

Tabela 3: Ranking x faturamento Apolo em 2007.

Ranking	% sobre o total de itens	% do Faturamento
1- 95	20%	79,0%
96-477	80%	21,0%

Fonte: Elaboração própria a partir de dados fornecidos pela empresa.

O quadro exemplifica esta situação: uma loja com 100 itens, se os 10 primeiros representarem 50% das vendas, significa que 10% dos itens são responsáveis por 50% das vendas, enquanto em uma outra loja com 10000 itens, os 10 primeiros poderiam representar 1% das vendas, analisando-se em termos percentuais precisaríamos de 1000 itens e se eles representassem 80% das vendas, a concentração pareceria muito maior, mas em termos absolutos estariam muito mais espalhados, então a concentração dos produtos mais vendidos ficará menor pois a variedade é muito maior.

Tabela 4: Comparativo ranking absoluto x relativo

	Vendas	
Ranking	Loja A Física	Loja B Virtual
	Ranking absoluto	
1- 10	50%	1%
11 e mais	50%	99%
	Ranking relativo	
Primeiros 10%	50%	80%
Restantes 90%	50%	20%

Fonte: Elaborado pelo autor.

Comparando as vendas da VinhosNet com a loja física (tabela 5), não se percebe este efeito. O principal motivo é que entre as duas empresas não há diferença em termos de magnitudes na variedade de produtos oferecidos como no caso de livros e música. Mesmo que o número de itens oferecidos da loja virtual seja maior, o número de itens efetivamente vendidos foi menor e a concentração absoluta que aparece é maior na loja virtual.

Tabela 5: Comparativo dos rankings absolutos entre VinhosNet e Apolo.

Ranking	VinhosNet	Apolo
1- 10	31,2%	25,7%
11-100	47,3%	54,2%
100 e mais	21,5%	20,1%

Fonte: Elaboração própria a partir de dados fornecidos pelas empresas.

A variedade de produtos ofertados pelo comércio eletrônico e pelo comércio físico de vinhos, não se apresenta na mesma proporção que produtos como livros, música e filmes. De acordo com levantamento efetuado em lojas virtuais de vinho, mesmo a wine.com[8] e a Mistral[9], que estão entre as maiores no comércio virtual de vinhos, apresentam em seu catálogo, uma quantidade próxima de 1.000 rótulos.

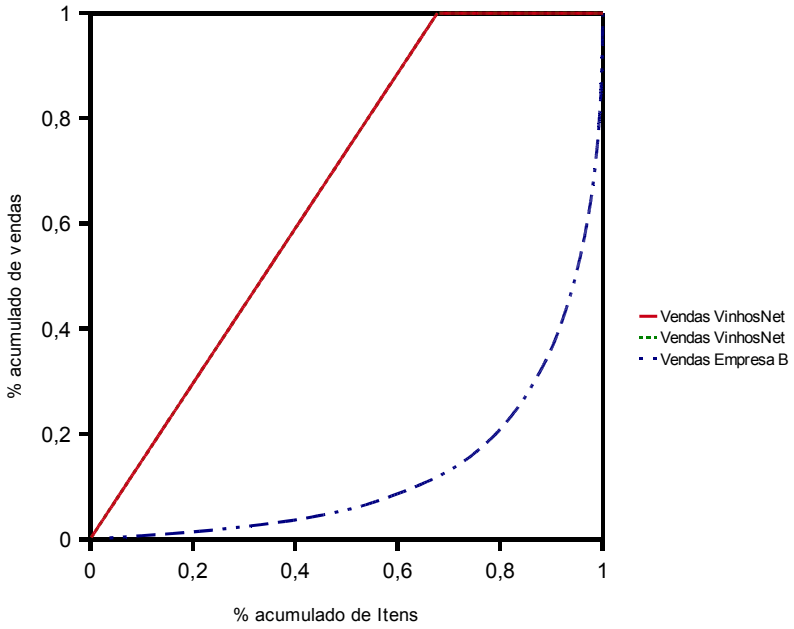
4.1.2 Coeficiente de Gini e Curva de Lorenz

Para o estudo de caso, foram desenhas as Curvas de Lorenz e calculados os Coeficientes de Gini como forma de comparar a concentração das vendas para as empresas participantes do estudo de caso, a VinhosNet e o Apolo e o resultado é apresentado no gráfico 4 que exibe as diferenças entre as distribuições de vendas entre a empresa de comércio online e a outra empresa de tijolo e cimento.

Ambas as curvas que representam as vendas de cada empresa, estão afastadas da reta diagonal, mas a de comércio virtual, está menos afastada e apresenta um índice menor de desigualdade. O cálculo do Coeficiente Gini para a VinhosNet resultou em 0,597, enquanto para o Apolo, resultou em 0,643. Este comportamento se apresenta de acordo com a teoria do l*ong tail*, que prevê uma distribuição mais homogênea das vendas *online* (ANDERSON, 2006).

8 Acessível em www.wine.com
9 Acessível em www.mistral.com.br

Gráfico 4: Curva de Lorenz comparando VinhosNet e Apolo em 2007.



Fonte: Elaborado pelo autor.

4.2 ESTUDO DE CASO 2

O segundo estudo de caso baseia-se em avaliar a forma que os usuários utilizam as recomendações e estudar como elas levam o visitante para a cauda. Este estudo baseia-se em dados fornecidos por um outro site especializado em registrar e compartilhar avaliações de vinhos e produtos relacionados. Ele deve sugerir recomendações de produtos baseando-se em avaliações de usuários. O trabalho consistiu em implementar estes sistemas de recomendação automatizados e acompanhar seu uso por meio de registros de acessos.

A Enoteca é um misto de comunidade, banco de dados de avaliações e sistema de suporte a recomendações conforme descrito no capítulo 4. Baseada no conceito Web 2.0 (O'REILLY, 2005), a comunidade reúne enófilos do país inteiro e conta mais de 1.000 usuários cadastrados, 1.100 produtos e 1.300 avaliações. O sistema possibilita que seus usuários mantenham um banco de dados com o registro de suas avaliações. Estas avaliações por serem públicas, podem ser consultadas por outros membros da comunidade e visitantes do *site*.

Figura 10: Página do vinho na Enoteca com formulário para avaliação de produto.

Fonte: www.enoteca.com.br

Para o registro das avaliações, a Enoteca possibilita duas formas: pode ser através de uma ficha de avaliação ou simplesmente um comentário, em ambas o usuário deve informar uma nota de 50 a 100 conforme a escala fornecida (figura 10). Se for utilizada a ficha de degustação, a avaliação do vinho envolve analisar suas características visuais, olfativas e gustativas.

4.2.1 Implementação

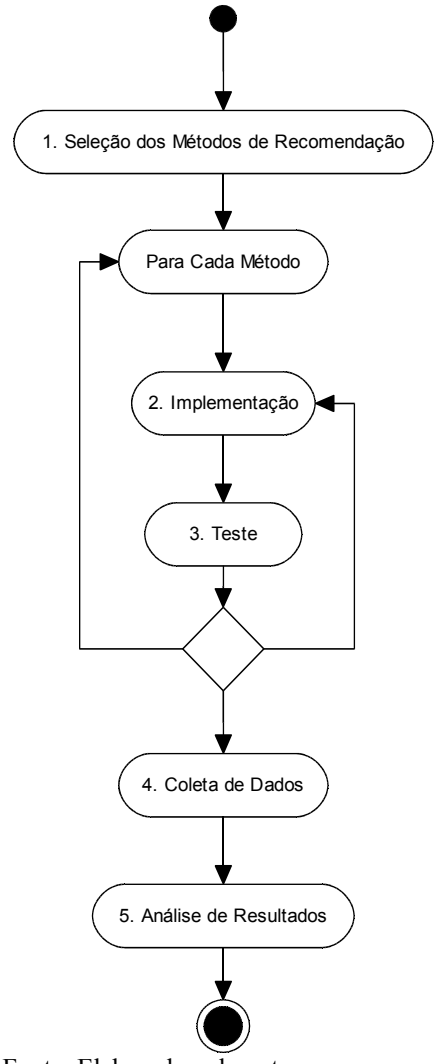
Este estudo de caso, consistiu em adaptar sistemas de recomendação conforme estudados no capítulo 3 e acompanhar a sua utilização através dos registros de acesso. Os sistemas de recomendação implementados são: 1) pela demografia, os vinhos são recomendados por possuírem o mesmo produtor, mesmo tipo de uva ou serem produzidos na mesma região; 2) por avaliações de outros usuários, neste caso, outras avaliações do mesmo usuário são usadas como recomendações; 3) por filtragem colaborativa utilizando-se o algoritmo Slope One (LEMIRE e MACLACHLAN, 2005) e 4) recomendações aleatórias, alguns produtos são sorteados e exibidos como recomendação.

A implementação seguiu de acordo com o diagrama da figura 11. Na primeira etapa foram selecionados os métodos de recomendação. O critério para a seleção foi a 1) disponibilidade da informação necessária para sua implementação, como no caso da recomendação demográfica, o registro do vinho possui as informações sobre a origem, como a vinícola, a uva utilizada na elaboração e a região produtora de origem, 2) o grau de dificuldade para a implementação, pois o desenvolvimento de um algoritmo específico poderia levar mais tempo do que o disponível para o projeto e 3) o desempenho do algoritmo, pois alguns algoritmos que exigissem muito consumo de processador e memória poderiam

inviabilizar o próprio sistema.

A seleção resultou em três categorias de recomendações sendo que a primeira é a recomendação conforme as características do item ou demográfico, ele possui os critérios de disponibilidade de informação, simplicidade de implementação e desempenho aceitável. O método seguinte selecionado foi a recomendação de outras avaliações registradas pelo usuário. Um mesmo usuário pode registrar quantas avaliações ele desejar e estas serão usadas como recomendações. Este método atende parcialmente aos critérios da seleção pois alguns usuários possuem apenas uma avaliação e neste caso elas não seriam exibidas. Quanto ao

Figura 11: Diagrama de atividades do estudo de caso 2.



Fonte: Elaborado pelo autor.

desempenho e a simplicidade de implementação, ele apresenta o mesmo nível de dificuldade do método demográfico.

O terceiro método selecionado foi a filtragem colaborativa com o algoritmo S*lope*

One. Ele atende aos critérios de disponibilidade de informação, pois sua exigência é uma matriz de avaliações numéricas de produtos para cada usuário. Na Enoteca o usuário deve avaliar o produto com uma nota de 50 a 100 baseada na escala de Robert Parker (PARKER, 2001). Como existem algoritmos com exemplos de implementações (LEMIRE e MCGRATH, 2005; O'SULLIVAN, 2006) e bibliotecas de código aberto testadas (TASTE, 2005; VOGOO, 2005), isso fez com que o critério de dificuldade de implementação fosse atendido tornando possível a sua adaptação para o projeto.

O *Slope One* também atende ao critério de desempenho pois sua utilização de recursos do sistema é moderada, suas tabelas podem ser atualizados dinamicamente a medida que as avaliações são efetuadas já trazendo as previsões, as consultas são rápidas e podem ser efetuadas pelos sistemas de banco de dados comerciais em uso e não apresentam consumo excessivo de memória e fornecem resultados válidos mesmo com poucas avaliações (LEMIRE e MACLACHLAN, 2005).

Finalmente, foi implementado um quarto método na qual as recomendações são sorteadas por um algoritmo que busca todos os produtos do banco de dados e seleciona cinco deles de forma aleatória. Como se espera que os resultados não sigam qualquer padrão para a escolha das recomendações, este método servirá para comparações.

Os passos seguintes consistiram em escrever e testar o código para o fornecimento das recomendações pelo *site*. Para cada um dos métodos, além do algoritmo para a seleção dos itens, foi necessário gerar a saída em tela com a assinatura necessária para o registro das leituras quando ela é selecionada pelo visitante. Esta assinatura contém o tipo de recomendação seguido e o nível pelo qual o visitante está navegando. Por exemplo, antes da implementação, um item qualquer possuía em sua URL[10] apenas o código do item, ex: www.enoteca.com.br/?id=1234 após a implementação, as URLs foram configuradas para permitir o registro das recomendações de acordo com o quadro 6.

Quadro 6: Formato da URL após a implementação do registro de recomendações

www.enoteca.com.br/ domínio	?id=9999 item	&nivel=x nível	&rcm=t tipo

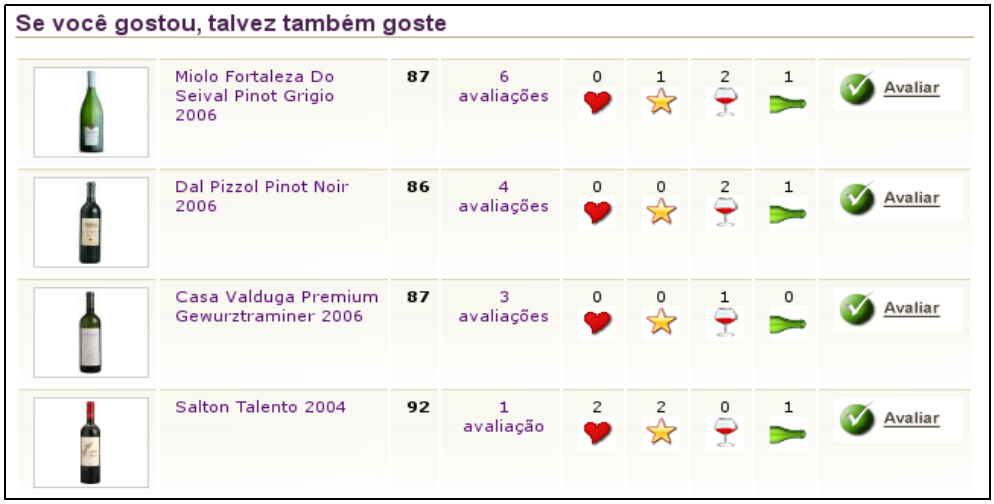
Fonte: Elaborado pelo autor.

Os registros são coletados a medida que os visitantes anônimos e usuários do *site* navegam pelas páginas. Toda vez um usuário acessa uma recomendação, é gravado um

10 URL – Uniform Resource Locator. endereço de internet.

registro contendo o item original consultado, o item selecionado pela recomendação, o tipo de recomendação escolhido e o nível da navegação. A cada nova seleção, o nível é incrementado de um, desta forma, é possível contar a profundidade da navegação do visitante.

Figura 12: Recomendações oferecidas ao consultar um vinho na Enoteca.



Fonte: www.enoteca.com.br

O período de coleta de dados durou dois meses e ocorreu durante o período de dezembro de 2007 a janeiro de 2008. A figura 12 é um exemplo de como as recomendações são apresentadas.

4.2.2 Resultados

Ao todo foram registradas 20.132 leituras de recomendações seguidas. Este número merece as seguintes considerações: somente são contadas as visualizações a partir do segundo nível, isto é, ao visualizar um produto, o visitante do *site* deve antes encontrar o produto, isso pode ser através do campo de busca ou pelos recomendados na página de entrada. Isso significa que ao selecionar uma recomendação, o usuário obrigatoriamente já consultou outro item, desta forma, visitantes que acessaram por acaso ou percebem que o conteúdo não é de seu interesse, já saem sem prosseguir. Esta situação é medida pelo índice taxa de rejeição que apresentou o percentual de 58,15% no período.

Os dados obtidos estão apresentados na tabela 6. A partir do oitavo nível os resultados tornam-se escassos e são agrupados. O método que apresentou o maior número de leituras foi

o demográfico seguido pelo aleatório, depois as avaliações de outros usuários e por último a filtragem colaborativa. Pelo percentual acumulado, 74,96% das leituras são até o segundo nível e até o quarto nível já atinge 94,01%. A partir do quinto nível em diante é atingido por menos de 5% dos visitantes. A média ponderada do número de níveis por usuários é 2,07.

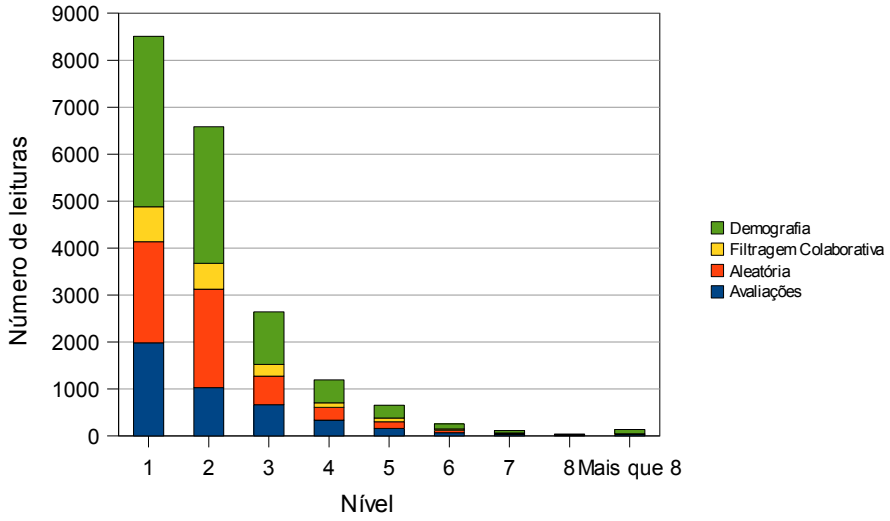
Tabela 6: Dados das leituras registradas por tipo de recomendação e nível resultantes do acompanhamento de recomendações no *site* Enoteca no período de dezembro de 2007 a janeiro de 2008.

Nível	Avaliações de usuários	Filtragem Colaborativa	Demografia	Aleatória	Total	% Sobre Total	% Acumulado
1	1983	745	3629	2150	8507	42,26%	42,26%
2	1031	550	2912	2091	6584	32,70%	74,96%
3	666	247	1119	608	2640	13,11%	88,07%
4	337	97	488	273	1195	5,94%	94,01%
5	159	82	271	141	653	3,24%	97,25%
6	68	32	109	50	259	1,29%	98,54%
7	28	13	54	21	116	0,58%	99,12%
8	12	3	18	9	42	0,21%	99,32%
> 8	30	12	87	7	136	0,68%	100,00%
Totais	4314	1781	8687	5350	20132	100,00%	

Fonte: Elaborado pelo autor a partir de leituras registradas pelos sistemas de recomendação.

O gráfico 5 apresenta o histograma da distribuição das leituras da tabela 6 agrupados por nível e permite visualizar as proporções para cada uma das categorias de recomendações e o seu decréscimo conforme aumenta a profundidade da navegação do usuário.

Gráfico 5: Leituras obtidas pelo acompanhamento de recomendações no *site* Enoteca no período de dezembro de 2007 a janeiro de 2008.



Fonte: Elaborado pelo autor.

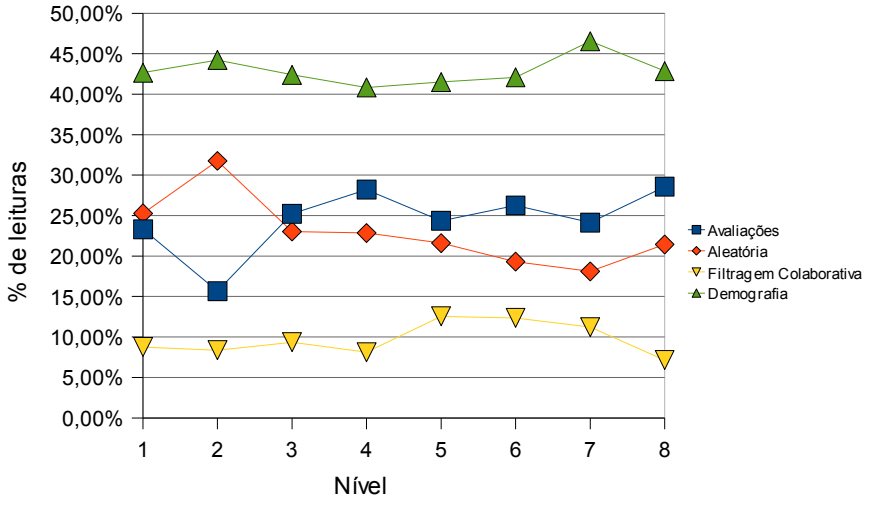
Tabela 7: Distribuição das observações em % registradas por tipo de recomendação e nível resultantes do acompanhamento de recomendações no *site* Enoteca no período de dezembro de 2007 a janeiro de 2008.

Nível	Avaliações	Filtragem Colaborativa	Demografia	Aleatória
1	23,31%	8,76%	42,66%	25,27%
2	15,66%	8,35%	44,23%	31,76%
3	25,23%	9,36%	42,39%	23,03%
4	28,20%	8,12%	40,84%	22,85%
5	24,35%	12,56%	41,50%	21,59%
6	26,25%	12,36%	42,08%	19,31%
7	24,14%	11,21%	46,55%	18,10%
8	28,57%	7,14%	42,86%	21,43%
Mais que 8	22,06%	8,82%	63,97%	5,15%
Média	24,20%	9,63%	45,23%	20,94%
Variância	0,13%	0,03%	0,45%	0,45%
Desvio Padrão	3,63%	1,83%	6,81%	6,71%

Fonte: Elaborado pelo autor a partir de leituras registradas pelos sistemas de recomendação.

Com o objetivo de avaliar se são mantidas as proporções de consultas em cada um dos níveis, foi calculado o percentual que representa cada uma das categorias em cada nível. A tabela 7 contém a distribuição em percentuais para cada nível, apresenta também a variância e o desvio padrão para cada categoria. Esta visão permite observar que os métodos demográficos e aleatórios são os que resultaram em maior variância e desvio padrão.

Gráfico 6: Distribuições das leituras em % conforme o nível de profundidade da visita no site Enoteca no período de dezembro de 2007 a janeiro de 2008.



Fonte: Elaborado pelo autor.

O gráfico 6 ilustra os dados da tabela 7 e mostra que há uma troca de posições no segundo nível entre as exibições aleatórias e as avaliações de usuários. Também é possível ver que, exceto no nível dois os resultados se mantêm em uma faixa relativamente fixa.

4.2.3 O decaimento também segue uma lei de potência?

Como o gráfico 5 apresenta um formato que se assemelha à uma lei de potência, também se analisou se este decaimento das recomendações conforme o nível segue esta mesma lei. Após efetuada a regressão, foi observado que o decaimento do número de leituras por nível não segue esta distribuição, ela apresenta-se na forma exponencial, na qual a relação segue a equação 19.

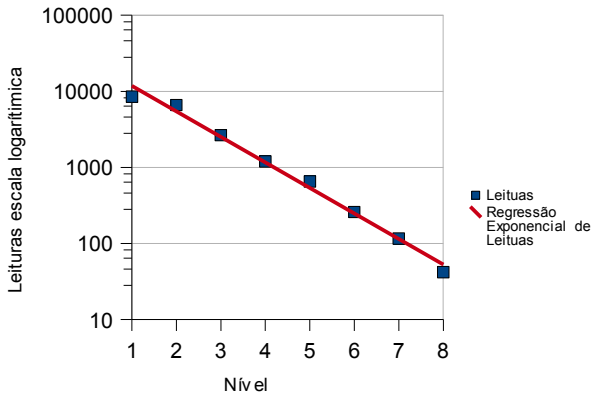
$$\hat{y} = c\,a^{x} \tag{19}$$

O expoente x representa o nível de profundidade de navegação e o parâmetro a é fracionário fazendo com que $\hat{y}$ diminua a medida que x cresce. Pela regressão com o método dos mínimos quadrados, seus parâmetros estimados resultam em (20) com precisão $r^2 = 0{,}9906$.

$$\hat{y} = ٢٥٤٨٠ * ٠.٤٦١٦^{x} \tag{20}$$

O gráfico mostra o diagrama de dispersão e reta resultante do ajuste obtido pela regressão das freqüências das leituras utilizando-se para isso os dados das leituras até o 8 nível.

Gráfico 7: Regressão em escala logarítmica de leituras x nível para Enoteca no período de dezembro de 2007 a janeiro de 2008.



Fonte: Elaborado pelo autor.

Os coeficientes individualizados por método podem ser vistos na tabela 8. O coeficiente *a* pode ser visto como o coeficiente de inclinação da reta de regressão e um índice maior representa um decaimento menor. Para o estudo de caso, as avaliações de usuários apresentaram o maior índice *a*=0,4816.

Tabela 8: Coeficiente *a* e precisão r² por método de recomendação na Enoteca para o período de dezembro de 2007 a janeiro de 2008.

Método	*a*	r²
Avaliações	0,4816	0,9902
Demográfico	0,4631	0,9901
Filtragem Colaborativa	0,4689	0,9680
Aleatório	0,4372	0,9874
Média	0,4627	
Desvio Padrão	0,0162	

Fonte: Elaborado pelo autor a partir de leituras registradas pelos sistemas de recomendação.

4.2.4 Dados ajustados conforme quantidade de exibições

A cada consulta efetuada em algum produto, o sistema oferece até cinco recomendações por método, exceto para a demográfica, na qual o sistema sugere vinhos da mesma vinícola, região e uva e pode apresentar até quinze sugestões. Em casos em que não existam vinhos que satisfaçam as condições do filtro, o número de sugestões é menor. Em casos em que existam mais do que cinco itens, eles são sorteados entre os resultados de forma que, para um mesmo produto possam ser apresentadas recomendações diferentes a cada consulta efetuada.

Os métodos demográficos e aleatórios, apresentam uma quantidade muito maior de recomendações do que o método de recomendações de usuários. Para ponderar a quantidade de recomendações sugeridas com a quantidade de leituras efetuadas, foram registradas as quantidades exibidas para cada método. Assim, o sistema, além de apresentar as recomendações, ele armazena a quantidade de itens exibidos, por exemplo, um vinho pode ter outros 5 da mesma vinícola ou variedade, mas somente outros dois avaliadores. O total de recomendações apresentadas, em percentagem, são mostrados na tabela 9.

Tabela 9: Percentagem de recomendações oferecidas por método na Enoteca no período de dezembro de 2007 a janeiro de 2008.

Tipo	%
Demografia	47,64%
Aleatório	26,18%
Filtragem Colaborativa	21,01%
Avaliações	5,17%
Total	100,00%

Fonte: Elaborado pelo autor a partir de leituras registradas pelos sistemas de recomendação.

A quantidade de recomendações demográficas, por ser a mais genérica e oferecer mais possibilidades, contribuiu com aproximadamente a metade das recomendações. Enquanto as avaliações de outros usuários resultaram em apenas 5% das recomendações. Isto se deve ao fato de que os produtos possuem em média 1,2 avaliações cada e está de acordo com o esperado. Este número representa aproximadamente um quinto da quantidade de recomendações aleatórias que são sempre exibidas.

Tabela 10: Comparativo de observações por categoria de recomendação ponderado por número de exibições na Enoteca no período de dezembro de 2007 a janeiro de 2008.

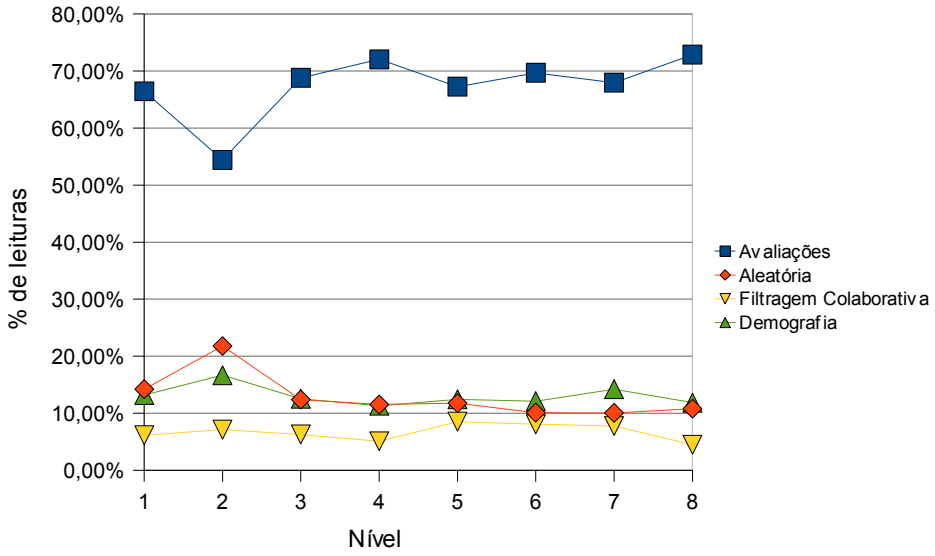
Nível	Avaliações	Filtragem Colaborativa	Demografia	Aleatória
1	66,44%	6,14%	13,19%	14,22%
2	54,40%	7,14%	16,67%	21,79%
3	68,79%	6,28%	12,54%	12,40%
4	72,05%	5,10%	11,32%	11,52%
5	67,26%	8,53%	12,44%	11,78%
6	69,69%	8,07%	12,12%	10,12%
7	67,96%	7,76%	14,22%	10,06%
8	72,87%	4,48%	11,86%	10,79%
Mais que 8	68,54%	6,74%	21,56%	3,16%
Média	67,56%	6,69%	13,99%	11,76%
Variância	0,26%	0,02%	0,09%	0,21%
Desvio Padrão	5,05%	1,27%	3,07%	4,57%

Fonte: Elaborado pelo autor a partir de leituras registradas pelos sistemas de recomendação.

Estas percentagens permitem ponderar as recomendações acessadas pelos usuários de acordo com a quantidade de recomendações apresentadas. Para efetuar o ajuste, foi utilizada uma regra de três simples dividindo-se o percentual total de exibições da categoria pelo número absoluto de observações da mesma categoria. O resultado deste ajuste é apresentado na tabela 10.

Agora as avaliações de usuários se destacam tendo em média 67,56% das leituras de recomendações seguidas pelos visitantes. Após, em um distante segundo lugar vem a demografia com 13,99%, a aleatória em terceiro com 11,76% e por último a filtragem colaborativa com 6,69%. O gráfico 8 ilustra este resultado, mostrando que as recomendações de usuários ficam distantes das demais.

Gráfico 8: Comparativo de leituras por categoria de recomendação ponderado por número de exibições para o *site* Enoteca no período de dezembro de 2007 a janeiro de 2008.

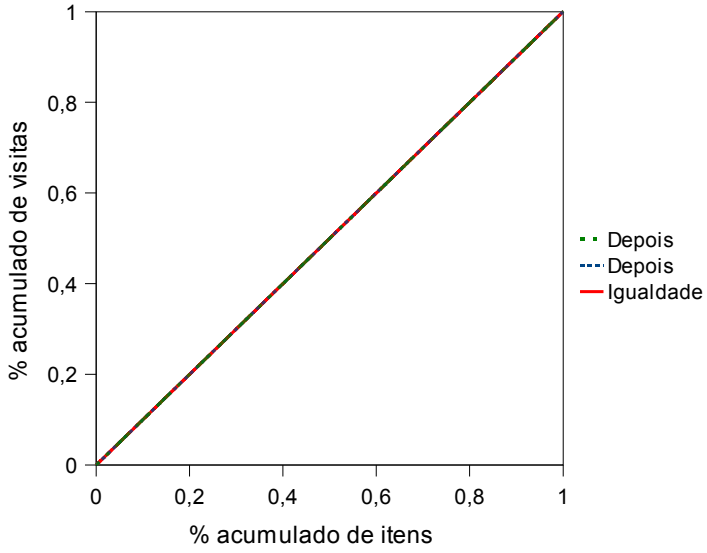


Fonte: Elaborado pelo autor.

4.2.5 Os sistemas de recomendação conduzem ao *long tail*?

Até aqui, foi possível ver como os visitantes do *site* utilizam os sistemas de recomendação, mas ainda é necessário responder a questão da pesquisa que pede se os sistemas de recomendação possibilitam atingir o *long tail*. Para responder a isto, foram analisadas as distribuições das visualizações de páginas de produtos antes e após a implementação dos sistemas de recomendação. Para efetuar esta análise, foram utilizados os dados coletados anteriormente à implementação dos sistemas de recomendação, durante os meses de setembro a novembro de 2007. O resultado é mostrado no gráfico 9.

Gráfico 9: Curva de Lorenz comparativa de distribuição das visualizações de itens antes e depois da implementação dos sistemas de recomendação no *site* Enoteca nos períodos de setembro e outubro de 2007 e dezembro de 2007 e janeiro de 2008.



Fonte: Elaborado pelo autor.

Pela curva de Lorenz (9), é possível visualizar que após a implementação dos sistemas de recomendação a curva apresenta menor desigualdade. Para se obter uma medida desta diferença, foi calculado o índice Gini utilizando a fórmula (14) para a distribuição das visitas antes e depois da implementação resultando respectivamente em 0,608 e 0,470, o que representa uma redução de 22,79%.

4.3 ANÁLISE DOS RESULTADOS

O primeiro estudo de caso mostrou que as vendas de uma empresa de comércio eletrônico e uma outra de comércio tradicional, possuem comportamentos diferentes quando comparados em termos de concentração de vendas utilizando a Curva de Lorenz e o Coeficiente de Gini como parâmetros. Esta diferença estaria de acordo com a teoria *long tail* e

pode indicar que ela seja a sua causa. Mesmo que as duas não tenham grande discrepância nas quantidades de produtos oferecidos como acontece, por exemplo, com livros em que a Amazon apresenta uma média de 57 vezes mais títulos que uma livraria típica (BRYNJOLFSSON, HU, SMITH, 2003). Mas, pelo fato de estarem sendo comparados resultados de duas empresas diferentes e sendo uma delas um supermercado, que não oferece qualquer tipo de sugestão ao cliente além das ofertas e promoções, estas diferenças também devem ser atribuídas a outros fatores, como a quantidade e produtos, seus preços, as negociações com os fornecedores e formas de venda.

Segundo Anderson (2006), o efeito *long tail* seria causado principalmente pelo enorme número de itens oferecidos pelas lojas virtuais que não estariam sujeitas às restrições econômicas da escassez em produtos que possam ser facilmente reproduzidos. A explicação para o fato do vinho, que mesmo com pouca diferença na quantidade de itens oferecidos, consiga apresentar um efeito no estilo *long tail*, é que ele também é causado pela distribuição dos consumidores, que no caso do comércio eletrônico estão representados por uma base mais distribuída, enquanto em um comércio local, tradicionalmente apenas um grupo de consumidores locais, que moram próximos, que compram sempre os mesmos produtos de sua preferência.

Por que no caso do vinho, ocorre esta pequena diferença na quantidade de itens do catálogo entre lojas virtuais e físicas? Este fenômeno pode ser explicado pela logística do produto. Enquanto um livro, música, vídeo pode ser traduzido e reproduzido em qualquer país ou local, esta situação não ocorre com o vinho fino que precisa ser transportado e armazenado. O transporte é delicado e caro pois o produto é frágil e além de possuir prazo de validade, como a maioria dos itens de alimentação, deve ser armazenado em locais climatizados. As empresas produtoras encontram-se distribuídas pelo mundo, algumas em locais de difícil acesso e cada uma produz pequeno número de produtos, uma média de 5 a 10 no Brasil. Estas questões, tornam difícil mesmo para lojas virtuais que não possuem estoque, manterem um catálogo amplo mesmo que estimativas de *experts* apontem para a existência de mais de 100 mil rótulos de vinhos a venda no mundo todo (LOGALDI, 2008).

A matemática envolvida para o ajuste das curvas de potência se revelou relativamente complexa, trabalhos recentes como o de Clauset, Shalizi e Newman (2007) propõe várias metodologias que podem ser exploradas em um trabalho mais aprofundado. Neste trabalho optou-se por utilizar a metodologia proposta por Newman (2006) e também a regressão linear simples via extração dos logaritmos (DOWNING e CLARK, 1998) e ambas apresentaram um

ajuste r^2 maior que 0.9 considerado alto suficiente para demonstrar que trata-se de uma lei de potência.

Por este motivo, o uso do Coeficiente de Gini como forma de comparar as curvas de vendas é muito mais simples de implementar e vários trabalhos voltados ao estudo do *long tail,* têm utilizado este indicador para sua medição, por exemplo em Brynjolfsson, Yu e Simester (2007), Elberse (2007) e Fleder e Hosanagar (2007). Assim como nas análises dos percentuais de vendas, o Gini também apresentou resultado de acordo com a teoria *long tail* demonstrando uma menor concentração das vendas na loja virtual. Mesmo que não seja possível afirmar que exista *long tail* para o vinho, os resultados apontam para indícios de que este fenômeno esteja ocorrendo neste estudo de caso.

O segundo estudo de caso, que permitiu explorar o uso de sistemas de recomendação em um *site* voltado para a recomendação de vinhos efetuando comparativos, trouxe alguns *insights* interessantes. O resultado mais importante é que as recomendações de outros usuários se destacaram das demais. O que está de acordo com a intuição pois é mais provável que as pessoas aceitem recomendações de outras pessoas que tenham gostos semelhantes ou de *experts* do que arriscar uma recomendação gerada por um computador sem uma explicação aparente. Herlocker, Konstan e Riedl (2000) encontraram evidências de que as recomendações feitas por filtragem colaborativa obtém melhor aceitação quando são explicadas.

O mesmo estudo também pode auxiliar a entender o outro resultado que chama a atenção. Se trata do último lugar em desempenho obtido pela filtragem colaborativa. Mesmo que autores afirmem que é a tecnologia de recomendações de maior sucesso (SARWAR et al., 2001). Seu fraco desempenho pode ser explicado pelo pequeno número de recomendações por produto e usuário apesar do *Slope One* afirmar ser adequado para uma situação de dados espalhados (LEMIRE e MACLACHLAN, 2005). Mesmo em *sites* que utilizam filtragem colaborativa como no caso da Amazon, estas recomendações são mostradas na forma de outros usuários na forma de clientes que compraram estes itens, também compraram (LINDEN, SMITH e YORK, 2003) o que a torna auto explicativa.

A curva de decaimento por nível em forma exponencial, também com um ajuste da curva que resultadou em r^2 alto, próximo de 0,99 pode ser usado como uma medida da eficácia das recomendações em sites de comércio eletrônico. As medidas comumente utilizadas são a quantidade média de páginas visitadas e o tempo médio no site e para a obtenção da profundidade da navegação exige uma implementação mais complexa. O

coeficiente de inclinação da curva, indica que as avaliações de usuários possuem um decaimento menor, também ficando de acordo com o resultado em que as pessoas preferem as avaliações de outras pessoas.

Finalmente, a comparação entre os coeficientes Gini, calculados para as distribuições de antes e depois da implementação dos sistemas de recomendação, mostra que houve uma redistribuição da concentração de visitas entre os itens. Esta redistribuição ocorreu sem que houvesse um aumento na quantidade de itens em ordens de magnitudes no período. Este resultado, que está de acordo com a proposição do *long tail*, pode ser um indicador de que esta diferença nos coeficientes tenha sido causada pelos sistemas de recomendação implementados.

CONCLUSÃO

Produtos e empreendimentos que seriam inviáveis em um mercado puramente local, tornam-se viáveis em âmbito global. O *long tail* prevê um futuro onde se venderá menos de mais. Ele mostrou que há desejo do consumidor de experimentar o diferente e que a internet e o comércio eletrônico podem auxiliar a tornar isto possível, mas não é tão simples. As regras não mudaram completamente. Não é suficiente criar um produto esquisito e esperar que o consumidor *long tail* o encontre. O consumidor é levado até eles seguindo um caminho seguro e se arriscando aos poucos obtendo conselhos e recomendações. Por isso os mais vendidos não vão desaparecer e nem deixarão de ser importantes mesmo em lojas virtuais.

O primeiro estudo de caso apresentou uma diferença nas distribuições de vendas entre uma loja virtual e outra de tijolo e cimento. Enquanto a primeira obteve 67% do seu faturamento pelos 20% mais vendidos, a segunda, obteve 79% do seu faturamento para a mesma proporção. Chegando muito próximo ao 80/20 de Pareto. Esta diferença está de acordo com a teoria *long tail* e pode ser explicada pelo consumidor em busca de produtos diferentes levado por *sites* de recomendação, comunidades virtuais e o tradicional boca-a-boca. Enquanto na loja física, os consumidores locais variam menos as suas escolhas.

É importante perceber que uma loja virtual, pode ter um catálogo maior que uma empresa física mesmo que não venda todos os itens, o custo de manter este catálogo é muito menor do que o de manter os produtos. Este catálogo deve ser usado como um recurso de vendas pois é através dele que o visitante vai encontrar a empresa e a partir deste momento, pode receber outras recomendações que podem ser aceitas.

O segundo estudo de caso mostrou que houve uma redistribuição das freqüências dos acessos aos itens no período após a implementação dos sistemas de recomendação. Como este fato está de acordo com a previsão do *long tail*, isso pode ser um indicador de que eles tenham sido os causadores desta mudança.

Os resultados também mostram que os visitantes do *site* de recomendações de vinho Enoteca, consultam em média duas recomendações e preferem recomendações baseadas no

gosto de outras pessoas do que recomendações simplesmente baseadas em semelhanças de produtos. Não que estas recomendações genéricas deixem de ser importantes, pois em muitas ocasiões a pessoa está simplesmente em busca de algo ligeiramente diferente do que já conhece, como um outro produto do mesmo fornecedor no qual confia. As escolhas baseadas em opiniões de outros fornecem a sensação de segurança, pois diminuem o estresse de efetuar uma escolha errada. Quem escolhe baseando-se em opinião de outro, de certa forma está transferindo uma parte da responsabilidade de acertar ou da tirania da escolha para uma outra pessoa.

Outro resultado é o decaimento nível a nível que segue uma lei exponencial, a sua inclinação pode ser medida e utilizada como um indicador da efetividade das recomendações. Ele pode complementar os sistemas de análise de visitantes, chamados de *web analytics* que registram o número médio de visualizações de páginas e o tempo médio que o visitante permaneceu no seu *site*.

Os resultados deste trabalho permitem entender melhor a forma que os visitantes do *site* seguem as recomendações e podem servir para as empresas que pretendam recomendar produtos de forma automatizada. Também mostra que é importante deixar claro o motivo das recomendações sugeridas.

Consultar opiniões de outras pessoas para tomar uma decisão é um evento do cotidiano das pessoas. A internet trouxe enorme facilidade quando esta decisão envolve a aquisição de um produto, pois possibilita a busca das opiniões positivas e negativas de forma automática. O desafio das empresas agora é utilizar este recurso da internet ao seu favor.

Idéias para Trabalhos Futuros

Outros estudos sobre a existência do *long tail* para o vinho na internet podem ser temas de trabalhos futuros, um caminho para isso seria através da comparação de dados de empresas de vinho que possuam tanto loja virtual quanto física, tomando as vendas *online* e *offline.* Neste caso, a diferença entre os coeficientes seria um indicador mais consistente. O coeficiente Gini também pode ser utilizado em outras comparações para lojas somente virtuais e somente físicas.

Para a obtenção de mais resultados que fortaleçam e expliquem as variações de

concentração no coeficiente Gini, há a possibilidade de remoção das recomendações de forma total ou parcial, deixando apenas um tipo de recomendação.

O primeiro estudo foi efetuado utilizando-se dados de um supermercado, geralmente este tipo de comércio, não possui alguém que aconselhe vinhos aos clientes, como ocorre nas lojas especializadas. No entanto, existiria a possibilidade de instalação de um quiosque de recomendações onde o consumidor pudesse procurar vinhos de acordo com seus gostos e consultar recomendações de outros consumidores. Esta experiência exigiria um certo investimento em equipamentos mas poderia testar a redistribuição das vendas causadas por sistemas de recomendação em ambientes *offline*.

Fleder e Hosanagar (2007) tratam a questão de como os sistemas de recomendação influenciam a concentração de vendas e seu estudo que eles podem regular a forma da curva de vendas, causando tanto uma concentração maior, para o caso dos rankings dos mais vendidos quanto uma diluição nos nichos, como no caso da Amazon. Um trabalho futuro poderia testar como os sistemas de recomendação moldam a curva por oferecer somente um tipo de recomendações por vez.

Seguindo pela idéia de que as pessoas preferem recomendações de outras pessoas, é possível dar um peso para o usuário que recomenda, como expert, médio ou iniciante e compará-las de forma a entender como estas devam ser balanceadas nos sites de comércio eletrônico.

Neste estudo foi utilizado um algoritmo baseado em filtragem colaborativa, um trabalho futuro poderia também ser a criação de algoritmos especializados em vinho baseados em modelos que utilizem técnicas como as redes Bayesianas ou redes neurais artificiais.

BIBLIOGRAFIA

ADAMIC, L. **Zipf, Power-laws, and Pareto - A Ranking Tutorial**. Disponível em : <http:// www.hpl.hp.com/research/idl/papers/ranking/ranking.html>. Acesso em : 30 jan 2008

AGGARWAL, C.; WOLF, J.; WU, K.; YU, P. Horting Hatches an Egg: A New Graph-Theoretic Approach to Collaborative Filtering. **Proceedings of ACM SIGKDD International Conference on Knowledge Discovery & Data Mining**. San Diego, CA: ACM, 1999.

ANDAM, R. **E-Commerce and E-Business**. 1. ed. Kuala Lumpur, Malaysia: UNDP-APDIP, 2003. 47 p.

ANDERSON, C. **The Long Tail**. Disponível em : <http://www.wired.com/wired/archive/12.10/tail.html>. Acesso em : 20 fev 2008.

ANDERSON, C. **An Apparent Long Tail Paradox**. Disponível em : <http://www.thelongtail.com/the_long_tail/2006/04/an_apparent_lon.html>. Acesso em : 10 fev 2008.

ANDERSON, C. **A Cauda Longa**: do Mercado de Massa para o Mercado de Nicho. Tradução Afonso Celso da Cunha Serra. 2. ed. Rio de Janeiro: Elsevier, 2006. 240 p.

ATKINSON, R.; MCKAY, A. **Digital Prosperity:** Understanding the Economic Benefits of the Information Technology Revolution. 1a. ed. Washington, DC: ITIF, 2007. 78 p.

BALABANOVIC, M.; SHOHAM, Y. Fab: Content-based, Collaborative Recommendation. **Communications of the ACM**, New York, NY, USA, v40: 66-72. 1997

BERRY, M.; LINOFF, G.; **Data Mining Techniques:** For Marketing Sales, and Customer Relationship Management. 2a. ed. Indianapolis, Indiana: Wiley Computer Publishing, 2004. 672p.

BERSON, A.; SMITH. S.; THEARLING, K.; **Building Data Mining Applications for CRM**. . : McGraw-Hill Companies, 1999. 488p.

BILLSUS, D.; PAZZANI, M. Learning Collaborative Information Filters. **Proceedings of the International Conference on Machine Learning**. Madison, WI: Morgan Kaufmann, 1998.

BLOCH, M.; PIGNEUR, Y,; SEGEV, A. **On the Road of Electronic Commerce -- a Business Value Framework, Gaining Competitive Advantage and Some Research Issues**. Disponível em : <http://inforge.unil.ch/yp/Pub/ROAD_EC/EC.HTM>. Acesso em : 20 nov. 2007

BRAGA, P. **Indrodução à Mineração de Dados**. 2a. ed. Rio de Janeiro: E-Papers, 2005. 212p.

BREESE, J.; HECKERMAN, D.; KADIE, C.; Empirical Analysis of Predictive Algorithms for Collaborative Filtering. **Uncertainty in Artificial Intelligence. Proceedings of the Fourteenth Conference**. University of Wisconsin Business School, Madison, Wisconsin, USA: Morgan Kaufman, 1998. pp. 43-52

BRESNAHAN, T.; TRAJTENBERG, M. General Purpose Technologies: 'Engines of Growth'?. **Journal of Econometrics**, Cambridge, MA, 65: p. 88-108. 1995

BRYNJOLFSSON, E.; HU, Y.; SIMESTER, D. **Goodbye Pareto Principle, Hello Long Tail:** the effect of search costs on the concentration of product sales (working paper). Disponível em: <http://ssrn.com/abstract=953587>. Acesso em : 08 mar 2007.

BRYNJOLFSSON, E.; HU, Y.; SMITH, M. D. Consumer Surplus in the Digital Economy: Estimating the Value of Increased Product Variety at Online Booksellers. **Management Science**, Hannover, v. 49: pp. 1580–1596. nov. 2003

CALDERÓN-BENAVIDES, L.; GONZÁLEZ-CARO, C. **Recommender Systems Through Collaborative Filtering**. Disponível em : <jones.tc/ta-wiki-2005/images/c/cc/Recommender_Systems_through_Collaborative_Filtering.pdf>. Acesso em : 15 jan 2008.

CARVALHO, A. **Datamining - A Mineração de Dados no Marketing, Medicina, Economia, Engenharia, e Administração**. 1a ed. São Paulo, SP: , 2001. 236 p.

CETIC.BR. **Pesquisa Sobre o Uso das Tecnologias da Informação e da Comunicação no Brasil 2006**. São Paulo: Comitê Gestor da Internet no Brasil, 2007, 322 p. Disponível em <http://cetic.br>. Acesso em: 15 mai. 2007.

CLAUSET, A.; SHALIZI, C.; NEWMAN, M. **Power-law Distributions in Empirical Data.** Disponível em <http://arxiv.org/abs/0706.1062v1>. Acesso em 10 jan 2008. .

COOPER, D.; SCHINDLER, P. **Métodos de Pesquisa em Administração**. 7a ed. Porto Alegre, RS: Bookman, 2003. 640 p.

DOWNING, D.; E CLARK, 1998 **Estatística Aplicada**. 1a ed. São Paulo: Saraiva, 1998. 455 p.

DROUX, S. **Vogoo - Recommendation Engine and Collaborative Filtering**. Disponível em : <http://www.vogoo-api.com/>. Acesso em : 03 nov. 2007

E-BIT. **Web Shoppers**. 2006. 30 p. Disponível em : <http://www.webshoppers.com.br>. Acesso em: 28 mar 2007.

ELBERSE, A. **A Taste for Obscurity? An Individual-Level Examination of "Long Tail" Consumption.** Harvard Business School Working Paper, No. 08-008, August 2007.

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

No. 07-015. Disponível em: <www.people.hbs.edu/aelberse/papers/hbs_07-015.pdf>. Acessado em: 17 dez. 2007.

ELLENBERG, J. **This Psychologist Might Outsmart the Math Brains Competing for the Netflix Prize**. Disponível em : <http://www.wired.com/techbiz/media/magazine/16-03/mf_netflix/?currentPage=all#>. Acesso em : 28 fev. 2008

FGV-EAESP. **Pesquisa FGV-EAESP de Comércio Eletrônico no Mercado Brasileiro**. 2007. 18 p. 8. ed.

FLEDER, D.; HOSANAGAR, K. Recommender Systems and their Impact on Sales Diversity. **EC '07: Proceedings of the 8th ACM conference on Electronic commerce**. New York, NY, USA: ACM, 2007. p. 192-199.

GARTNER GROUP. **Understanting Hype Cycles**. Disponível em : <http://www.gartner.com/pages/story.php.id.8795.s.8.jsp>. Acesso em : 14 jun 2007

GHOSE, A.; GU, B. **Search Costs, Demand Structure and Long Tail in Electronic Markets: Theory and Evidence.** NET Institute Working Paper No. 06-19. Disponível em: <http://ssrn.com/abstract=941200>. Acesso em: 20 dez. 2007.

HERLOCKER ET AL. Evaluating Collaborative Filtering Recommender Systems. **ACM Transactions on Information Systems**, New York, NY, USA, vol. 22: pg. 5-53. 2004

HERLOCKER, J. **Understanding and Improving Automated Collaborative Filtering Systems**. 2000. 148 f. Ph.D. Thesis - Faculty of The Graduate School, University of Minnesota, 2000.

HERLOCKER, J.; KONSTAN, J.; RIEDL, J. Explaining Collaborative Filtering Recommendations. **Proceedings of the ACM 2000 Conference on Computer Supported Cooperative Work**. Philadelphia, PA: ACM, 2000. p. 241-250

HERVAS-DRANE, A. **Word of Mouth and Recommender Systems: A Theory of the Long Tail**. NET Institute Working Paper No. 07-41 Disponível em: <http://ssrn.com/abstract=1025123>. Acesso em: 15 jan. 2008.

HILL, W.; STEAD, L.; ROSENSTEIN, M. e FURNAS, G. Recommending and Evaluating Choices in a Virtual Community of Use. **Proceedings of CHI'95**. Denver CO: ACM Press, 1995.

JANSEN, W.; STEENBAKKERS, G.C.A.; JÄGERS, H.P.M. **Electronic Commerce and Virtual Organizations**. Universiteit van Amsterdam. Department of Accountancy & Information Management. Amsterdam, 1999. Disponível em : <http://primavera.fee.uva.nl/PDFdocs/99-17.pdf>. Acessado em : 05 jun 2007

KALAKOTA, R.; ROBINSON, M. **E-Business**: Estratégias para Alcançar o Sucesso no Mundo Digital. 2.ed. Porto Alegre: Bookman, 2002. 470 p.

LEMIRE, D.; MACLACHLAN, A.; Slope One Predictors for Online Rating-Based Collaborative Filtering. **Proceedings of SIAM Data Mining (SDM'05)**. Newport Beach,

California: SIAM, 2005. 8p.

LEMIRE, D.; MCGRATH, S. **Implementing a Rating-Based Item-to-Item Recommender System in PHP/SQL**. Disponível em : <http://www.daniel-lemire.com/fr/documents/publications/webpaper.pdf>. Acesso em : 03 nov. 2007

LINDEN, G.; SMITH, B.; YORK, J. Amazon.com Recommendations - Item-to-item Collaborative Filtering. **IEEE Internet Computing**, Piscataway, NJ, USA, v.7: p. 76-80. 2003

LOGALDI, A. **Quantos vinhos existem?**. Disponível em : <http://www.enoteca.com.br/index.php?id=11005>. Acesso em : 05 fev 2008

LOHSE G., SPILLER P. Electronic Shopping - Designing online stores with effective customer interfaces has a critical influence on traffic and sales. **Communications of ACM**, New York, v. 7: p. 81-88. jul. 1998

MATOS, C. **Econometria Básica**. 3. ed. São Paulo: Atlas, 2000. 330 p.

NASDAQ. **Netflix Inc. Company Financials**. Disponível em : <http://www.nasdaq.com/asp/ExtendFund.asp?selected=NFLX&symbol=NFLX>. Acesso em : 20 jan. 2008

NETFLIX. **Netflix Prize**. Disponível em : <http://www.netflixprize.com/>. Acesso em : 20 jan. 2008.

NEVES, JOSÉ LUIS Pesquisa Qualitativa : Características, Usos e Possibilidades. **Caderno de Pesquisas em Administração**, São Paulo, v.1: p. 1-5. 2. sem. 1996

NEWMAN, M. Power Laws, Pareto Distributions and Zipf's Law. **Contemporary Physics**, Ann Arbor, MI, 46: p. 323-351. 2006

O'REILLY, T. **What Is Web 2.0 - Design Patterns and Business Models for the Next Generation of Software**. Disponível em : <http://www.oreilly.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html>. Acesso em : 20/02/2008

O'SULLIVAN, B. **Collaborative Filtering Made Easy**. Disponível em : <http://www.serpentine.com/blog/2006/12/12/collaborative-filtering-made-easy/>. Acesso em : 1 nov. 2007

OWEN, S. **Taste: Collaborative Filtering for Java**. Disponível em : <http://taste.sourceforge.net/>. Acesso em : 03 nov. 2007

PARKER, R. **The Wine Advocate Rating System**. Disponível em : <http://www.erobertparker.com/info/legend.asp>. Acesso em : 15 jan. 2008

RAPIDMINER. **Rapidminer 4.0:** User Guide, Operator Reference, Devoloper Tutorial. 1a. ed. Germany: Rapid-1, 2007?. 545p. Disponível em <http://sourceforge.net/project/showfiles.php?group_id=114160>. Acesso em 20/12/2007

RESNICK, P.; IACOVOU, M.; SUCHAK, P.; BERGSTORM, P.; RIEDL, J. GroupLens: An Open Architecture for Collaborative Filtering of Netnews. **ACM 1994 Conference on Computer Supported Cooperative Work**. Chapel Hill, North Carolina: ACM, 1994. 175-186

SARWAR, B. et al. Analysis of Recommendation Algorithms for E-Commerce. **Proceedings of the 2nd ACM conference on Electronic commerce**. New York, NY, USA: ACM, 2000. páginas 158 - 167

SARWAR, B.; et al. Item-Based Collaborative Filtering Recommendation Algorithms. **Proceedings of the 10th international conference on World Wide Web**. Hong Kong, Hong Kong: ACM, 2001.

SCHAFER, J.; KONSTAN, J.; RIEDL, J. Recommender Systems in E-Commerce. **EC '99: Proceedings of the 1st ACM conference on Electronic commerce**. New York: ACM Press, 1999. p. 158-166

SCHWARTZ, B. **The Tyranny of Choice**. Disponível em : <http://www.sciammind.com/article.cfm?articleID=00056941-1933-1196-906983414B7F0000&pageNumber=1>. Acesso em : 10 jun. 2007.

SHARDANAND, U.; MAES, P. Social Information Filtering: Algorithms for Automating "Word of Mouth". **Proceedings of CHI'95**. Denver CO: ACM Press, 1995.

SNOOTH. **Snooth Launches Public Beta Version of its Website**. Disponível em : <http://blog.snooth.com/2007/06/06/snooth-launches-public-beta-version-of-its-website/>. Acesso em : 20 jan. 2008.

TERVEEN, L.; HILL, W.; AMENTO, B.; MCDONALD, D.; CRETER, J. PHOAKS: A System for Sharing Recommendations. **Communications of the ACM**, New York, NY, USA, v40: p59-63. 1997

TERVEEN. L.; HILL. W. Beyond Recommender System: Helping People Help Each Other. In JACK CARROLL. **HCI In The New Millennium**. 1 ed. New YorkAddison-Wesley. 2001. 22, p.487.

TUCKER, C.; ZHANG, J. **Long Tail or Steep Tail? A Field Investigation into How Online Popularity Information Affects the Distribution of Customer Choices**. MIT Sloan Research Paper No. 4655-07 Disponível em: <http://ssrn.com/abstract=1002763>. Acesso em: 20 jan. 2008.

US CENSUS BUREAU. **E-Stats**. 2007. 26 p. U.S. Departament of Commerce Economics and Statistics Administration. Disponível em <http://www.census.gov/estats>. Acesso em 14 jun. 2007.

VIEGAS, F.; DONATH, J. Chat Circles. **Proceedings of CHI'99**. Pittsburgh, PA: ACM Press, 1999.

WYCKOFF, A.; COLECCHIA, A.; **The Economic and Social Impact of Electronic Commerce:** Preliminary Findings and Research Agenda. 1. ed. Paris, France: OECD