JOÃO LOURENÇO VIVAN BERNARTT

UM SISTEMA DE RECOMENDAÇÃO BASEADO EM FILTRAGEM COLABORATIVA

FLORIANÓPOLIS
2008

UNIVERSIDADE FEDERAL DE SANTA CATARINA

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

UM SISTEMA DE RECOMENDAÇÃO BASEADO EM FILTRAGEM COLABORATIVA

Dissertação submetida à
Universidade Federal de Santa Catarina
como parte dos requisitos para a
obtenção do grau de Mestre em Engenharia Elétrica.

JOÃO LOURENÇO VIVAN BERNARTT

Florianópolis, Setembro de 2008.

UM SISTEMA DE RECOMENDAÇÃO BASEADO EM FILTRAGEM COLABORATIVA

JOÃO LOURENÇO VIVAN BERNARTT

'Esta Dissertação foi julgada adequada para a obtenção do título de Mestre em Engenharia Elétrica, Área de Concentração em *Controle, Automação e Informática Industrial*, e aprovada em sua forma final pelo Programa de Pós-Graduação em Engenharia Elétrica da Universidade Federal de Santa Catarina.'

Guilherme Bittencourt, Doutor.
Orientador

Kátia Campos de Almeida, Doutora.
Coordenadora do Programa de Pós-Graduação em Engenharia Elétrica

Banca Examinadora:

Guilherme Bittencourt, Doutor
Presidente

Jean-Marie Farines, Doutor

Eduardo Camponogara, Doutor

Luis Otavio Campos Alvares, Doutor

A Moacir, Vilma, Karen, Bruno e Rafaela com amor e carinho...

AGRADECIMENTOS

Algumas pessoas contribuíram, direta e indiretamente, de forma essencial para a realização deste trabalho. À estas pessoas eu reservo este momento para um sincero agradecimento. A seguir, é apresentado em ordem cronológica das contribuições, cada um destes agradecimentos. Primeiramente sou grato ao amigo, colega, mentor e professor Prof. Jean-Marie Farines por ter me convencido a realizar o mestrado, a me orientar nas primeiras tentativas de realização do trabalho, ao auxílio da entrada ao término na PGEEL e pela consideração que sempre me prestigiou. Sem o seu incentivo, certamente este trabalho não existiria. Ao amigo Leopoldo Silva Xavier "Bolinho" agradeço por ter encontrado o caminho científico para o problema que lhe apresentei, fornecendo os subsídios iniciais para esta pesquisa. Agradeço o apoio e o incentivo da Fundação CERTI na figura de três pessoas: 1) Marcelo Ferreira Guimarães pela sua visão ao criar as regras que permitiram a liberação do tempo de trabalho para o mestrado; 2) Ricardo Henrique Teixeira, que no seu incondicional apoio, forneceu todas as condições para a conclusão da dissertação e 3) o Prof. Carlos Alberto Schneider, por ter criado uma instituição cujas crenças e valores incentivam seus colaboradores à pesquisa e à inovação. Ao meu orientador, Prof. Guilherme Bittencourt, sou grato por ser uma pessoa que dentro do seu *incrível* conhecimento (e permita-me aqui usar o "incrível"), conseguiu de forma simples colocar sua sabedoria para me direcionar e aconselhar. Obrigado por ter simplificado tudo o que compliquei, pelas incontáveis correções do texto, por ter me ensinado a usar o "aonde" e enfim, ter prezado sempre pela qualidade do meu trabalho. Aos amigos e colegas do time Maracatu Alceu de Medeiros e João Bosco Pereira Filho agradeço por terem sido os companheiros incansáveis neste trabalho, complementando todo o saber necessário para que estes resultados pudessem ser alcançados. Mais do que a ajuda nos resultados alcançados, a demonstração de amizade de vocês dois, talvez seja a coisa mais importante que levo deste trabalho. Outro grande amigo, Cristiano Kurt Ritzke, agradeço pela ajuda na elaboração do slide-mestre da apresentação da defesa, colocando o seu talento mais uma vez a disposição para uma melhor e mais bonita comunicação. Aos Francisco de Assis Neto, Adriano Naspolini e Bernardo de Castro agradeço por nunca terem me contado "como era" ser mestre, fazendo a curiosidade se transformar em motivação. Ao Ricardo Pralon, agradeço pela várias "consultorias" de C. Aos competidores do Netflix Prize agradeço pela colaboração, pela troca de conhecimentos que nortearam o desenvolvimento deste trabalho. Um agradecimento especial dedico a minha amada Rafaela, que sem dúvida foi a pessoa que mais teve influência para a realização deste trabalho. Agradeço pelo carinho, pelo incentivo, pela compreensão na ausência nos finais de semana de trabalho, pelo amor e por ser esta pessoa maravilhosa que me dá alegria e vontade de realizar. Quero agradecer a minha familia, meu pai, minha mãe, minha irmã e meu irmão não só pelo incentivo, pela prestatividade, pelo carinho e pelo amor mas também por tudo aquilo que já fizeram permitindo-me chegar até este momento. Agradeço por serem uma família maravilhosa modelo e direção para tudo o que faço e exemplo para tudo que pretendo ainda atingir. Ao Bruno agradeço ainda pelo auxílio com o resumo e o abstract. Finalmente agradeço a Deus por ter colocado todas estas pessoas no meu caminho.

Resumo da Dissertação apresentada à UFSC como parte dos requisitos necessários para obtenção do grau de Mestre em Engenharia Elétrica.

UM SISTEMA DE RECOMENDAÇÃO BASEADO EM FILTRAGEM COLABORATIVA

JOÃO LOURENÇO VIVAN BERNARTT

Setembro/2008

Orientador: Guilherme Bittencourt, Doutor
Área de Concentração: Controle, Automação e Informática Industrial
Palavras-chave: Filtragem colaborativa, data mining, sistemas de recomendação, aprendizado de máquinas
Número de Páginas: 87

Este trabalho tem como objetivo contribuir com a pesquisa na área de sistemas de recomendação, particularmente sistemas baseados em Filtragem Colaborativa, buscando promover o desenvolvimento de tecnologias informacionais no Brasil. Para tal, propõe-se desenvolver um sistema de recomendação completo para a competição promovida pela empresa Netflix, procurando obter uma precisão melhor que a do sistema em uso pela empresa – o Cinematch ®. Como resultado, primeiramente apresenta-se o estado da arte da pesquisa em sistemas de recomendação e a sistemática da competição. Na seqüência, a contribuição do autor é exposta através da descrição do algoritmo desenvolvido e dos resultados alcançados. Dentre estes, está a qualificação dentro da competição, o bom tempo computacional do algoritmo e a sua precisão que superou a do sistema Cinematch ®. Ao final, as conclusões acerca dos resultados alcançados são descritas e, estabelecem-se perspectivas para a continuidade do trabalho.

Abstract of Dissertation presented to UFSC as a partial fulfillment of the requirements for the degree of Master in Electrical Engineering.

A COLLABORATIVE FILTERING-BASED RECOMMENDER SYSTEM

JOÃO LOURENÇO VIVAN BERNARTT

September/2008

Advisor: Guilherme Bittencourt, Ph.D.
Area of Concentration: Control, Automation and Industrial Informatics
Key words: Collaborative filtering, data mining, recommender systems, machine learning
Number of Pages: 87

This work aims to contribute with the research in the recommender systems area, mainly, collaborative filtering-based systems, and intends to promote the development of informational technologies in Brazil. To accomplish this, the development of a complete system to the Netflix competition is proposed, looking for a better precision then the company´s own system – the Cinematch ®. As a result of this, firstly, the state-of-art of the recommender systems and the competition´s systematic are exposed. Further, the algorithm developed is described and its results presented and analyzed in details. Some of these results reached are, the qualifying rates on the competition, good running time for the computation of the algorithm and its precision, overcoming Cinematch's. At last, but not at least, one concludes on the proposed methodologies and states perspectives for further work.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

Acima de tudo, precisamos preservar a absoluta imprevisibilidade e a total improbabilidade de nossas mentes conectadas. Desta forma podemos manter abertas todas as opções, como no passado.

— Lewis Thomas

A presente dissertação descreve o trabalho desenvolvido na pós-graduação em Engenharia Elétrica especificamente na área de Inteligência Artificial – Aprendizado de Máquinas – entre os anos de 2006 a 2008, servindo como base para obtenção do grau de mestre na Universidade Federal de Santa Catarina. O trabalho esteve inserido dentro do contexto da competição promovida pela empresa Netflix, maior locadora de filmes pela internet do mundo, tendo como objetivo o desenvolvimento de um sistema de recomendação de filmes baseado em Filtragem Colaborativa.

Sistemas de Recomendação formam uma área de estudo recente, que possui suas raízes em pesquisas das ciências cognitivas, recuperação da informação e teoria da aproximação. Sistemas de recomendação procuram apresentar qual item (e.g. livros, filmes, imagens, notícias etc.) mais se adequaria para um determinado usuário [62]. Algoritmos de recomendação são conhecidos por serem utilizados em sites de *e-commerce*, dos quais são usadas informações como itens visualizados, dados demográficos, assuntos de interesse e avaliações para fazer uma lista personalizada de recomendações. Hoje, estes algoritmos se tornaram um diferencial competitivo entre as empresas de comércio eletrônico, sendo que algumas, como a Amazon, já obtêm mais de 30% de suas vendas oriundas de recomendações automáticas feitas pelo site [32].

Neste capítulo introdutório é apresentada a motivação para o tema, os objetivos traçados para o trabalho e uma breve contextualização sócio-econômica e tecnológica do mundo digital, procurando caracterizar a pertinência da pesquisa na área. Por fim é apresentada a metodologia empregada e a divisão de capítulos da presente dissertação.

1.1 Motivação

Vivemos em uma sociedade mundial interconectada de forma global e em tempo real, na qual um número grande de informações pode, em potencial, estar acessível para muitas pessoas simultaneamente. Nesta tempestade de informações, a cada ano cerca de 1,5 bilhão de Gigabytes de informações são produzidos e disponibilizados em aproximadamente 2 bilhões de sites na Internet. É a chamada Sociedade da Informação, que surgiu com o advento da rede mundial de computadores e está migrando para uma nova sociedade a qual alguns pensadores denominam de Sociedade do Conhecimento [13]. Esta nova era pós-industrial é caracterizada pela abundância da informação e pelo fato do diferencial competitivo ser definido pelo conhecimento, que em uma definição simplista seria a aplicabilidade destas informações. Dentro desta cultura cada vez mais cibernética, [1] o desafio maior é conseguir automatizar a geração (ou identificação) do conhecimento coletivo, dando significado aos milhões de dados e informações cujo volume é duplicado a cada seis meses na grande rede mundial – a Internet.

Esta automação é necessária porque o cérebro humano não é adaptado para gerenciar tamanha quantidade de informações, precisando de ferramentas que as compilem e filtrem. A mente é mal equipada para tratar de problemas que devem ser resolvidos de modo serial – um cálculo após o outro – uma vez que os neurônios precisam de um "tempo de recuperação" de cerca de cinco milésimos de segundo, o que resulta em uma capacidade de realizar apenas 200 cálculos por segundo, enquanto que os PCs atuais podem fazer milhões de cálculos por segundo [28]. Contudo, diferentemente da maioria dos computadores, o cérebro é um sistema paralelo de grande porte, com 100 bilhões de neurônios trabalhando todos ao mesmo tempo. Esse paralelismo permite que o cérebro protagonize admiráveis proezas de reconhecimento de padrões, proezas estas ainda a serem alcançadas pelos computadores digitais.

A deficiência humana de lidar com grandes quantidades de informações e a conseqüente necessidade de filtragem da informação e de automação do conhecimento se colocam como alguns dos principais problemas da era do Conhecimento. Para melhor compreender este desafio faz-se necessário estabelecer uma analogia com a ciência que até então buscou desenvolver os processos automatizados: a área de Controle e Automação. Esta área do conhecimento procura criar sistemas que transformem práticas manuais em sistemas automatizados, fazendo

[1] Cibernética aqui utilizada como adjetivo para caracterizar uma sociedade que interioriza sua comunicação e outras formas de transferência de informação através de processos como codificação, feedback e aprendizagem automática e digital.

uso de técnicas que modelam matematicamente decisões e habilidades antes só dominadas por humanos. Os processos industriais foram os primeiros a serem automatizados e controlados devido ao contexto sócio-econômico e tecnológico dos anos 70 a 90 que promoveu o desenvolvimento da área. Da mesma forma que ocorre em processos industriais, o mantra do controle "Medir, Comparar e Agir" ou o cerne da automação, que é a "Modelagem", podem ser novamente aplicados porém sob uma nova ótica. O medir, que para os sistemas clássicos é tratado através de sistemas de aquisição de dados, passa a ser feito através de sistemas de *data-mining*, que procuram identificar o que é efetivamente uma informação relevante; o comparar, feito pelos elaborados controladores, é substituído pela inteligência artificial, estatística e outras técnicas de reconhecimento de padrões e de representação do conhecimento; e finalmente o agir, desempenhado nos processos industriais pelos atuadores, passa a ser substituído por sistemas inteligentes que utilizam-se do conhecimento para desempenhar o seu propósito. A Automação deixa de ser modelada pelos formalismos determinísticos e passa a se deparar com sistemas complexos nos quais a modelagem simples de diversos agentes que se relacionam determinará um comportamento emergente delimitado por repulsores e atratores. Atuar dentro deste contexto buscando estender conceitos clássicos da área de Controle & Automação para um novo paradigma social, científico e tecnológico é a primeira motivação para o trabalho.

A pertinência da realização de pesquisa na área de Inteligência Artificial é ressaltada pelo fato da busca por padrões de equivalência em bases de dados ter evoluído dos conhecidos sistemas de *data-mining* corporativos para a busca de padrões na internet. Com o crescimento da rede mundial de computadores abriu-se grandes possibilidades para o desenvolvimento de sistemas emergentes e de identificação de conhecimento coletivo. Sendo que, diferentemente das bases corporativas, a internet está ao acesso de todos. O tipo de informação disponível na internet também se diferencia de bases de dados corporativas pelo fato de que a colaboração em massa permite que informações de caráter pessoal (e.g. dados de navegação, *ratings*, *profiles*, etc.) de uma infinidade de pessoas ou agentes possam ser correlacionadas e posteriormente analisadas.

A Filtragem Colaborativa é uma área distinta da Inteligência Artificial tradicional, procurando criar algoritmos que possibilitem o reconhecimento de padrões e o aprendizado de máquinas, conseguindo assim identificar "Comportamentos Emergentes". As estruturas emergentes são padrões que não são criados por um único evento ou regra. Não existe nada que comande o sistema para que ele forme um padrão, mas ao invés disso as interações de cada parte com o ambiente externo geram um processo complexo que leva à ordem [2].

Mesmo com reconhecida aplicabilidade comercial, estes algoritmos baseados em filtragem colaborativa possuem um pouco mais de 10 anos de existência, e portanto estão incipientes em sua total potencialidade tecnológica. Além disto, novas oportunidades aparecem

[2]Os Sistemas Emergentes serão definidos com maior propriedade na seção 1.2.1

com o advento da Web 2.0 onde a colaboração e as redes sociais permitem que novas informações sejam disponibilizadas contribuindo com a evolução de sistemas de recomendação e outros sistemas caracterizados pela emergência. O impacto que a aplicação destas tecnologias pode alcançar fica evidenciado no relatório **Emerging Technologies Hype Cycle 2007** do renomado Instituto Gartner [48], que aponta a área como uma das tecnologias com maior potencial de impacto nos próximos 10 anos, corroborando a pertinência da pesquisa na área.

1.1.1 A Internet no Brasil e no Mundo

Um aspecto que contribui significativamente para a motivação da pesquisa de sistemas associados à Internet é o cenário mundial e brasileiro da utilização de sistemas online. O Brasil possui um público de 39 milhões de usuários de internet (fonte: Ibope NetRatings), 102 milhões de celulares ativos (fonte: INFO Online), que logo migrarão para o 3G e 96,5% de casas com televisão (fonte: Folha Online), que em alguns anos migrarão para a TV Digital – TVD. A convergência digital permitirá que em poucos anos o acesso à internet seja feito em qualquer uma destas plataformas (TVD, celular, computador), contribuindo para que a taxa de crescimento da informação disponível na rede mundial seja ainda maior.

Segundo dados do Ibope NetRatings, o número total de pessoas com idade acima de 16 anos que acessaram a internet no Brasil em 2007 foi de 37 milhões de usuários. Desse total, 20 milhões são considerados internautas residenciais ativos, o que significa um crescimento de 47% em relação a 2006. Além disso, o número de brasileiros conectados à internet residencial em banda larga cresceu para 15,4 milhões, representando 76,4% dos acessos residenciais no País. E não é só o acesso à internet que está se popularizando no Brasil. Segundo balanço da Associação Brasileira da Indústria Elétrica e Eletrônica (Abinee), o mercado brasileiro encerrou o ano de 2007 com vendas de 10,1 milhões de PCs, um crescimento de 23% em relação a 2006. Ou seja, o acesso ao computador e à internet já é uma realidade no país, que se reflete, por exemplo, no aumento do comércio eletrônico. Não apenas a quantidade mas também a qualidade dos usuários brasileiros deve ser destacada: o brasileiro é líder mundial em horas de permanência online na internet com média de 20h39min mensais (NetRatings Jul/2006), é considerado o público mais ativo de sites de relacionamentos e possui umas das maiores taxas de adoção de novidades no mundo online.

Apesar de muito ativo na utilização da internet, o Brasil está longe de atuar no estado da arte do desenvolvimento de tecnologias informacionais online. Isto se deve talvez à complexidade do desenvolvimento destas tecnologias que se caracterizam pela multidisciplinaridade, alta performance e alto grau de inovação. Além disso, enquanto as potências mundiais financiam projetos na área, correndo contra o tempo para obterem o diferencial competitivo que estas tecnologias podem oferecer às suas nações, o Brasil ancora sua percepção de que o desenvolvimento na internet se limita aos sites feitos por *free-lancers* do final da década

de 90. Esta visão impede que o país se beneficie de possuir um povo ávido por tecnologia, perdendo a oportunidade de ditar as regras e padrões de uma nova cultura digital global.

A importância desta cultura digital, já preconizada por muitos, mostra sua força ano após ano quebrando paradigmas culturais graças à força da colaboração e às redes de conhecimento. Esta nova cultura permite que em pouco tempo pessoas fiquem famosas, notícias sejam difundidas e bases de conhecimento sejam geradas. A Wikipedia – a enciclopédia livre – é um dos diversos exemplos desta nova consciência. O poder desta revolução pode mudar a cultura de um povo (como já está sendo sentido na China); pode mudar todo um mercado industrial (como aconteceu na indústria fonográfica com o advento do mp3); ou jornalística (com os blogs e wikis) e estabelecer em pouco tempo marcas fortes como a da empresa Google que em apenas sete anos criou uma marca mais valiosa que as centenárias Ford ou Coca-Cola, e isto é apenas o começo de uma revolução ainda maior.

Certamente nesta revolução teremos dois lados, no primeiro estarão as nações que conduzirão o processo ditando as regras de uma nova era, e no outro as outras nações a margem do processo acatando decisões e importando as tecnologias, informações e conhecimentos existentes. O que determinará em qual lado estará uma nação será o grau de evolução de suas tecnologias informacionais, o grau de instrução de seu povo e o posicionamento mercadológico conquistado.

1.1.2 A Competição *Netflix Prize*

A realização desta pesquisa na área de Sistemas de Recomendação foi motivada pelos argumentos apresentados até aqui, culminando na inciativa de desenvolvimento de um projeto para a recomendação de bares e restaurantes. Entretanto, a dificuldade de validação do algoritmo, a inexistência de uma base de dados grande o suficiente para gerar comportamentos emergentes (não triviais) e a carência de parâmetros que possibilitassem aferir a precisão das predições calculadas dificultavam o desenvolvimento de um sistema efetivo.

Em outubro de 2006 a Netflix Inc. anunciou o **Netflix Prize**, uma competição que pagaria o prêmio de 1 milhão de dólares para aquele que conseguisse desenvolver um algoritmo capaz de fazer predições 10% melhores que o sistema de recomendação proprietário da empresa – o Cinematch ®. Para isto a Netflix forneceu uma gigantesca base de dados com avaliações de diferentes usuários sobre um extenso conjunto de filmes e uma métrica para caracterizar a eficiência do algoritmo.

Com a competição, as dificuldades inerentes ao desenvolvimento de um sistema de recomendação foram amenizadas através de métricas de precisão bem definidas, disponibilização de uma grande base de dados e uma sistemática que uniu os principais desenvolvedores da área no mundo. Através do fórum de competidores, pesquisadores de todos os continentes

trocam suas experiências e suas dúvidas, colaborando e competindo para a evolução dos algoritmos de recomendação. Uma oportunidade ímpar de aprendizado para pesquisadores iniciantes na área.

Desta forma, o trabalho que inicialmente estava direcionado à recomendação de bares e restaurantes, foi redirecionado para o problema de recomendação de filmes, respeitando as regras da competição promovida pela Netflix. Este fato, aliado ao suporte científico proporcionado pelo departamento de Automação e Sistemas, complementou as condições necessárias para a condução de um trabalho na área.

1.2 Contexto Tecnológico-Econômico

Nesta seção é apresentada uma contextualização tecnológica e econômica de novas tendências que surgiram com o mundo digital interconectado. Estas informações auxiliarão na compreensão das influências que promoveram o desenvolvimento científico dos sistemas de recomendação, objeto desta dissertação. A seguir são apresentados os **Sistemas Emergentes** que instigam uma nova abordagem na compreensão dos sistemas computacionais e da própria internet, e o chamado **Mercado de Nicho** que está influenciando o pensamento econômico, mostrando as diferenças nas estratégias de negócios com o advento da Sociedade de Informação e do Conhecimento.

1.2.1 Os Sistemas Emergentes

Emergência é o processo de formação de modelos complexos a partir de regras simples [28]. Este processo pode ser dinâmico (ocorrendo através do tempo), como a evolução do cérebro humano através de milhares de gerações sucessivas, ou pode ocorrer em escalas de tamanhos diversos, como as interações entre os neurônios produzindo um cérebro humano capaz de pensar (mesmo sabendo que neurônios individuais não têm consciência própria). Para um fenômeno ser nomeado emergente ele deve geralmente ser inesperado e imprevisível. Geralmente o fenômeno não existe ou existem apenas alguns traços no nível mais baixo. Assim, um fenômeno direto, como a probabilidade de achar uma uva seca em uma fatia de bolo, geralmente não requer a teoria da emergência para ser explicada. Pode ser no entanto útil considerar a emergência da textura do bolo como um resultado complexo do processo de cozimento e mistura dos ingredientes.

Um comportamento emergente ou propriedade emergente pode aparecer quando uma quantidade de entidades (agentes) simples opera em um ambiente, formando comportamentos complexos no coletivo. A propriedade em si é comumente imprevisível e original, e representa um novo nível de evolução dos sistemas. O comportamento complexo ou as suas propriedades

não são herdados de nenhuma entidade em particular, e também não podem ser previstos ou deduzidos dos comportamentos das entidades em nível baixo. O formato e o comportamento dos bandos de pássaros são bons exemplos de comportamentos emergentes.

Uma razão pela qual o comportamento emergente ocorre está no número de interações entre os componentes de um sistema, que aumenta exponencialmente com o número de componentes, permitindo que uma série de novos e diferentes tipos de comportamentos apareçam. Por exemplo, as possíveis interações entre grupos de moléculas crescem exponencialmente com o aumento do número de moléculas, de modo que é impossível para um computador contar o número de arranjos possíveis, mesmo para um sistema com apenas 20 moléculas.

Por outro lado, apenas a existência de um grande número de interações não é o suficiente para garantir o comportamento emergente. Muitas das interações podem ser previsíveis ou irrelevantes, e muitas podem cancelar as outras. Em alguns casos, um grande número de interações pode de fato trabalhar contra a emergência de comportamentos interessantes, criando uma grande quantidade de "ruído" que elimina qualquer "sinal" emergindo. O comportamento emergente pode precisar ser temporariamente isolado de outras interações antes de ter massa crítica o suficiente para poder se auto-suportar. Portanto, não é apenas o número de conexões que encoraja a emergência; também deve ser considerado o modo como estas conexões estão organizadas. Uma organização hierárquica é um exemplo que pode gerar um comportamento emergente, dependo das regras e da governança existente. O comportamento emergente pode também surgir de estruturas organizacionais mais descentralizadas, como acontece no mercado financeiro. Em muitos casos, o sistema tem que alcançar um nível de diversidade, organização e conectividade antes do comportamento emergente ocorrer.

Sistemas com propriedades emergentes podem parecer não seguir os princípios da entropia e a segunda lei da termodinâmica, pois eles se formam e crescem independente da falta de um comando ou controle central. Isto é possível porque sistemas abertos podem extrair as informações do seu ambiente.

A emergência ajuda a explicar porquê a falácia da divisão é uma falácia. De acordo com a perspectiva emergente, a inteligência emerge das conexões entre os neurônios e, desta perspectiva, não é necessário propor uma "alma" para sustentar o fato de que os cérebros podem ser inteligentes, mesmo pensando que os neurônios individuais que o compõe não o são.

Muitos podem associar a internet a um sistema emergente, porém é importante considerar algumas questões antes de fazer esta afirmação. Alguns sistemas, como a Web, são altamente eficiêntes em fazer conexões, mas pobres em estrutura. As tecnologias que suportam a internet são elaboradas para manusear aumentos drásticos de escala, mas são indiferentes à tarefa de criar uma ordem de nível superior. Entretanto, em meio ao caos de informações disponíveis, alguns observadores começaram a detectar macro-padrões no desenvolvimento da

Web, padrões que são invisíveis a seus usuários e, na maioria das vezes, inúteis. A distribuição dos sites da Web e seus públicos parecem seguir o que é chamado a lei de potência: os dez sites mais populares são dez vezes mais populares do que os próximos cem mais populares, que por sua vez, são dez vezes mais populares que os próximos mil. Outros macro-padrões também foram detectados, porém nenhum deles tornam a Web um sistema mais navegável ou informativo. Eles estão mais perto da complexidade de um floco de neve do que de uma rede neural cerebral: um floco de neve se auto-organiza em formas complicadas, mas é incapaz de se tornar um floco mais esperto ou efetivo, sendo apenas um padrão congelado.

O fato da Web tender mais para conexões caóticas do que para a inteligência emergente não é intrínseco a todas as redes de computadores. Mexendo com algumas suposições subjacentes à Web hoje, seria possível esboçar uma versão alternativa, que potencialmente, seria capaz de imitar a auto-organização dos sistemas emergentes conhecidos. A Web não é inerentemente desorganizada, ela foi construída assim. Se sua arquitetura for modificada, talvez ela seja capaz de alcançar comportamentos emergentes mais complexos. Muito trabalho vem sendo feito neste sentido dentro da área de estudo denominada Web Semantica. Uma mudança que poderia trazer grandes benefícios, seria a inserção do *feedback* para os nodos da internet. Os links baseados em HTML são unidirecionais, ou seja, você pode apontar para uma infinidade de outros sites em sua *homepage*, mas não há como essas páginas saberem que você apontou para elas. Os sistemas auto-organizáveis usam o feedback para evoluir para uma estrutura mais ordenada. Lembrando da analogia com os sistemas clássicos de controle e automação, sabe-se que nenhum sistema pode ser bem controlado se ele não for realimentado, e o *feedback* sugerido na estrutura da internet é justamente esta realimentação.

Mesmo com esta deficiência estrutural é possível desenvolver sistemas emergentes na Web sem com isto, ter que modificar seus protocolos. Brewster Kahle percebeu isto em meados de 1995 quando lançou o seu software Alexa. Este software tinha como objetivo correlacionar sites através de técnicas de Filtragem Colaborativa, de forma a auxiliar os usuários a navegarem na grande rede mundial de computadores. O que o software fazia era dizer: *"o site X é parecido com os sites Y e Z"*. Esta famosa frase ficou conhecida através do site Amazon que indicava livros da mesma forma. Isto não é coincidência, já que a Amazon comprou a empresa Alexa em 1999 e logo após lançou o serviço de recomendação de livros.

O programa Alexa e demais softwares que surgiram na seqüência não faziam nenhuma tentativa direta de simular a consciência ou inteligência humana, mas sim procuravam padrões em números que pudessem ressaltar algum comportamento emergente. Para isto entretanto era necessário que uma quantidade expressiva de pessoas (ou agentes) estivessem interconectados através de uma estrutura conhecida. Assim, nasceram os primeiros sistemas de recomendação na internet que possibilitaram o levantamento dos primeiros comportamentos emergentes úteis dentro da Web.

1.2.2 O Mercado de Nicho

Na economia de escassez, onde a oferta infinita de produtos é algo inexistente, a regra que conduziu os negócios foi a chamada *Regra dos 80/20*, que se aplica a praticamente todos os mercados. Esta regra sugere que 20% dos produtos respondem por 80% das vendas e, muitas vezes, por 100% dos lucros. Percebemos esta regra na busca obcecada pelos "sucessos" em todos os setores, dentre os quais estão: os *hits* na indústria fonográfica, os recordistas de bilheteria no cinema, os *best-sellers* nos livros e os campeões de audiência nas emissoras de rádio e televisão. Durante um século, uma triagem extremamente seletiva só deixava passar o que tinha condições de se transformar em um campeão de vendas, para utilizar de maneira mais eficiente possível as dispendiosas e escassas prateleiras, telas de cinema, canais de televisão e rádio. Desta forma estabeleceu-se uma **cultura de massa** que teve seu apogeu nos anos 70 e 80, quando as pessoas tinham acesso a meia dúzia de canais de TV, cujos principais programas eram vistos por todos, e a três ou quatro estações de rádio que impunham boa parte das músicas a serem ouvidas. As pessoas assistiam aos mesmos filmes de grande sucesso nos cinemas e recebiam as notícias pelos mesmos jornais e noticiários.

Embora ainda estejamos obcecados pelos "sucessos" do momento, esses já não são mais a força econômica de outrora. Com o advento da internet nos anos 90, a dominância da cultura de massa encontrou o início de seu declínio. Em uma nova era de consumidores em rede, na qual tudo é digital, a economia de distribuição está sendo mudada de forma radical à medida que a Internet absorve e distribui quase tudo, transmutando-se em loja, teatro e difusora por uma fração mínima de custo adicional. Os consumidores que antes avançavam como manada em uma única direção, agora se dispersam nas infinitas possibilidades de escolhas à distância de um clique, fragmentando o mercado em inúmeros nichos.

Esta mudança é percebida pelos números do mercado. Quase todos os cinqüenta álbuns musicais mais vendidos de todos os tempos foram gravados nas décadas de 1970 e 1980 (apogeu da cultura de massa) e nenhum deles é dos últimos cinco anos. A receita dos campeões de bilheteria de Hollywood diminuiu em dois dígitos em 2005, refletindo a realidade de que a quantidade de pessoas que vão aos cinemas ver o mesmo filme está caindo, apesar do aumento da população. Por outro lado, a mesma indústria de cinema aumenta a cada ano o número de filmes lançados, sem com isto aumentar o número de filmes que dão prejuízo. Todos os anos, as grandes redes de televisão perdem cada vez mais público para centenas de canais a cabo que se concentram em nichos do mercado. Os homens com idade entre 18 e 34 anos, o público mais almejado pelos anunciantes, estão começando a desligar de vez a televisão, dedicando parcelas cada vez maiores de tempo às telas eletrônicas da Internet e videogames.

Ainda existe demanda para a cultura de massa, mas esse já não é mais o único mercado. Os *hits*, hoje, competem com inúmeros mercados de nicho que variam de tamanho. E os consumidores estão exigindo cada vez mais opções de escolhas. A era do tamanho único está chegando ao fim, e em seu lugar está surgindo algo novo, o mercado de variedades, o mercado

de nicho. Esta nova visão muda completamente o entendimento do "sucesso" ou "fracasso" de um determinado produto, pois a maioria dos filmes não é recordista de bilheteria, a maioria das músicas não alcança as paradas de sucesso, a maioria dos livros não vira best-sellers e a maioria dos programas de televisão nem chega perto dos principais índices de audiência e nem se destina ao horário nobre. No entanto, muitas dessas produções atingem milhões de pessoas em todo o mundo conseguindo consideráveis lucros.

Neste novo contexto, no qual é possível fazer ofertas de muitos produtos, cria-se uma quantidade enorme de mercados de nicho que somados podem ser tão grandes ou até maiores que o mercado de massa. Isto foi percebido por várias empresas que reinventaram seus negócios e hoje se beneficiam da chamada *Regra dos 98%*. Esta regra sugere que independentemente do tamanho de seu catálogo de ofertas, 98% dos produtos poderão representar as vendas de sua empresa, contanto que haja mecanismos de busca ágeis e fáceis de se utilizar. A Apple divulgou que cada uma das então 2 milhões de faixas do iTunes é vendida pelo menos uma vez por trimestre. A Netflix estima que 95% de seus mais de 100 mil DVDs são alugados no mínimo uma vez por trimestre. O mesmo ocorre com a Amazon, que vende 98% de seus 200 mil livros no mesmo período. Os produtos menos procurados, assim como acontece no varejo físico, continuam vendendo pouco, mas estas vendas de poucos itens se tornaram tão numerosas que no todo constituem um grande negócio.

Chris Anderson, pesquisador e editor-chefe da revista Wired, focada em tecnologia e negócios, pesquisou o fenômeno do mercado de nichos e levantou as curvas de vendas das empresas de comércio varejista online, percebendo que sua distribuição era idêntica as "distribuições de cauda longa" e assim batizou um novo conceito no mundo de negócios: **The Long Tail** - (Cauda Longa em português). Este conceito foi apresentado em um artigo em 2004 e depois, com a colaboração de pesquisadores de Stanford, MIT e Harvard, transformou-se em livro [2] que virou referência para os negócios baseados em mercados de nicho.

A teoria da Cauda Longa pode ser resumida no movimento que a nossa cultura e a nossa economia está fazendo, migrando do foco nos "sucessos" relativamente pouco numerosos, que se encontram no topo da curva de demanda, para avançar em direção a uma grande quantidade de nichos na parte inferior, ou na cauda, da curva de demanda. Em uma era sem as limitações do espaço físico nas prateleiras e de outros pontos de estrangulamento da distribuição, bens e serviços com alvos estreitos podem ser tão atraentes em termos econômicos quanto os destinados ao grande público.

Porém apenas a inexistência das limitações de distribuição não permite a criação de um mercado de cauda longa, é necessário que a oferta seja seguida pela demanda. A Cauda Longa pode começar com um milhão de nichos, mas isto só terá algum significado econômico se estes nichos forem procurados e encontrados por pessoas que os almejem. Alguns fatos contribuem para que um mercado de Cauda Longa se estabeleça [2]. Em praticamente todos os mercados há mais nichos do que mercados de massa. Essa desproporção aumenta em ta-

xas exponenciais à medida que as ferramentas de produção se tornam mais baratas e difusas. Os custos de atingir esses nichos estão caindo drasticamente. Graças à distribuição digital, poderosas tecnologias de busca e a difusão da banda larga, os mercados online estão reconfigurando a economia do varejo, podendo oferecer uma maior variedade de produtos. A simples oferta de maior variedade, contudo, não é suficiente para descolar a demanda. Os consumidores devem dispor de maneiras para encontrar os nichos que atendam suas necessidades e interesses particulares. Um vasto espectro de ferramentas e técnicas - como **recomendações e classificações** - são eficazes para este propósito. Tais "**filtros**" são capazes de impulsionar a demanda ao longo da Cauda.

A redução de custos para que a demanda possa alcançar a oferta envolve, independentemente do mercado, a atuação de uma ou mais de três forças poderosas:

- **Democratização das ferramentas de produção**: O melhor exemplo desta força é o computador pessoal, que pôs todas as coisas, desde as máquinas de impressão até os estúdios de produção de filmes e músicas, nas mãos de todos. Os PCs permitiram a façanha de que qualquer indivíduo pudesse fazer o que há poucos anos atrás era possível de ser feito apenas por profissionais, multiplicando os potenciais "produtores". O resultado é que o universo de conteúdo disponível hoje está crescendo mais rápido do que em qualquer outra época. Essa característica atua na Cauda alongando-a, conforme a oferta de bens aumenta;

- **Democratização das ferramentas de distribuição**: Esta segunda força está reduzindo os custos de consumo dos bens. O PC transformou todas as pessoas em produtores e editores, mas foi a Internet que converteu todo mundo em distribuidores. Desta forma, torna-se mais barato alcançar mais pessoas aumentando efetivamente a liquidez do mercado na Cauda, o que por sua vez, se traduz em consumo, elevando efetivamente o nível da linha de vendas e ampliando a área sob a curva;

- **Ligação entre oferta e procura**: Esta última força busca unir esta grande quantidade de novos consumidores a uma infinidade de bens agora disponíveis com maior facilidade. Esta força pode assumir desde a forma de um sistema de busca, como o Google, até a de um sistema de recomendações de músicas feitas pelo iTune, de livros pela Amazon, ou de filmes pela Netflix. Além destas formas "automáticas," a propaganda boca a boca ampliada pelos blogs e *reviews* de clientes também é uma demonstração desta força.

Em economia, custo de busca é qualquer dificuldade que interfira na descoberta do que se tem em mira. Alguns destes custos são não-monetários como perda de tempo, aborrecimentos e confusão. Outros têm expressão financeira como comprar algo errado, de má qualidade ou pagar preço excessivo por não encontrar alternativas mais baratas. Em geral, os outros consumidores são quem fornecem a melhor orientação, pois seus incentivos estão

alinhados com aqueles que as pessoas estão procurando. O Netflix e o Google exploram a sabedoria coletiva dos consumidores, observando-os aos milhões e traduzindo as informações daí decorrentes em resultados de busca ou em recomendações relevantes. O desenvolvimento de tecnologias que ligam os consumidores é o que impulsiona a demanda da cabeça para a cauda da curva, permitindo que interesses se desmembrem em comunidades de afinidades cada vez mais estreitas, que se aprofundam cada vez mais nas respectivas preferências, como sempre ocorre quando as mentes atuam em conjunto.

Para demonstrar uma distribuição de Cauda Longa é apresentado na figura 1.1 um gráfico com as vendas de livros da empresa Amazon extraído de [18]. Neste gráfico podemos perceber que a área cinza que representa as vendas de livros menos populares está superando o número das vendas dos *best-sellers*, demonstrando a importância da economia da cauda-longa para os novos negócios.

Source: Brynjolfsson, Hu, and Smith, "Consumer Surplus in the Digital Economy," *Management Science*, November 2003.

Figura 1.1: Vendas da Amazon - Novembro de 2003

1.3 Objetivos

O presente trabalho objetiva contribuir com a pesquisa na área de sistemas de recomendação, particularmente sistemas baseados em Filtragem Colaborativa, buscando promover o desenvolvimento de tecnologias informacionais no Brasil. Para isto, propõe-se desenvolver um sistema completo para a competição promovida pela empresa Netflix, procurando obter uma precisão melhor que a do sistema proprietário da empresa - o Cinematch ®. Têm-se ainda como objetivos secundários para este trabalho:

1. Alcançar um tempo computacional viável para uma aplicação comercial;
2. Compreender o estado da arte da pesquisa em sistemas de recomendação;

1.4 Metodologia

O método científico utilizado para o desenvolvimento deste trabalho foi iniciado com a caracterização do problema em questão, feita através da observação, estudo e síntese dos sistemas de recomendação como tema genérico, e sua especialização dentro do domínio do assunto de filmes, procurando conhecer diferentes sistemas de recomendação e diferentes medidas de precisão dos mesmos.

Na seqüência foram levantadas hipóteses e objetivos para o trabalho. Estas hipóteses consideraram uma predição completa para a competição promovida pela Netflix, melhorias em tempo computacional de algoritmos já existentes e melhorias conceituais para alcance de maior precisão. Estes objetivos foram descritos neste capítulo na seção 1.3.

Para que as hipóteses pudessem ser comprovadas e alcançadas com êxito, o desenvolvimento do algoritmo respeitou regras da competição *Netflix Prize* que forneceu, além da base de dados para os testes, uma sistemática para a aferição dos resultados e a possibilidade de comparação com outros algoritmos desenvolvidos por pesquisadores de todo mundo. Desta forma, todo experimento realizado ficou livre de tendência ou de qualquer manipulação dos resultados.

A fundamentação teórica foi embasada em fontes de estudo como livros, artigos, blogs e fóruns sobre o assunto, além da troca de experiência com outros pesquisadores da área. A análise e a interpretação dos resultados foram favorecidas pela sistemática da competição, tornando-se simples e objetivas através da análise do RMSE (Root Mean Square Error). Técnicas de *cross-validation* permitiram o ajuste ótimo de parâmetros e a identificação de *overfitting*.

Finalmente os resultados foram descritos nesta presente dissertação, dividida em 6 capítulos que descrevem todas as fases de desenvolvimento do trabalho. No capítulo introdutório foram apresentadas as motivações e os objetivos do desenvolvimento do tema. A segunda parte da dissertação, composta pelos capítulos 2 e 3, procura fundamentar os conhecimentos teóricos necessários para a implementação de um sistema de recomendação. No capítulo 2 – "Sistemas de Recomendação" – é apresentado o estado da arte destes sistemas. No capítulo 3 – "k-Nearest Neighbor" – são apresentados os fundamentos do algoritmo kNN, sua definição formal, vantagens e desvantagens de sua utilização e todas as etapas de uma implementação genérica. Na terceira parte da monografia, composta pelos capítulos 4, 5 e 6, é apresentada a contribuição do autor para a pesquisa na área, na qual é descrita a implementação de um algoritmo baseado em kNN para a competição promovida pela Netflix e os resultados alcançados. Por fim no capítulo 7 –"Conclusões e Perspectivas"– são apresentadas as conclusões e perspectivas do trabalho.

1.5 Conclusão

A utilização de sistemas que filtram informações tende a se intensificar na medida em que a quantidade de informações e de pessoas que utilizam a internet aumenta a cada ano. A teoria da Cauda Longa categoriza estes sistemas como uma das três principais forças que reduzem o custo para se alcançar os nichos econômicos. Nesta teoria, um novo modelo econômico explica o fenômeno do comércio varejista online e do novo modelo de mídias e propaganda presente na Web que representou o crescimento exponencial das empresas do setor.

Comportamentos emergentes ocorrem na internet na medida em que milhões de interações ocorrem todos os dias sobre sua estrutura de rede. Sistemas de recomendação foram os primeiros sistemas a identificar de forma automática comportamentos emergentes úteis na rede mundial de computadores. Por este motivo e pelo fato dos sistemas de recomendação contribuirem para a viabilização dos mercados de cauda-longa, a área é apontada como uma das que possuem maior potencial de impacto nos próximos 10 anos, corroborando a pertinência da pesquisa na área.

O presente trabalho objetiva contribuir com a pesquisa na área de sistemas de recomendação, particularmente sistemas baseados em Filtragem Colaborativa, buscando promover o desenvolvimento de tecnologias informacionais no Brasil. Para isto, propõe-se desenvolver um sistema completo para a competição promovida pela empresa Netflix, procurando obter uma precisão melhor que a do sistema proprietário da empresa – o Cinematch ®.

Neste capítulo introdutório foram apresentadas as motivações para o desenvolvimento do tema, os objetivos e metodologia do trabalho e uma breve contextualização sócio-econômica e tecnológica do mundo atual, que justificaram a pertinência da pesquisa na área de *data-mining* e filtragem colaborativa.

Capítulo 2

Sistemas de Recomendação

Logo que, numa inovação, nos mostram alguma coisa de antigo, ficamos sossegados

— Friedrich Nietzche

Na última década, muito trabalho vem sendo realizado tanto pela indústria quanto pela academia no desenvolvimento de novos métodos para Sistemas de Recomendação. O interesse na área é grande, graças à vasta aplicabilidade em problemas que ajudam usuários a lidar com a abundância de informações. Este capítulo apresenta um panorama da área de sistemas de recomendação e descreve os atuais métodos que são usualmente classificados pela literatura dentro de três principais categorias: baseados em conteúdo, filtragem colaborativa e métodos híbridos. Uma explicação mais detalhada é apresentada dentro da categoria filtragem colaborativa, objeto de estudo desta dissertação, cujo estado da arte com os métodos desenvolvidos nos últimos anos são discorridos em maiores detalhes.

2.1 Contextualização

As pessoas podem hoje escolher entre centenas de canais de televisão, milhões de vídeos, livros, CDs, entre outras possibilidades disponíveis e facilmente acessíveis através da internet. Porém, não apenas existe uma grande variedade de escolhas, mas também uma grande variedade de qualidade. Avaliar todas estas possibilidades, todavia, continua levando o mesmo tempo e esforço que nos tempos em que esta disponibilidade não existia. O ser humano não evoluiu na mesma velocidade em que cresceu a oferta de informação. Assim, muitas vezes as pessoas possuem pouca ou quase nenhuma experiência pessoal para realizar escolhas entre

as várias alternativas que lhes são apresentadas. Outras vezes é impossível avaliar todas as possibilidades a não ser que restrinjam severamente o seu campo de escolha, ou então, que façam a utilização de filtros de informações ou peçam recomendações para outras pessoas.

Recomendações podem ser adquiridas de forma direta (*word of mouth*) [50], através de cartas de recomendações, muito utilizadas em pleito de empregos, de críticos (e.g. opiniões de filmes e livros em jornais e revistas), e também de sistemas que indicam preferência. Talvez a categoria mais simples destes sistemas seja as listas de preferências (Top 10, Top 100 etc.), nas quais as preferências de um conjunto de pessoas são demonstradas através de um ranking. Recomendações podem ser extraídas ao escolher, por exemplo, um entre os "dez livros mais vendidos" em uma loja online. Este tipo de sistema, entretanto, não faz nenhuma recomendação personalizada e também não indica necessariamente os livros mais adequados àqueles que se está procurando.

Em resposta a esta dificuldade, sistemas de recomendação computacionais emergiram como forma de suporte, mediação e automação do processo de recuperação da informação. A evolução destes sistemas e o fato deles trabalharem com bases grandes de informações, permitiram que recomendações emergentes (não triviais) pudessem ser alcançadas, proporcionando ainda maior credibilidade que uma recomendação humana.

Os proponentes de um dos primeiros sistemas de recomendação denominado Tapestry [23], desenvolvido no início dos anos 90s, criaram a expressão "Filtragem Colaborativa", visando designar um tipo de sistema específico no qual a filtragem da informação era realizada com o auxílio humano, ou seja, através da colaboração entre os grupos interessados. Vários pesquisadores acabaram adotando esta terminologia para denominar qualquer tipo de sistema de recomendação subseqüente. Resnick, no seu artigo [50], defendeu o termo "*sistemas de recomendação*" como terminologia mais genérica do que filtragem colaborativa, já que sistemas de recomendação podem existir sem nenhuma colaboração entre as pessoas.

2.2 Os Sistemas de Recomendação como Área de Pesquisa

As raízes dos sistemas de recomendação podem ser encontradas nos trabalhos extensivos das ciências cognitivas [51], teoria de aproximação [45], recuperação da informação [53], teoria de previsões [3] e também possuem influências das ciências de administração e marketing [36], [31]. A área de sistemas de recomendação emergiu como uma área de pesquisa independente, no meio da década de 90, quando a partir do sistema de recomendação Tapestry [23], os pesquisadores passaram a focar em problemas de recomendação que explicitamente invocavam estruturas de avaliação (*ratings*). A partir deste momento surgiram os primeiros artigos sobre filtragem colaborativa [24], [49], [56], nos quais o problema foi formalizado primeiramente e desde então estudado intensamente.

2.2.1 Formalização do Problema

Definição 2.2.1 *: Seja C o conjunto de todos os usuários de um determinado sistema, e seja S o conjunto de todos os possíveis itens que podem ser recomendados como livros, filmes, restaurantes etc. Seja u a função utilidade que mede o quão útil é um determinado item s para um determinado usuário c, i.e., $u : C \times S \longrightarrow R$, onde R é um conjunto totalmente ordenado. Então, para cada usuário $c \in C$, procura-se um item $s' \in S$ que maximiza a utilidade do usuário. Isto pode ser expressado pela equação abaixo:*

$$\forall c \in C, \quad s'_c = \arg\max_{s \in S} u(c, s) \tag{2.1}$$

Em um sistema de recomendação a utilidade de um item é geralmente representada por uma avaliação que indica o quanto um determinado usuário gosta de um item (e.g. Rafaela deu ao filme "Harry Potter" a nota 7 em 10). No entanto, conforme descrito na definição 2.2.1, a função de utilidade pode ser uma função arbitrária.

Cada elemento do espaço de usuários C pode ser definido através de um *profile* que inclui as características do usuário, como a sua idade, sexo, estado civil, renda, etc. No caso mais simples, o profile pode conter um único elemento como o User ID. Da mesma forma, cada item do espaço S pode ser definido por um conjunto de características. Por exemplo, na recomendação de filmes, na qual S é a coleção de filmes, cada filme pode ser representado não apenas pelo seu ID, mas também pelo seu título, gênero, diretor, ano de lançamento, atores principais, etc.

O problema central dos sistemas de recomendação reside no fato da utilidade u geralmente não ser definida em todo o espaço $C \times S$, mas apenas em um subconjunto deste. Isto significa que u precisa ser extrapolado para todo o espaço $C \times S$. Geralmente em sistemas de recomendação, a utilidade é definida através de avaliações, e estas são definidas apenas nos itens previamente avaliados pelos usuários. Deste modo, o algoritmo de recomendação deve ser capaz de estimar (predizer) as avaliações não realizadas para os pares *usuário-item* e de fazer recomendações apropriadas baseadas nestas predições.

A extrapolação de avaliações conhecidas para avaliações inexistentes é geralmente feita pela 1) especificação de *heurísticas* que definem a função utilidade e validam empiricamente sua performance e 2) pelas *estimativas* da função utilidade através da otimização de algum critério de performance como o *mean square error*. Assim que avaliações desconhecidas são estimadas, o sistema de recomendação seleciona aquelas com maiores avaliações para serem recomendadas.

A predição de avaliações de itens ainda não avaliados pode ser feita de diferentes formas utilizando métodos de aprendizado de máquinas, teorias de aproximação e vários tipos de heurísticas. Os sistemas de recomendação são classificados de acordo com o método de

predição utilizado, o que auxilia na pesquisa da área. Esta classificação é feita usualmente em três categorias propostas inicialmente por [58]. Contudo, a explicação de cada uma delas vem sendo complementada pelos diversos trabalhos conduzidos desde então [12], [55], [34], [57]:

- *Recomendações Baseadas em Conteúdo*: O usuário receberá recomendações de itens similares a itens preferidos no passado;
- *Recomendações Colaborativas*: O usuário receberá recomendações de itens que pessoas com gostos similares aos dele preferiram no passado. Este método é subdividido em duas catergorias: a primeira chamada de *memory-based*, e a segunda chamada de *model-based*;
- *Métodos Híbridos*: Estes métodos combinam tanto estratégias de recomendação baseadas em conteúdo quanto estratégias baseadas em colaboração.

Maiores detalhes sobre cada uma destas categorias serão apresentados nas sessões 2.3, 2.4 e 2.5.

2.2.2 Pesquisas Complementares

A fim de complementar os sistemas de recomendação que predizem valores absolutos de avaliações que indivíduos dariam para itens ainda não visualizados, cabe citar os trabalhos feitos em *filtragem baseada em preferências* [14], [19], [46], [27] através da qual a predição *relativa* das preferências dos usuários são calculadas. Por exemplo, em uma aplicação de recomendações de filmes, técnicas de filtragem baseadas em preferências buscam prever a correta ordem relativa dos filmes, ao invés das notas individuais de cada um. Pelo fato destes tipos de sistemas fugirem da definição formal proposta para este trabalho, eles não serão vistos em detalhes.

Alguns autores, como Montaner em [34], destacam que existe um tipo especial de filtragem de informação denominada filtragem demográfica. A filtragem demográfica utiliza a descrição de um indivíduo para aprender o relacionamento entre um item particular e o tipo de indivíduo que poderia se interessar por ele. Este tipo de abordagem utiliza as descrições das pessoas para conseguir aprender o relacionamento entre o item e a pessoa. O perfil de usuário é criado pela classificação dos usuários em estereótipos que representam as características de uma classe de usuários. Como exemplo, o autor cita o método implantado em LyfeStyle Finder [29]. Porém, analisando este sistema é possível classificá-lo como um método híbrido que se utiliza de métodos colaborativos e baseados em conteúdo para fazer suas predições, não necessariamente definindo uma nova categoria e, por este motivo, não foi considerado na classificação apresentada na seção anterior.

2.2.3 Avaliação de Sistemas de Recomendação

Na caracterização dos sistemas de recomendação como uma área de pesquisa científica, é fundamental entender as metodologias para avaliação dos sistemas. Uma forma eficaz para a avaliação de sistemas de recomendação é através da comparação das predições realizadas com as respectivas avaliações reais de usuários para as instâncias preditas. Este tipo de avaliação era considerada custosa e difícil antes do surgimento das grandes empresas de comércio eletrônico, que possuem grandes bases de dados com milhares de usuários e itens. Desta forma, com a supressão de uma determinada avaliação já feita, o sistema de recomendação é utilizado para fazer uma determinada predição [1] e, então, esta é comparada com a avaliação real.

A obtenção de métricas para aferir o desempenho de um sistema de recomendação antes de uma ampla utilização comercial é fundamental para verificar se as predições realizadas serão adequadas para o propósito em questão. A seguir são apresentadas três das principais métricas utilizadas na literatura para a avaliação de sistemas de recomendação. Elas foram tiradas de uma lista mais completa proposta por dois trabalhos [34] e [54]:

- Precisão I (*Precision*) [34]: A precisão de um sistema indica a quantidade de itens recomendados que são do interesse do usuário em relação ao conjunto de todos os itens que lhe são recomendados.
- Precisão II [37]: A precisão de um sistema indica o quanto uma predição é próxima da avaliação real feita pelo usuário. Esta aferição só é possível de ser feita conforme ensaio explicado no início desta seção.
- Recuperação (*Recall*): O índice de recuperação indica a quantidade de itens de interesse do usuário que aparecem na lista de recomendações.
- Cobertura (*Coverage*): A cobertura é a proporção de itens que são passíveis de serem recomendados em relação ao conjunto de todos os itens conhecidos pelo sistema de recomendação.

2.3 Métodos Baseados em Conteúdos - MBC

Recomendações baseadas em conteúdos (*content-based* em inglês) empregam a comparação entre o conteúdo dos itens, de forma a recomendar itens parecidos àqueles que o usuário gostou no passado. Essa abordagem tem suas origens nas técnicas empregadas em sistemas de recuperação de informações [4], [53] e nas pesquisas em filtragem de informações

[1] Importante ressaltar aqui que para esta avaliação ser válida, o sistema deve ser treinado com as notas suprimidas para não facilitar inferências que auxiliem no resultado final.

[5]. Pelo fato destas áreas de pesquisa terem se desenvolvido bastante na área de aplicações baseadas em textos, existem hoje vários sistemas com foco em recomendação de itens contendo informações textuais, como documentos, web-sites e notícias.

Nos MBCs, a utilidade $u(c, s)$ de um item s para um usuário c é estimada baseada nas utilidades $u(c, s_i)$ já assinaladas pelo usuário c aos itens $s_i \in S$ e que são similares ao item s. Para exemplificar, em uma aplicação de recomendação de filmes, no intuito de recomendar filmes para o usuário c, o algoritmo baseado em conteúdo tenta entender as características dos filmes que o usuário c avaliou no passado com notas altas (atores específicos, diretores, gêneros, etc.). Assim, apenas os filmes que possuem grande grau de semelhança com estes filmes avaliados com altas notas serão recomendados.

A técnica baseada em conteúdo busca uma compreensão passível de descrição tanto dos itens quanto dos usuários. A descrição de interesses do usuário pode ser obtida através de informações fornecidas por ele próprio ou através de suas ações como a seleção, visualização ou aquisição de itens. Muitas ferramentas desta abordagem aplicam técnicas como indexação de freqüência de termos [53]. Neste tipo de indexação, informações dos documentos e necessidades dos usuários são descritas por vetores com uma dimensão para cada palavra que ocorre na base de dados. Cada componente do vetor corresponde à freqüência que uma respectiva palavra ocorre em um documento ou na consulta do usuário. Assim, os vetores dos documentos que estão próximos aos vetores da consulta do usuário são considerados os mais relevantes para ele.

Além das heurísticas tradicionais que são baseadas na sua maioria em métodos de recuperação de informação, outras técnicas para sistemas baseados em conteúdos também vêm sendo usadas, como classificadores bayesianos [35], [41] e também várias técnicas de aprendizado de máquinas, incluindo clustering, decision trees e redes neurais artificiais [41]. Estas técnicas diferem da maioria dos métodos baseados em recuperação da informação, já que calculam a utilidade das predições não baseadas na fórmula que implementa a heurística (e.g. *cosine similarity*), mas sim no modelo aprendido através de técnicas estatísticas e de aprendizado de máquinas sobre os dados de treinamento.

Os sistemas de recomendação baseados em conteúdo podem, em princípio, ser utilizados em qualquer domínio no qual se deseja gerar recomendações, mas na prática este método costuma ser utilizado em domínios cujos itens tenham quantidade considerável de informação armazenada de forma textual. Isso ocorre pelo fato de que a técnica baseada em conteúdo se limita às features associadas explicitamente aos objetos que se deseja recomendar. E para se ter uma quantidade suficiente de features, o conteúdo precisa estar em uma forma que seja possível analisá-lo automaticamente por um computador (e.g. um texto). Por isso a pesquisa associada à extração de conhecimento de informações textuais é bastante avançada e a extração de conteúdos multimídias como videos, áudio e outros formatos ainda é bastante incipiente.

Outra desvantagem inerente à limitação de análise do conteúdo encontra-se no caso da existência de dois itens diferentes descritos pelas mesmas features. Neste caso, eles são indistinguíveis e o conteúdo deveria ser revisto para ser descrito por um conjunto mais completo de features.

Além das limitações de análise do conteúdo, os MBCs também possuem problemas associados a super-especialização. Como os sistemas *content-based* fazem recomendações baseados nas notas altas já dadas, o usuário pode ficar limitado a receber recomendações apenas de itens similares aos já avaliados no passado, dependendo de como o sistema for estruturado. Desta forma, uma pessoa que jamais experimentou uma comida açoriana, poderá não receber uma recomendação de um restaurante açoriano mesmo que exista um excelente em sua cidade. Este problema vem sendo estudado em outros campos de estudo e algumas técnicas como inserção de aleatoriedade, generalização e uso de algoritmos genéticos procuram amenizar este tipo de problema.

O problema da super-especialização não ocorre apenas pelo fato de um MBC não conseguir recomendar um item que é diferente de tudo o que usuário já viu. Em certas aplicações, quando um item é muito similar a outro já visto pelo usuário, este não deve ser recomendado. Este caso pode ser visto em um sistema de recomendação de notícias que recomendaria vários textos que trazem o mesmo fato. Assim, a *diversidade* de recomendações é sempre desejável em um sistema de recomendações. Idealmente, o usuário deveria ser apresentado a uma variedade de opções e não a um conjunto homogêneo de escolhas.

O usuário precisa avaliar um número suficiente de itens antes que o sistema de recomendação baseado em conteúdo consiga realmente fazer recomendações associadas com as preferências do usuário. Este problema, conhecido como o *Problema do Novo Usuário*, faz com que usuários com poucas avaliações não consigam receber recomendações precisas.

2.4 Métodos Baseados em Colaboração - Filtragem Colaborativa - FC

O sistema Grundy [51] é considerado por alguns autores como o primeiro sistema de recomendação que se tem notícia. Neste sistema, livros são recomendados com base em estereótipos construídos através da análise das características dos usuários. Alguns anos depois, o sistema Tapestry [23] fazia a recomendação de documentos baseado em usuários com gostos semelhantes. Esta semelhança é encontrada através da declaração manual das preferências feita por cada usuário (criação de filtros). Muitos autores [21], [47] citam o Tapestry como o primeiro sistema de recomendação construído. Além disto, em [44], o sistema Tapestry é classificado como um sistema híbrido pelo fato da criação das regras ser

uma forma de caracterização do conteúdo (modelo baseado em conteúdo) e da colaboração das recomendações ser uma característica de sistemas de filtragem colaborativa.

Uma nova evolução aconteceu com a criação de alguns sistemas como o GroupLens [49], Vídeo Recommender [24] e Ringo [56], que *automatizaram* o processo de predição. Nestes sistemas, a grande diferença com o Tapestry é que neste último a vizinhança precisava ser definida manualmente através da criação de regras pelos usuários, enquanto que no GroupLens, Ringo e Video Recommender esta vizinhança é criada automaticamente por algoritmos de correlação. Com a automatização do processo de predição e a evolução dos sistemas de informação foi possível o desenvolvimento de sistemas de recomendação para fins comerciais. Uma das aplicações mais conhecidas é a recomendação de livros feita pelo site da Amazon [33].

Diferentemente dos sistemas baseados em conteúdo, os sistemas baseados em colaboração (ou baseados em *filtragem colaborativa*) tentam predizer a utilidade dos itens para um usuário particular com base nos itens previamente avaliados por *outros* usuários que também participam do sistema. Mais formalmente, a utilidade $u(c, s)$ do item s para o usuário c é estimada baseada nas utilidades $u(c_j, s)$ assinaladas para o item s pelos usuários $c_j \in C$ que são "similares" ao usuário c.

Apesar de existirem muitos métodos já usados para a implementação de sistemas baseados em filtragem colaborativa, os mais conhecidos e estudados são os baseados na formação de uma vizinhança entre os usuários ou itens [55]. Nestas abordagens procura-se identificar uma *similaridade* entre itens ou entre usuários definindo uma *vizinhança* sobre a instância a ser predita e em seguida calcula-se a predição do valor que o usuário alvo da recomendação daria para o item candidato à recomendação. Esta predição baseia-se nos valores das avaliações que os vizinhos do usuário alvo deram ao respectivo item.

Conforme proposto em [12], algoritmos baseados em filtragem colaborativa podem ser agrupados em duas classes gerais: algoritmos baseados em memória (também chamado *heuristic-based approach*) e os algoritmos baseados em modelo. A escolha por uma abordagem ou outra depende da natureza dos dados a serem recomendados. Nas sessões a seguir são apresentadas as características de cada abordagem.

2.4.1 Algoritmos Baseados em Memória

Algoritmos baseados em memória são essencialmente heurísticas que fazem predições de avaliações baseados no conjunto total de todos os itens já avaliados por todos os usuários. Formalmente, o valor de uma avaliação desconhecida $r_{c,s}$ do usuário c para um item s é usualmente computada como uma função de agregação das avaliações de alguns outros (usualmente os N mais similares) usuários para o mesmo item s:

$$r_{c,s} = \underset{c' \in \hat{C}}{\text{aggr}}\, r_{c',s} \tag{2.2}$$

onde $\hat{C}$ denota o conjunto dos N usuários mais similares ao usuário c e que avaliaram o item s (N pode variar entre 1 e todos os usuários do conjunto e são denominados usualmente por "vizinhos"). Alguns exemplos de uma função agregação são:

$$r_{c,s} = \frac{1}{N} \sum_{c' \in \hat{C}} r_{c',s} \tag{2.3}$$

$$r_{c,s} = k \sum_{c' \in \hat{C}} sim(c, c') \times r_{c',s} \tag{2.4}$$

$$r_{c,s} = \bar{r}_c + k \sum_{c' \in \hat{C}} sim(c, c') \times (r_{c',s} - \bar{r}_{c'}) \tag{2.5}$$

onde o multiplicador k serve como um fator de normalização e é usualmente estabelecido como $k = 1/\sum_{c' \in \hat{C}} |sim(c, c')|$ e $\bar{r}_c$ denota a média da avaliação do usuário c na equação 2.5 e é definido como[2]: $\bar{r}_c = (1/|S_c|) \sum_{s \in S_c} r_{c,s}$, onde $S_c = \{s \in S | r_{c,s} \neq \oslash\}$

A função de agregação pode ser uma simples média aritmética como sugerido na equação 2.3, entretanto a abordagem mais utilizada é a soma ponderada descrita na equação 2.4. A medida de similaridade entre os usuários c e c', $sim(c, c')$, é essencialmente a distância medida entre os usuários, sendo esta utilizada dentro de uma função de peso para a predição de $r_{c,s}$. Assim, quanto mais similares forem os usuários c e c', mais a avaliação $r_{c',s}$ contribuirá para a predição de $r_{c,s}$. Um problema na utilização da soma ponderada é que esta não leva em consideração que usuários diferentes podem utilizar as escalas de avaliação de forma diferente. A média ponderada ajustada, descrita na equação 2.5, é bastante utilizada para atender a esta limitação. Nesta abordagem, em vez de utilizar os valores absolutos das avaliações, a soma ponderada utiliza a diferença da média da avaliação do usuário correspondente.

Vários métodos vêm sendo propostos para o cálculo da função da similaridade. De forma geral a similaridade entre dois usuários é baseada nas avaliações que *ambos* tenham feito para determinado item. Em grandes bases de dados, a existência de vários itens avaliados por ambos usuários, chamados de *itens-comuns* (ou *common-items* em inglês), é fundamental para estabelecer se uma correlação é ou não confiável.

Os dois métodos mais populares para o cálculo da similaridade são a *correlação* e o baseado em *coseno*. Para apresentar estes métodos, considerar-se-á S_{xy} como o conjunto de todos os itens avaliados tanto por x quanto por y, ou seja, S_{xy} representa os *itens-comuns* entre x e y. Mais formalmente: $S_{xy} = \{s \in S | r_{x,s} \neq \oslash \;\&\; r_{y,s} \neq \oslash\}$. Em sistemas de recomendação colaborativos, S_{xy} é geralmente usado como um resultado intermediário para

[2]Utilizamos $r_{c,s} = \oslash$ para representar que o item s não foi avaliado por c.

o cálculo dos "vizinhos" do usuário x, geralmente computado através da busca pela intersecção dos conjunto S_x e S_y.

Em métodos baseado em correlação, o coeficiente de correlação de Pearson é bastante utilizado [39], [49], [56], e será explicado em maiores detalhes no capítulo 3. Nos métodos baseados em coseno [12], [55], os dois usuários x e y são tratados como dois vetores em um espaço m-dimensional, onde $m = |S_{xy}|$. Então a similaridade entre dois vetores pode ser medida pelo coseno do ângulo entre eles.

O entendimento desta similaridade, entretanto, ficou restrito por muito tempo à similaridade entre os usuários $sim(c, c')$, ao invés de também considerar a similaridade entre os itens $sim(s, s')$. A consideração dual da similaridade para resolução do problema de recomendação pode ser análoga a uma pesquisa indexada por itens ou por usuários no espaço $C \times S$ que define as avaliações $r_{c,s}$. O cálculo da medida da similaridade por itens foi primeiramente proposto em [55], no qual é sugerido que tanto correlações quanto a medida do coseno podem ser usadas na medida de similaridades de itens para a computação de recomendações. A idéia foi posteriormente estendida em [17] sugerindo a recomendação pelos N itens com maior similaridade.

2.4.2 Algoritmos Baseados em Modelo

Os algoritmos baseados em modelo [41], [12], [23], [40], [7] utilizam o conjunto de avaliações $C \times S$ para aprender um *modelo*, que é então utilizado para fazer predições. Em [12] uma abordagem probabilística é proposta para a filtragem colaborativa, onde as avaliações desconhecidas são calculadas como:

$$r_{c,s} = E(r_{c,s}) = \sum_{i=0}^{n} i \times \Pr(r_{c,s} = i | r_{c,s'},\ s' \in S_c) \tag{2.6}$$

assumindo que os valores das avaliações são inteiros variando de 0 a n e que a expressão probabilidade é a probabilidade de que o usuário c dará uma avaliação específica para o item s baseado nas avaliações dadas previamente a outros itens. Para estimar esta probabilidade, diferentes métodos podem ser utilizados como clusterização [12], redes bayesianas [15] ou métodos probabilísticos gaussianos [25] entre inúmeros outros. Estes métodos possuem algumas limitações como a dificuldade de representar um item (ou usuário) em diferentes clusters.

Além das abordagens probabilísticas, existem técnicas baseadas em aprendizado de máquinas, como redes neurais artificiais (e.g. RBM e SOMs) e aprendizado baseado em instâncias que podem ser utilizadas em sistemas de recomendação [61]. Nestes métodos, modelos são aprendidos com o treinamento dos algoritmos sobre o universo de dados disponível

e utilizados para gerar recomendações. Em [11] o autor propõem um método de filtragem colaborativa no qual várias técnicas de aprendizado de máquinas (diferentes redes neurais) são acopladas com técnicas de extração de features (singular value decomposition - SVD) gerando bons resultados. Nos trabalhos [12], [11], os respectivos métodos baseados em modelo são comparados com métodos baseados em memória e reportam que em determinadas aplicações os resultados dos algoritmos baseados em modelo superam em precisão as recomendações feitas pelos baseados em memória. É importante salientar no entanto, que estas comparações são puramente empíricas e nenhuma evidência teórica sustenta estas afirmações.

A grande diferença entre os sistemas colaborativos baseados em modelos e os sistemas colaborativos baseados em heurísticas é que as técnicas baseadas em modelos calculam a predição da avaliação baseadas não em regras empíricas mas, diferentemente disto, baseadas em um modelo estatístico ou matemático extraído dos dados de treinamento. Diversos estudos como [7], [43] vêm propondo a combinação de técnicas baseadas em memória e em modelo, demonstrando empiricamente que estas abordagens podem alcançar melhores recomendações que um algoritmo baseado em apenas uma das técnicas.

Uma importante característica dos sistemas colaborativos é que diferentemente dos sistemas baseados em conteúdos, os algoritmos desenvolvidos possuem aplicabilidade em qualquer domínio do conhecimento, já que se baseiam puramente nas avaliações dos itens não se preocupando em entendê-los. Esta característica é um dos principais motivos do fato da pesquisa relacionada aos sistemas colaborativos estar em um estágio mais desenvolvido que a pesquisa dos sistemas baseados em conteúdos.

Algumas desvantagens dos sistemas colaborativos são idênticas a sistemas baseados em conteúdo. O Problema do Novo-Usuário também ocorre nos sistemas baseados em modelo já que este deve aprender as preferências do usuário a partir da correlação de avaliações feitas no passado em relação às avaliações de outros usuários. Além disto, novos itens são constantemente adicionados ao sistema e como sistemas colaborativos calculam suas predições baseados nas preferências dos usuários, até este novo item não ser avaliado por um número substancial de usuários, o sistema não estará apto a recomendá-lo.

Em qualquer sistema de recomendação o número de avaliações já obtidas dentro do espaço total de possibilidades de avaliações é muito pequeno. Isto significa que sempre haverá muitos itens para serem recomendados, porém isto significa também que pode ser difícil conseguir uma massa crítica de dados mínima para se fazer uma boa recomendação. Conforme descrito na seção 2.4.1, para que sistemas colaborativos possuam uma boa confiabilidade é necessário vários *usuários-comuns* ou *itens-comuns* dentro do espaço $C \times S$. Assim, se este espaço é preenchido *esparsamente* o sistema poderá fazer recomendações fracas. Em um sistema de recomendações de filmes, por exemplo, quando um usuário possui um gosto que difere da maioria dos usuários, este pode estar fadado a receber sempre recomendações pobres. Estes filmes que poucas pessoas avaliam, mesmo que recebam altas notas, serão pouco

recomendados. Alguns métodos foram propostos para lidar com o problema da esparsidade. Em [11] e [40], um método de redução dimensional denominado Singular Value Decomposition é proposto para reduzir a esparsidade das matrizes $C \times S$ conseguindo alcançar bons resultados. Maiores detalhes sobre este método são apresentados ainda neste capítulo.

2.5 Métodos Híbridos

As limitações existentes nos métodos baseados em conteúdo e baseados em colaboração podem ser amenizadas através da combinação de ambas abordagens. Pelo fato de existir uma complementariedade entre estes métodos, soluções híbridas vêm obtendo bons resultados em diferentes estudos realizados [7], [21], [57].

Existem diferentes formas de combinar técnicas colaborativas e técnicas baseadas em conteúdo em um sistema híbrido de recomendação. Podemos dividir estas formas das seguintes maneiras:

1. implementando os métodos colaborativos e baseados em conteúdo em separado e combinando suas predições;
2. incorporando algumas características de sistemas baseados em conteúdo dentro de abordagens colaborativas;
3. incorporando algumas características de sistemas colaborativos dentro de abordagens baseadas em conteúdo, e
4. construindo um modelo unificado que incorpora ambas características.

Na solução 1, um sistema de recomendação híbrido pode implementar separadamente diferentes abordagens (colaborativas ou baseadas em conteúdo) e então combiná-las de diferentes maneiras: uma forma é fazer a regressão linear das avaliações encontradas, ou utilizar o método denominado *voting scheme* [42]. Outra forma para fazer a combinação entre as avaliações é escolher uma em um dado momento e a outra em outra situação. A escolha por uma ou outra avaliação dependerá de uma métrica pré-estabelecida para medir a "qualidade" da predição. Assim itens que possuam uma vizinhança com forte correlação com outros itens poderão ser preditos através de uma abordagem colaborativa, já outros itens com baixa correlação poderão ser recomendados para usuários que possuam um estereótipo bem conhecido dentro de uma representação já conhecida (e.g. ontologia) em um sistema de recomendação baseado em conteúdo.

2.6 Estudos Recentes em Filtragem Colaborativa

Em função da competição promovida pela Netflix – *Netflix Prize* [37] – a pesquisa em sistemas de recomendação baseados em filtragem colaborativa evoluiu consideravelmente nos últimos anos. Nas sessões a seguir é apresentado o resumo de alguns destes estudos.

2.6.1 K-Nearest Neighbor

O *k-nearest neighbor – kNN–* é talvez o algoritmo mais utilizado na composição de sistemas de recomendação colaborativos baseados em memória [3]. O sucesso deste método depende, entre outros fatores, da escolha dos pesos que cada vizinho contribuirá para a predição das avaliações desconhecidas. Em [9], os autores apresentam os resultados encontrados após o desenvolvimento de duas técnicas de otimização. A primeira trata-se de uma etapa de pré-processamento, na qual os chamados "*global effects*" são removidos do conjunto de dados permitindo que as avaliações sejam comparadas com maior facilidade e conseqüentemente melhorando a precisão do algoritmo.

A segunda contribuição dos autores está associada à normalização de dados. Normalização é essencial para os métodos kNN, já que a correlação entre avaliações pertencentes à usuários ou itens não-normalizados produzem resultados inferiores. O trabalho descreve dez efeitos facilmente observáveis no conjunto de dados da Netflix que causam considerável variabilidade e mascaram as relações fundamentais existentes entre as avaliações. Empregando técnicas como, fatoração, *double centering* e remoção de efeitos globais, os autores empiricamente comprovam melhorias na precisão de métodos já utilizados no passado sobre o conjunto de dados da Netflix.

2.6.2 Singular Value Decomposition - SVD

A Decomposição em Valores Singulares (*Singular Value Decomposition* ou *Latent Semantic Indexing* em inglês), ou redução/projeção dimensional é uma técnica algébrica utilizada há algum tempo como método efetivo no processamento de linguagem natural e também em algoritmos de compactação e redução de matrizes esparsas.

Este método é uma conseqüência direta de um teorema da álgebra linear que coloca:

Teorema 2.6.1 *: Uma matriz A de dimensão $M \times N$ pode ser definida como o produto de três outras matrizes*

$$A_{mxn} = U_{mxm} S_{mxn} V^T_{nxn} \tag{2.7}$$

[3] Esta abordagem é apresentada em detalhe no próximo capítulo.

onde U é uma matriz ortogonal que satisfaz $U^TU = I_{mxm}$, S é uma matriz diagonal com elementos positivos que representam os valores singulares da matrix A, e V^T é a matriz V transposta que satisfaz $V^TV = I_{nxn}$

A partir deste teorema pode ser deduzido que uma matriz qualquer A pode ser aproximada através da decomposição em outras três matrizes sendo considerado os valores singulares da matriz S como uma representação dos valores originais. Esta propriedade permite a redução de grandes matrizes esparsas em n valores singulares.

No problema de recomendações, estes valores singulares podem ser interpretados como os parâmetros não explícitos que definem as preferências de um determinado usuário, e ao mesmo tempo o quanto que um determinado filme é pontuado em cada um destes parâmetros. Verificando os parâmetros com altos valores tanto para um usuário quanto para um filme (através de uma multiplicação vetorial) é possível verificar a similaridade entre eles.

Além da pertinência para criar um modelo de recomendação, a técnica SVD permite a redução de matrizes esparsas em matrizes compactas mais fáceis de serem trabalhadas. Em [20] este método é utilizado primeiramente dentro da competição promovida pela Netflix conseguindo bons resultados. Posteriormente o método é aprimorado em [40], onde o autor propõe a utilização da técnica SVD combinada com métodos de pós-processamento e da utilização de métodos lineares. Estas abordagens de sistemas de recomendação baseado em modelos quando utilizadas em combinação com outras técnicas resultaram em predições mais precisas que as baseadas em memória (kNN).

2.6.3 Restricted Boltzman Machines - RBM

As Máquinas Restritas de Boltzman são redes neurais estocásticas caracterizadas pela sua capacidade de aprenderem representações internas e de resolverem problemas combinatórios complexos [64]. Em [52], o autor apresenta que a utilização deste método na composição de um sistema de recomendação resulta em bons resultados mesmo quando utilizado sobre uma grande base de dados. Além disto, o autor mostra que os modelos RBM chegam a valores mais precisos que as abordagens SVD puras. Em [38], diversos competidores reportaram a complementariedade do método em relação a abordagem kNN demonstrando empiricamente que a combinação linear de predições de ambas abordagens resultam em resultados mais precisos que um sistema baseado puramente em RBM.

2.6.4 Soluções Híbridas

A utilização das três técnicas descritas nesta seção resultam hoje na abordagem de um algoritmo híbrido que resultou nas melhores predições alcançadas dentro da competição *Netflix Prize*. Em [7] o autor apresenta as técnicas utilizadas na composição de um preditor único

para a competição, utilizando modelos kNN, SVD, RBM e justifica a complementariedade entre estas abordagens utilizadas.

2.7 Conclusão

Em resposta à dificuldade de avaliação, em um mundo com cada vez mais possibilidade de escolhas, sistemas de recomendação computacionais emergiram como forma de suporte, mediação e automação do processo de recuperação da informação. A evolução destes sistemas e o fato destes trabalharem com bases grandes de informações, permitiram que recomendações emergentes (não triviais) pudessem ser alcançadas, proporcionando ainda maior credibilidade que uma recomendação humana.

Os sistemas de recomendação representam hoje uma área independente de pesquisa, resultando em inúmeros trabalhos na academia e na indústria. Apesar de ser uma área nova, com um pouco mais de uma década de existência, a sua aplicabilidade e multidisciplinaridade atrai a atenção de pesquisadores de diferentes áreas e de diferentes setores econômicos.

Sistemas de recomendação são divididos em três grandes categorias: sistemas baseados em conteúdo, sistemas baseados em filtragem colaborativa e os sistemas híbridos. Os sistemas baseados em filtragem colaborativa ainda são subdivididos em sistemas colaborativos baseados em memória e os sistema colaborativos baseados em modelos. Os algoritmos baseados em memória são os mais amplamente utilizados, porém nos últimos anos sistemas baseados em modelos vêm ganhando força graças aos bons resultados alcançados.

A competição *Netflix Prize* promoveu a pesquisa em sistemas de recomendação baseados em filtragem colaborativa gerando inúmeros trabalhos na área. Otimizações de algoritmos clássicos foram realizadas e novas abordagens, como a utilização de SVD e RBM em algoritmos de filtragem colaborativa, foram sugeridas. Sistemas de recomendação híbridos que utilizam as principais abordagens em conjunto conseguem alcançar os melhores resultados.

Este capítulo apresentou um panorama da área de sistemas de recomendação e descreveu os atuais métodos que são usualmente classificados pela literatura dentro de três principais categorias: baseados em conteúdo, filtragem colaborativa e métodos híbridos. Uma descrição mais completa foi apresentada dentro da categoria filtragem colaborativa, objeto de estudo desta dissertação, cujo estado da arte com os métodos desenvolvidos nos últimos anos foram discorridos em maiores detalhes.

Capítulo 3

K Nearest Neighbor

Eu gosto de jogos de palavras. Mas elas sempre roubam e acabo mudo.

— Bruno Vivan Bernartt

A medida de similaridade em um sistema de recomendação baseado em filtragem colaborativa procura identificar itens ou usuários que possam auxiliar no cálculo de uma determinada predição. Neste capítulo é apresentado o método de aprendizagem de máquina baseado em instâncias denominado k-Nearest Neighbor – kNN. Este método, pelas suas características de aprendizado "não-parametrizadas", possibilita a identificação de comportamentos complexos em grandes volumes de dados. A abordagem kNN é um dos mais populares algoritmos utilizados nas implementações de sistemas de recomendações baseados em filtragem colaborativa. A escolha desta abordagem para o desenvolvimento de um sistema de recomendação é justificada no capítulo 5. Ao longo do texto a seguir assume-se que o leitor é familiarizado com a terminologia e com os conceitos clássicos de algoritmos de aprendizado de máquinas e, mais particularmente, algoritmos baseados em instâncias.

3.1 Contextualização

O algoritmo ***k-nearest neighbor*** é um dos mais simples métodos de aprendizagem de máquina existentes hoje. Dentro do campo de aprendizagem estatística enquadra-se na família dos algoritmos não-parametrizados denominados **instance-based learning** ou **memory-based learning**. Este nome é dado pelo fato do aprendizado ser feito através das próprias instâncias de entrada do algoritmo. Por não criar nenhum modelo de classificação

ou predição, mas apenas explorar a analogia do novo objeto a ser classificado em relação aos dados de treinamento, ele é denominado um *lazy algorithm* ou método preguiçoso. Este método foi descrito primeiramente nos anos 50, mas foi somente a partir da década de 60, quando computadores mais potentes começaram a surgir, que o método ganhou popularidade. Ele tem sido muito usado desde então em diversas aplicações no campo de *data-mining*, em métodos estatísticos de reconhecimento de padrões, em processamento de imagens entre outros.

3.2 O Modelo

A idéia central do modelo ***k-nearest neighbor*** ou **kNN** é baseada na distância mínima entre a instância desejada e as amostras de treinamento que determinam os vizinhos mais próximos desta entrada. Desta forma o algoritmo deve encontrar os k – parâmetro a ser definido no algoritmo – vizinhos mais próximos da entrada x que posteriormente poderão ser utilizados para uma predição, classificação ou *smoothing*. [1]

Exemplificando dentro do contexto da classificação: Considere um espaço bidimensional definido pelos atributos a_1 e a_2 onde a classe c_1 e a classe c_2 estão representadas por 6 e 5 instâncias, respectivamente. Desejamos determinar a qual das duas classes pertence uma instância x, de classe desconhecida. Esta situação é ilustrada na Figura 3.1, onde os quadrados representam a classe c_1, os triângulos a classe c_2 e a bola a instância de classe desconhecida x. Basicamente, as técnicas kNN assumem como classe de x, a classe da instância que está mais próxima de x. No exemplo da Figura 3.1, x assume a classe c_2, considerando neste caso que $k = 3$ representado na figura através do primeiro círculo que contém um número maior de triângulos (classe c_2) do que quadrados (classe c_1). Já se considerarmos o segundo círculo, estamos considerando $k = 5$ e neste caso a classe de x será predita pela classe c_1 já que existe um número maior de quadrados do que triângulos na proximidade de x. E assim o algoritmo faz uma determinada classificação/predição com base na definição do parâmetro k.

[1] Smooth em estatística e em processamento de imagem, significa criar uma função que consiga identificar padrões em um conjunto de dados, deixando de lado ruídos e dados indesejáveis

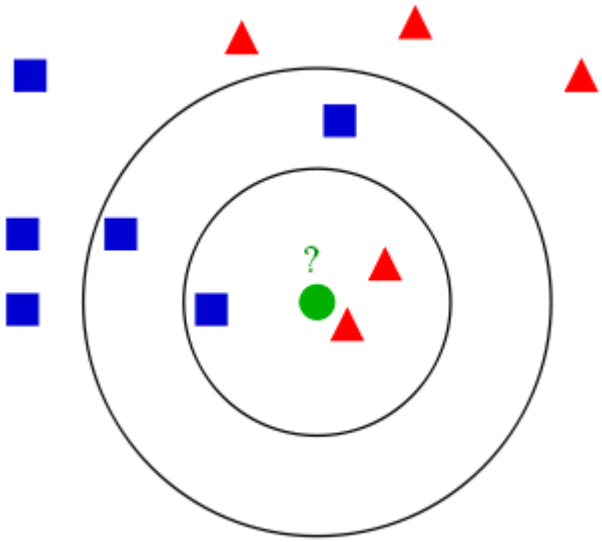

Figura 3.1: Exemplo de Classificação kNN

A seguir apresentamos uma definição formal do método NN, baseada na definição proposta por [22], e a sua generalização para o método kNN. Considere:

1. Um espaço n-dimensional de atributos;
2. M classes numeradas $1, 2, 3, ..., M$;
3. Seja $T_{NN} = \{(x^1, \theta_1), (x^2, \theta_2), ..., (x^p, \theta_p)\}$ o conjunto de p pares de treinamento, cada par expresso por (x^i, θ_i), para $1 \leq i \leq p$, onde:
   (a) x^i é uma instância de treinamento expressa pelo vetor multidimensional com um atributo em cada dimensão $x^i = (x_1^i, x_2^i, ..., x_n^i)$
   (b) $\theta_i \in 1, 2, ..., M$ denota a classe correta da instância x^i

   Então:

Definição 3.2.1 *: Dada uma instância desconhecida x, a regra de decisão do algoritmo NN decide que x está na classe θ_j se, e somente se,*

$$D(x, x^j) \leq D(x, x^i)$$

$$1 \leq i \leq p$$

onde D é alguma métrica n-dimensional de distância.

A regra $D(x, x_j) \leq D(x, x_i)$ apresentada na definição 3.2.1 é mais apropriadamente chamada de regra 1-NN, uma vez que usa apenas um vizinho mais próximo, i.e., calcula

a distância da nova instância a cada ***um*** dos exemplos do conjunto de treinamento. Assim sendo, ao considerarmos k vizinhos, teremos a variante conhecida como k-NN, a qual identifica as k instâncias mais próximas $x^1, x^2, ..., x^k$ e decide pela escolha da classe que comparece com maior freqüência no conjunto $\theta_{i1}, \theta_{i2}, ..., \theta_{ik}$. Uma forma de proceder é fazer uma validação cruzada (*cross-validation*) utilizando diferentes valores de k, e escolher aquele que obtiver o melhor resultado. Formalmente definimos kNN:

Definição 3.2.2 *: Dada uma instância desconhecida x, a regra de decisão do algoritmo kNN decide que x está na classe θ_j se, e somente se, a classe θ_j for a mais freqüente do conjunto $\theta_{i1}, \theta_{i2}, ..., \theta_{ik}$ entre os k vizinhos de x onde a vizinhança é definida pelas k menores distâncias calculadas por*

$$D(x, x^j) \leq D(x, x^i)$$

$$1 \leq i \leq p$$

onde D é alguma métrica n-dimensional de distância.

Nas seções a seguir, apresentamos uma breve consideração sobre as vantagens e desvantagens gerais do método kNN de forma a subsidiar futuras considerações sobre a utilização do mesmo no desenvolvimento do sistema de recomendação.

3.2.1 Vantagens do Método

O aprendizado baseado em instâncias favorece o aprendizado incremental a partir da experiência, uma vez que é sempre mais fácil aprender através da retenção da memória de uma experiência concreta do que generalizar a partir dela. Esta é uma importante vantagem para o desenvolvimento de aplicações, pois permite sistemas iniciarem sua operação com um pequeno conjunto de padrões e adicionarem cobertura através do armazenamento de novos padrões se for demonstrado que são necessários. Como não realiza nenhum tipo de generalização do conhecimento, não incorre no risco de esquecer detalhes à medida que o conhecimento cresce.

Algoritmos do tipo kNN possuem um custo de atualização relativamente reduzido e o aprendizado baseado em instâncias permite que o sistema resultante seja bastante robusto. Além disso, sistemas com aprendizado kNN são verificáveis (o que não se pode dizer da maioria dos modelos de redes neurais, por exemplo), pois permitem que se consulte os padrões concretos responsáveis pelo comportamento classificatório adquirido pelo sistema após o aprendizado.

De forma geral, a fácil implementação do algoritmo kNN e o rápido aprendizado são as suas principais vantagens, o que é evidenciado pela sua popularidade e grande utilização nas

mais diversas aplicações. Em contraste com métodos de aprendizagem paramétricos, o kNN permite que a complexidade da hipótese levantada cresça com o crescimento dos dados.

Como é demonstrado na seqüência deste capítulo, o método kNN quando combinado com métodos de eliminação de ruídos e de *weighting* possui uma performance boa em comparação com outros métodos, tanto no contexto de classificação quanto no contexto de predição, podendo manipular dados incompletos ou com ruídos praticamente em tempo real.

Do ponto de vista matemático, todos os algoritmos da família kNN incrementalmente aprendem aproximações parcialmente lineares de conceitos em um espaço n-dimensional através da medida de similaridade, o que é uma forma bastante natural de se aproximar uma grande gama de distribuições de variáveis.

3.2.2 Desvantagens do Método

Apesar de sua relativa simplicidade de implementação, o algoritmo possui algumas desvantagens quando aplicado em grandes bases de dados. Isto se justifica pelo fato de que calcular a distância de uma determinada entrada com todos os possíveis atributos a serem associados, pode levar a um tempo de computação muito elevado. Vários métodos de pré-processamento foram propostos para tornar esta etapa mais eficiente, porém a maioria destes métodos não consegue ser escalável quando o problema possui mais do que uma dimensão.

Espaços multi-dimensionais colocam um desafio a mais ao método kNN, já que quando tratamos com espaços vetoriais de muitas dimensões, os vizinhos acabam se tornando muito distantes não conseguindo explicitar o propósito básico do algoritmo, que é identificar similaridades dos poucos vizinhos próximos à entrada em questão. Para ilustrar, pegaremos o exemplo retirado de [59], onde encontramos um conjunto de dados de tamanho N de um hipercubo [2] de dimensão d, assumindo vizinhos hipercúbicos de lado b e volume b^d . Para conter k pontos, a média da vizinhança deverá ocupar a fração k/N do volume total, que aqui definiremos como 1, assim $b^d = k/N$, ou $b = (k/N)^{1/d}$. Determinando que a dimensão d seja 100, ou seja, um espaço de muitas dimensões, e que a vizinhança k seja de 10 vizinhos, e ainda que o tamanho total do conjunto de dados seja $N = 1.000.000$, teremos um $b \approx 0.89$. Este exemplo mostra que para conseguir definir a vizinhança, seria necessário percorrer quase todo o espaço vetorial de entrada. Isto sugere que o método **k-nearest neighbor** não é confiável para dados com muitas dimensões. Com menos dimensões não existe o problema, o que pode ser percebido considerando no exemplo acima $d = 2$ resultando em $b = 0,003$.

A performance deste algoritmo é prejudicada quando há a existência de dados com ruído no conjunto de testes, pois a determinação da classe de uma determinada instância de entrada é feita através de seus vizinhos mais próximos sem nenhum cálculo adicional

[2] O mesmo raciocínio poderia ser feito para hiperesferas; mas a fórmula de volume de esferas seria mais complicada para demonstração.

que pudesse eliminar os ruídos. Também encontramos problemas quando os atributos dos vizinhos afetam em magnitudes diferentes a saída procurada. O caso extremo é quando alguns atributos são completamente irrelevantes e acabam contribuindo igualitariamente na fórmula da distância.

3.3 Etapas de Implementação

A seguir são feitas algumas considerações mais detalhadas para cada etapa de desenvolvimento de um sistema genérico que utiliza o algoritmo kNN e, conseqüentemente, é discorrido com maiores detalhes sobre cada parte do algoritmo, demonstrando como algumas implementações podem contornar as desvantagens que o algoritmo básico kNN possui.

3.3.1 Pré-Processamento

A etapa de pré-processamento possui dois objetivos principais: (1) diminuir o número de exemplares de forma a reduzir o tempo de computação e melhorar a performance do sistema, e (2) eliminar exemplares ruidosos que diminuem a performance do algoritmo.

A existência de um conjunto de dados muito grande pode ocasionar um tempo computacional impraticável. Este foi um dos motivos que fizeram com que o algoritmo kNN fosse conhecido como um algoritmo lento antes dos anos 90. Nesta época estruturas sofisticadas de dados como *kD-trees* começaram a ser utilizadas de forma a reduzir consideravelmente o tempo computacional. A estrutura de *kD-tree* é uma árvore binária que divide um espaço vetorial de entrada com um hiperplano e divide cada partição recursivamente, armazenando um conjunto de dados k-dimensional, onde k é o número de atributos. Maiores detalhes sobre esta estrutura de dados podem ser encontrados em [26]. Na prática, as árvores *kD-trees* se tornam ineficientes com o aumento da dimensão do espaço e sua validade também só é justificada quando o número de atributos é pequeno ≈ 10. Estruturas de dados como as *ball-trees* e outras estruturas de dados mais recentes trabalham com sucesso centenas de dimensões.

Em vez de armazenar todas as instâncias de treinamento, é possível comprimí-las em regiões. Uma técnica muito simples é gravar apenas um intervalo de valores observados nos dados de treinamento para cada atributo e categoria. Dada uma instância de teste, é determinado um intervalo onde caem os valores dos atributos, e escolhe-se a categoria com o maior número de intervalos corretos para uma determinada instância. Uma técnica mais elaborada é construir intervalos para cada atributo e usar o conjunto de treinamento para contar o número de vezes que cada categoria ocorre para cada intervalo em cada atributo. Atributos numéricos podem ser discretizados em intervalos, e "intervalos" consistindo de um único ponto, podem ser usados para denominá-lo. Assim sendo, dada uma instância de teste,

você pode determinar em qual intervalo ela reside e classificá-la através de votação, método este chamado de *voting feature intervals*. Apesar de serem aproximações, estes métodos são extremamente rápidos e podem ser úteis para uma análise inicial do conjunto de dados em questão.

Normalizações podem ser efetuadas para conseguir retirar efeitos indesejados sobre o conjunto de dados de treinamento. Em [9] são apresentadas algumas técnicas de normalização utilizadas sobre o conjunto de treinamento fornecido pela empresa Netflix para a competição Netflix Prize. Através destas técnicas o autor consegue alcançar consideráveis melhorias com o seu algoritmo kNN implementado.

É possível realizar pré-cálculos com a base de treinamento e salvar os resultados em arquivos intermediários para depois serem utilizados pelo algoritmo no cálculo de recomendações. Esta estratégia permite que alguns cálculos sejam feitos apenas uma vez ao invés de serem repetidos várias vezes. Assim o algoritmo é otimizado proporcionado tempos computacionais mais baixos para as recomendações finais.

Exemplares com ruído, inevitavelmente, reduzem a performance de qualquer esquema baseado em kNN que não os suprimem, já que estes ruídos provocam o efeito de repetidamente desclassificar novas instâncias, ou aumentarem os erros em predições. Existem duas formas de lidar com este problema. Uma destas formas é em vez de localizar um único vizinho, localizar os k vizinhos mais próximos, sendo k uma constante pré-determinada e considerar a classe que ocorrer majoritariamente nos vizinhos, como a classe da instância desconhecida. O único problema aqui é determinar um valor adequado para k. Quanto mais ruído, maior será o valor ótimo de k.

3.3.2 Determinação dos Vizinhos

Uma importante parte do algoritmo kNN é a escolha da função de distância mínima a ser utilizada para a definição dos ***vizinhos***. Este passo está totalmente associado à natureza do problema a ser resolvido.

Para identificar os vizinhos mais próximos de uma instância desconhecida, é necessária uma métrica n-dimensional de distância, aqui denominada $D(x_1, x_2)$ (ver definição 3.2.1). Uma métrica bastante utilizada é calcular a **distância Euclidiana**, definida como:

Definição 3.3.1 *: A distância entre os pontos* $P = (p_1, p_2, ..., p_n)$ *e* $Q = (q_1, q_2, ..., q_n)$ *no n-espaço Euclidiano é definido como:*

$$\sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + ... + (p_n - q_n)^2} = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2} \tag{3.1}$$

A utilização da distância Euclidiana é inapropriada quando cada dimensão do espaço mede algo diferente, como por exemplo o peso e a altura. Isto se deve ao fato de que ao mudar a escala de uma dimensão é mudado também o conjunto dos vizinhos mais próximos. Uma solução para isto é padronizar a escala para cada dimensão. Para fazer isto, é necessário medir o desvio padrão de cada atributo sobre todo o conjunto de dados e expressar os valores dos atributos como múltiplos do desvio padrão de cada atributo. Este é um caso especial da distância de Mahalanobis, que leva em consideração a covariância dos atributos.

Em estatística, a **distância de Mahalanobis** é uma medida baseada na correlação entre duas variáveis que possuem diferentes modelos que podem ser identificados e analisados. É uma medida muito utilizada para determinação de similaridades, e por isto, pode ser utilizada como medida de distância no algoritmo kNN. Formalmente pode ser definida como:

Definição 3.3.2 *: A distância de Mahalanobis entre os vetores $\vec{x}$ e $\vec{y}$, ambos com mesma distribuição da matrix covariante Σ é definida como:*

$$D_M(\vec{x}, \vec{y}) = \sqrt{(\vec{x} - \vec{y})^T \Sigma^{-1} (\vec{x} - \vec{y})} \tag{3.2}$$

Se a matrix covariante é a matrix identidade, a distância de Mahalanobis se reduz a distância Euclidiana. Se a matrix covariante for diagonal, então a distância resultante medida é chamada de *distância Euclidiana Normalizada*:

$$d(x, y) = \sqrt{\sum_{i=1}^{p} \frac{(x_i - y_i)^2}{\sigma_i^2}} \tag{3.3}$$

onde σ_i é o desvio padrão de x_i sobre o conjunto de amostra.

Na distância de Mahalanobis, percebemos a utilização da *correlação*, na procura de similaridades entre duas variáveis. Em estatística, **correlação** indica a força e a direção do relacionamento linear entre duas variáveis aleatórias. No uso estatístico geral, *correlação* ou *co-relação* se refere à medida da relação entre duas variáveis, embora esta correlação não indique causalidade [3]. Neste sentido geral, existem vários coeficientes medindo o grau de correlação, adaptados à natureza dos dados. O coeficiente mais utilizado em algoritmos kNN é o **coeficiente de correlação de Pearson**, também chamado de coeficiente de correlação produto-momento ou simplesmente de "r de Pearson". Formalmente:

Definição 3.3.3 *: O coeficiente de correlação de Pearson entre o conjunto de variáveis $x_1, x_2, ..., x_n$ e $y_1, y_2, ..., y_n$ pode ser obtido dividindo a covariância entre as variáveis x, y, $Cov(x, y)$ pelo produto de seus desvios padrão:*

[3] Causalidade aqui referida no seu sentido filosófico ou físico.

$$r = \frac{\text{Cov}(x, y)}{\sigma_x \sigma_y} \tag{3.4}$$

onde,

$$\text{Cov}(x, y) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})$$

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

$$\sigma_y = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \overline{y})^2}$$

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

e

$$\overline{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

Desta forma, uma equação simplificada pode ser encontrada:

$$r = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \overline{x})^2} \sqrt{\sum_{i=1}^{n} (y_i - \overline{y})^2}} \tag{3.5}$$

Através da definição acima é possível perceber que os valores de r assumem apenas valores entre -1 e 1. Desta forma, os valores podem ser interpretados da seguinte forma:

- $r = 1$ significa uma correlação perfeitamente positiva entre as duas variáveis;
- $r = -1$ significa uma correlação perfeitamente negativa entre as duas variáveis. Ou seja, se uma aumenta a outra sempre diminui;
- $r = 0$ significa que as duas variáveis não possuem uma dependência linear. No entanto, pode existir uma dependência não-linear, e neste caso o resultado $r = 0$ deve ser investigado por outros meios.

Se generalizarmos a função da distância euclidiana chegaremos a função de similaridade de **Minkowski** que é definida por:

Definição 3.3.4 *: A distância de Minkowski entre os pontos* $P = (p_1, p_2, ..., p_n)$ *e* $Q = (q_1, q_2, ..., q_n)$ *no n-espaço euclidiano é definida como:*

$$d_q(p, q) = \sqrt[q]{\sum_{i=1}^{n-1} w_i |p_i - q_i|^q} \tag{3.6}$$

Se considerarmos na equação de Minkowski $q = 2$ e cada peso w_i igual a 1 teremos a distância Euclidiana definida em 3.3.1. Alternativamente, se definirmos $q = 1$ e $w_i = 1$ teremos a **distância de Manhattan**, que apesar de ser utilizada em alguns algoritmos kNN, possui algumas limitações.

Outros modelos ainda podem ser utilizados dependendo da natureza dos dados, e.g: quando o problema é identificar a distância de atributos discretizados a **distância de Hamming**, que define $D(x_1, x_2)$ como sendo o número de features nas quais x_1 difere de x_2, é uma boa escolha a ser feita.

3.3.3 Treinamento - atribuindo pesos

Na maioria dos domínios de estudos, alguns atributos presentes no conjunto de dados, ou mais especificamente presentes no conjunto de vizinhos mais próximos *(k-nearest neighbors)*, são irrelevantes. Dentro do conjunto dos atributos relevantes podemos encontrar ainda alguns que são mais importantes que os outros. Um grande avanço no aprendizado baseado em instâncias é aprender a relevância de cada atributo incrementalmente através da atualização dinâmica dos pesos dos atributos.

De forma a permitir o aprendizado da relevância de cada atributo, a distância métrica deve incorporar os pesos $w_1, w_2, ..., w_n$ em cada dimensão. Utilizando como exemplo a distancia Euclidiana, temos a seguinte equação:

$$\sqrt{w_1^2(x_1 - y_1)^2 + w_2^2(x_2 - y_2)^2 + ... + w_n^2(x_n - y_n)^2}.$$

Todos os pesos dos atributos são atualizados após cada classificação de uma instância de treinamento, e o exemplar mais similar é utilizado como a base para a definição desta atualização. Chamaremos a instância de treinamento de x e o exemplar mais similar de y. Para cada atributo i, a diferença $|x_i - y_i|$ é a medida da contribuição daquele atributo na decisão. Se a diferença é pequena, então o atributo contribui muito; quando a diferença é grande ele contribui menos. A idéia básica é fazer a atualização do peso w_i com base no tamanho desta diferença, independentemente se a classificação é ou não correta. Se a classificação é correta, o peso associado aumenta, e se for incorreta o peso diminui, sendo que o quanto o peso aumenta ou diminui é definido pelo tamanho da diferença de forma inversamente proporcional: se a diferença é pequena a atualização (positiva ou negativa) é grande e vice-versa. A mudança dos pesos geralmente é seguida de uma etapa de renormalização. Uma estratégia simples, que pode ser igualmente efetiva à descrita aqui, é a de não atualizar os pesos se a decisão é correta e se a decisão for incorreta aumentar os pesos para aqueles atributos que diferirem mais, acentuando a diferença [1].

Outra forma de atualizar os pesos é determinando uma fórmula para ele que utilize variáveis que identifiquem o quão relevante ele é para a predição. Diversos estudos como [9],[10], sugerem estratégias para a determinação destas variáveis e da atualização dos pesos.

Um bom teste para saber se o método de atribuições de pesos está funcionando conforme desejado é adicionar atributos irrelevantes aos exemplares do conjunto de dados. Idealmente, a introdução de atributos irrelevantes não deverá afetar nem a qualidade das predições nem o número de exemplares armazenados.

3.3.4 Função de Classificação ou Predição

Após passar a etapa de pré-processamento, de implementar a função de distância métrica para encontrar os vizinhos mais próximos e, por fim, fazer o treinamento do algoritmo atribuindo pesos para os diferentes atributos existentes, é finalmente possível fazer a classificação ou a predição para uma instância desconhecida. Pelo fato do algoritmo ser utilizado tanto para a classificação quanto para a predição de valores numéricos, esta etapa difere conforme o problema a ser resolvido.

Quando o problema em questão é a classificação da instância para algum valor simbólico ou uma classe [4], geralmente a predição se dá através da **Maioria Simples** dos valores simbólicos ou classes dos vizinhos mais próximos. Outra forma também utilizada é a utilização de **votação ponderada**, na qual cada classe ou valor simbólico possui um peso de votação que é contabilizado na determinação final da classe a ser determinada.

Quando estamos tratando com predições numéricas, a escolha do valor a ser predito geralmente é feita através do cálculo da média aritmética simples dos valores dos vizinhos mais próximos. Entretanto, dependendo da dimensionalidade e complexidade do problema, outras formas de cálculo de uma tendência central dos valores dos vizinhos podem ser utilizadas. As formas mais comuns para cálculo de uma tendência central são a **Média**, a **Mediana** e a **Moda**.

A *Média* é uma medida que calcula o valor aritmético médio do conjunto da amostra, somando todos os valores e dividindo esta soma pelo número de amostras existentes.

A *Mediana* é uma medida de localização do centro da distribuição dos dados, definida do seguinte modo: Ordenados os elementos da amostra, a mediana é o valor (pertencente ou não à amostra) que a divide ao meio, isto é, 50 por cento dos elementos da amostra são menores ou iguais à mediana e os outros 50 por cento são maiores ou iguais à mediana. Para a sua determinação utiliza-se a seguinte regra, depois de ordenada a amostra de n elementos,

[4] Os termos **Classe** e **Categoria** são utilizados coloquialmente de forma quase sinônima. Em RBC e outras áreas que envolvem Representação de Conhecimento estas palavras representam conceitos distintos. Sendo que o termo "classe" possui limites bem definidos e o termo "categoria" determina um espaço próximo a instância em questão.

se n é ímpar, a mediana é o elemento médio. Se n é par, a mediana é a média dos dois elementos médios.

Define-se *Moda* como sendo o valor que surge com mais freqüência se os dados são discretos, ou, o intervalo de classe com maior freqüência se os dados são contínuos. Assim, da representação gráfica dos dados, obtém-se imediatamente o valor que representa a moda ou a classe modal. Esta medida é especialmente útil para reduzir a informação de um conjunto de dados qualitativos apresentados sob a forma de nomes ou categorias, para os quais não se pode calcular a média e, por vezes, nem a mediana.

Quando se trabalha com uma distribuição normal, os valores de Média, Mediana e Moda podem coincidir, entretanto, quando se trabalha com distribuições assimétricas, estes valores não coincidem. Muito cuidado, portanto, deve ser tomado na escolha da forma de cálculo para a predição de algum valor numérico durante a utilização de uma ou outra forma de buscar uma tendência central. Deve-se lembrar que o cálculo da média ignora a distribuição dos valores, e portanto a presença de algum vizinho que contenha ruídos pode complicar a predição. O uso da Mediana, apesar de lidar com este problema, pode eliminar um importante vizinho que deveria influenciar a predição.

Para predições numéricas, outras técnicas podem ser combinadas com o kNN de forma a gerar preditores mais precisos. Uma delas é o Slope One, introduzido em [30] que sugere: Dados dois conjuntos de avaliações v_i, e w_i com $i = 1, ..., n$, procura-se pelo melhor preditor na forma $f(x) = x + b$ para predizer wa partir de v pela minimização $\sum_i (v_i + b + w_i)^2$. Derivando em relação a b e igualando a derivada a zero, obtém-se $b = \sum_i w_i - v_i \times n^{-1}$. Em outras palavras, a constante b deve ser escolhida para ser a diferença média dos dois conjuntos. Assim, dado um conjunto de treinamento x e dois itens j e i com avaliações u_j e u_i respectivamente em alguma avaliação do usuário u (cuja notação será feita por $u \in S_{j,i}(x)$), considera o desvio médio do item i em relação ao item j como:

$$dev_{j,i} = \sum_{u \in S_{j,i}(x)} \frac{u_j - u_i}{card(S_{j,i}(x))} \tag{3.7}$$

Importante notar na equação 3.7, que qualquer avaliação u que não contenha ambos u_j e u_i não é incluído no somatório. A matriz simétrica definida como $dev_{j,i}$ pode ser calculada uma vez e facilmente atualizada quando novos dados são inseridos. Desta forma podemos definir a regra Slope One da seguinte maneira:

Definição 3.3.5 *: Sendo $u_i + dev_{j,i}$ uma boa predição para u_j, dado u_i, um preditor razoável para um conjunto de possíveis avaliações dos vizinhos similares a u seria:*

$$P(u)_j = \frac{1}{card(R_j)} \sum_{i \in R_j} (dev_{j,i} + u_i) \tag{3.8}$$

onde, $R_j = \{i | i \in S(u), i \neq j, card(S_{j,i}(x)) > 0\}$ *é o conjunto de todos os itens relevantes (vizinhos selecionados).*

Há uma aproximação que pode simplificar o cálculo da predição baseada no Slope One. Para o conjunto de dados denso o suficiente, no qual quase todos os pares de itens possuem classificação, isto é, onde $card(S_{j,i}(x)) > 0$ para quase todo j, i, a maioria das vezes $R_j = S(u)$ para $j \neq S(u)$ e$R_j = S(u) - \{j\}$ quando $j \in S(u)$. Assim,

$$\bar{u} = \sum_{i \in S(u)} \frac{u_i}{card(S(u))} \simeq \sum_{i \in R_j} \frac{u_i}{card(R_j)}$$

Assim simplifica-se a fórmula 3.8 para:

$$P'(u)_j = \bar{u} + \frac{1}{card(R_j)} \sum_{i \in R_j} dev_{j,i} \tag{3.9}$$

3.4 Exemplo

Para ilustrar todas as etapas descritas na seção 3.3, é apresentado a seguir um problema de predição numérica [5] a ser resolvido pelo algoritmo kNN. As escolhas tanto da problemática quanto dos detalhes inerentes ao algoritmo são de baixa complexidade para permitir uma fácil compreensão do algoritmo.

Considere cinco pares de dados (x, y) como mostrados na tabela 3.1. Os dados são quantitativos por natureza. O problema é estimar o valor de y baseado nos k vizinhos mais próximos quando $x = 6,5$.

Tabela 3.1: Exemplo 1 - Conjunto de pares (x, y)

X	Y
1	23
1,2	17
3,2	12
4	27
5,1	8
6,5	**?**

Definir-se-á o parâmetro k – número de vizinhos mais próximos – como sendo 2. Isto significa que para a predição do valor de y é considerado dois (2) vizinhos mais próximos. Para encontrar estes vizinhos é necessário calcular a distância mínima entre a instância de

[5] Escolhemos o problema de predição ao invés de classificação, pelo fato deste trabalho lidar com este tipo de problema, fazendo com que o leitor se familiarize com a sua resolução.

entrada e os exemplos de teste. Será utilizado como métrica a distância definida em 3.3.1, a chamada distância Euclidiana. Ela foi escolhida por se tratar de um problema simples, de uma dimensão, cujas escalas não variam. Desta forma as distâncias calculadas para a entrada $x = 6,5$ são dadas pela tabela 3.2.

Tabela 3.2: Exemplo 1 - Distâncias da entrada

Cálculo	**Distância**
$\|6,5 - 1\|$	5,5
$\|6,5 - 1,2\|$	5,3
$\|6,5 - 3,2\|$	3,3
$\|6,5 - 4\|$	2,5
$\|6,5 - 5,1\|$	1,4

Analisando as distâncias encontradas, percebemos que as duas entradas $x = 4$ e $x = 5,1$ são as que estão mais próximas da instância em questão, logo, são os seus vizinhos mais próximos (lembre-se que definimos o $k = 2$ e, portanto, queremos selecionar apenas dois vizinhos). Com base nos vizinhos selecionaremos os valores de y dos mesmos: $y = \{27, 8\}$ e faremos então a predição através da média aritmética.

$$\frac{27 + 8}{2} = 17,5$$

Desta forma obtivemos como valor de predição para a variável y o valor de 17,50. Poderíamos mudar o valor de k e obteríamos diferentes valores. Como não temos um conjunto de teste e de validação que permitam realizar o método *cross-validation*, não temos como prever qual valor de k seria ótimo para este problema.

3.5 Conclusão

O algoritmo k-nearest neighbor é um dos mais utilizados métodos de aprendizagem de máquina existentes hoje, com grande aplicação no campo de data mining, reconhecimento de padrões, processamento de imagens, entre outros. Suas principais vantagens são a fácil implementação, a aplicabilidade em problemas complexos e um aprendizado rápido, incremental e robusto. Suas principais desvantagens: dificuldades de aplicação com grandes bases de dados, influência de atributos redundantes e irrelevantes e a ineficiência com problemas com muitas dimensões do espaço vetorial. Sua aplicação pode ser dividida principalmente em problemas de predição e de classificação.

Ao longo deste capítulo foi apresentado o modelo matemático do algoritmo kNN com as suas vantagens e desvantagens. Foram demonstradas diferentes abordagens, bem como todas as etapas de implementação do algoritmo. Por fim, a teoria foi apresentada através de

um exemplo de predição, demonstrando de forma simplificada como o algoritmo poderia ser utilizado em um sistema de recomendação, objeto de estudo deste trabalho.

Capítulo 4

Netflix Prize

O ânimo tranqüilo de um justo pode descobrir mais coisas que todos os sábios

— Sófocles

Conforme apresentado no capítulo 2, quando as possibilidades para se fazer uma determinada escolha pessoal são muito numerosas, então um sistema que "conhece" o gosto de uma determinada pessoa pode auxiliar. Estes sistemas são chamados de sistemas de recomendação. Uma situação que cabe nesta categoria é o problema da escolha de um determinado filme. Neste capítulo este problema é caracterizado de forma a ser resolvido através de um sistema de predição. A motivação para a escolha deste tema está na competição promovida pelo Netflix, que disponibilizou uma extensa base de dados com mais de 100 milhões de avaliações de filmes feitas por mais de 480 mil usuários distintos. Com estes dados e uma sistemática de validação, é possível verificar propriedades emergentes que diferentes algoritmos combinados podem encontrar, além de ser possível verificar a evolução da eficiência da recomendação encontrada através de uma métrica pré-estabelecida.

4.1 O Problema

Segundo a Associação de Cinema dos Estados Unidos - MPAA[1] - em seu relatório anual de 2006 sobre os dados da indústria cinematográfica, foram lançados 599 filmes no mercado estadunidense, e a freqüência foi recorde ao cinema, com cerca de 1,5 bilhão de

[1] A MPAA é uma organização sem fins lucrativos formada por seis grandes estúdios com o objetivo de trabalhar em nome da indústria cinematográfica. Em seu site [http://www.mpaa.org/], a MPAA se apresenta como "Porta-Voz e Defensora das Indústrias de Cinema, Vídeo e Televisão dos EUA".

ingressos vendidos. A indústria cinematográfica dos EUA é a maior fornecedora de filmes para o mercado mundial, dando uma margem global aos dados apresentados. Além do grande número de filmes lançados a cada ano, muitos são os meios que um usuário/cliente pode utilizar para adquirir um filme. As locadoras, os cinemas e as lojas de filmes (em geral, DVDs) apresentam-se em grande número nas mais diversas partes do Brasil e do mundo de forma física e online, sendo que para atingir todos estes meios de vendas, é necessária uma cara e complexa logística de canais de distribuição.

Se pelo lado da indústria cinematográfica atingir um público-alvo nem sempre é barato, do lado do usuário escolher um bom filme está cada vez mais difícil. Quando vai ao cinema, locadora ou loja de filmes, o consumidor depara-se com um universo cada vez mais vasto de possibilidades. A leitura das sinopses, a avaliação do elenco, os prêmios recebidos e as críticas são algumas das variáveis consideradas na hora da escolha. É também muito comum o consumidor guiar-se pela indicação de alguém.

Gosto é um critério, antes de mais nada, pessoal. As pessoas possuem gostos e preferências diferentes. É devido a esse caráter individual dos gostos, já descrito no capítulo 2, que a recomendação de filmes pode não ser bem sucedida. Um filme considerado bom pelo balconista da locadora não necessariamente será agradável ao cliente. Uma boa alternativa ao balconista é obter indicação de amigos. Amigos normalmente conhecem os gostos uns dos outros, o que os coloca numa posição melhor que a do balconista no aspecto da indicação. No entanto, o que ocorre é que para uma boa indicação é necessário, além do conhecimento dos gostos, entender que aspectos a outra pessoa leva em consideração para definir se um filme é bom ou ruim. Tudo fica mais fácil quando este amigo possui "aparentemente" o mesmo gosto que o outro, o que garantirá a confiabilidade de sua recomendação. Porém, novos filmes podem surgir, gostos podem mudar ao longo do tempo, e nem sempre este amigo terá visto todos os filmes que poderiam ser sugeridos. Outro problema é que seria necessário esperar sempre este amigo assistir um filme para que ele pudesse ser recomendado, e este amigo poderia estar ocupado por um longo período de tempo, como por exemplo desenvolvendo o seu mestrado, o que impossibilitaria novas recomendações.

Uma forma de conseguir encontrar diferentes pessoas com gostos similares é através de comunidades que se formam fisicamente com encontros periódicos, ou através de comunidades virtuais na internet. Estas comunidades acabam se organizando através de gostos específicos, possibilitando a troca de informações e recomendações sobre filmes.

A participação em uma comunidade requer dedicação e tempo, o que a maioria das pessoas não dispõem. Na maioria das vezes, o que se quer é simplesmente uma indicação de um filme para poder desfrutar um momento de descanso. Assim, se houvesse uma forma automática de conseguir recomendações atualizadas [2] e certeiras de filmes ainda não vistos,

[2] Atualizadas aqui referente ao que foi ou não visto por uma determinada pessoa, e não apenas de filmes recentes.

conseguiríamos ocupar todo o tempo livre apenas com filmes que nos agradasse. Um sistema de recomendação de filmes visa exatamente prover este tipo de serviço, resolvendo o problema acima exposto.

O correto funcionamento de um sistema de recomendação não dependerá apenas de um bom algoritmo implementado, mas também da utilização conjunta do maior número de usuários, que mesmo agindo muitas vezes individualmente no sistema, inserindo apenas as suas preferências, acaba contribuindo essencialmente para as recomendações de filmes para outras pessoas. Desta forma, indiretamente ele estará participando de uma comunidade ativa sobre filmes.

Podemos perceber que a formação de comunidades consciente ou inconscientemente é fator fundamental para a existência de um algoritmo capaz de fazer recomendações de filmes. Na seção a seguir vamos entender o porquê desta necessidade e porque hoje é possível a criação indireta destas comunidades.

4.2 Formação de Comunidades e a Web 2.0

A quantidade de informação armazenada pela mente humana é limitada. Uma forma encontrada para reter mais informação foi através da organização em sociedade. Se não fossem as complexas organizações sociais, as pessoas seriam apenas como os demais primatas. A partir do momento em que o homem criou máquinas capazes de armazenar uma quantidade impressionante de informação – os computadores –, foi possível que a sociedade começasse então a armazenar seu conhecimento em um repositório único. Entretanto, estas informações não foram armazenadas de uma forma organizada de modo a extrair conhecimento das mesmas. Para entender este problema é necessário diferenciar um computador que recebe passivamente a informação dada e um computador que aprende por si mesmo [65]. A área da ciência da computação destinada a estudar esse paradigma é a Inteligência Artificial (IA).

Conforme visto no capítulo 2, a Filtragem Colaborativa (FC) é uma área distinta da IA tradicional, criando métodos para fazer predições automáticas (filtragem) sobre o interesse de um usuário através da coleta de informações do gosto de muitos usuários que estão colaborando. Esta necessidade se dá pelo fato de a FC estar fundamentada em métodos estatísticos para construção de funções de similaridade, e também pelo fato de que, para fazer predições para gostos muito peculiares, é necessário um volume grande de usuários para encontrar pessoas com gostos similares.

Observa-se que o termo **Filtragem Colaborativa** está totalmente sinérgico com as premissas sugeridas pela Web 2.0. Dentro da Web 2.0 o usuário interage cada vez mais com o sistema, colaborando de diferentes formas com uma comunidade virtual. O acesso livre da

internet permite que pessoas do mundo todo possam colaborar em um único sistema. Este grande número de interações permite que gostos muito particulares e específicos ecoem em outras pessoas facilitando o descobrimento de novos produtos e serviços. O mais interessante foi entender que este número de pessoas que fogem do gosto massificado e possuem preferências particulares é muito grande, ao ponto que somadas todas estas pessoas temos um novo grupo muitas vezes até maior que o da "massa". Para este fenômeno surgiu o termo **Long Tail** - cauda-longa em inglês. A empresa Amazon percebeu este novo paradigma e lucra 2 vezes mais com livros pouco conhecidos do que com os *best-sellers*. Além disto a Amazon percebeu que devido ao grande número de usuários que possuía, ela poderia desenvolver um sistema de recomendação de livros de forma a aumentar suas vendas. Segundo Greg Linden, desenvolvedor do sistema de recomendação da Amazon, em 2002 as vendas oriundas de recomendações personalizadas somaram 20 por centro das vendas totais do ano [33].

4.3 The Netflix Prize

A empresa Netflix é a maior locadora de filmes online do mundo, provendo mais de 100.000 títulos de DVDs, 10.000 séries de TV e tendo mais de 8 milhões de usuários apenas nos EUA. Foi a pioneira a explorar o aluguel de filmes pela internet e hoje é uma das maiores empresas do setor de e-commerce americano. Paralelo ao seu negócio de locação de filmes, ela começou desde o início de sua criação a desenvolver um *profile* para cada usuário, no qual é possível, entre outras coisas avaliar um filme com notas discretizadas de 1 a 5. Esta avaliação é utilizada pelo sistema de recomendação proprietário da empresa – Cinematch ®– que faz constantes recomendações de filmes não vistos pelo usuários, visando melhorar a experiência do usuário e, é claro, aumentar as locações de filmes.

Em outubro de 2006 a Netflix Inc. anunciou o **Netflix Prize**, uma competição cujo prêmio é 1 milhão de dólares para aquele que conseguir desenvolver um algoritmo capaz de fazer predições 10% melhor que o Cinematch ®. Para isto a Netflix forneceu uma gigantesca base de dados com 100.480.507 avaliações de 480.189 diferentes usuários sobre 17.770 filmes, referentes às avaliações coletadas pelo site entre outubro de 1998 e dezembro de 2005 e que representam todas as avaliações recebidas pelo Netflix de seus usuários desde a sua criação. Esta base de dados foi submetida na data do anúncio do prêmio ao algoritmo Cinematch ® para a predição de diversas avaliações suprimidas, e foi alcançada uma precisão medida através do RMSE de 0,9525. [3]

Desde então, diversos pesquisadores de diferentes partes do mundo vêm participando da competição e a área de Filtragem Colaborativa nunca evoluiu tanto em tão pouco tempo.

[3]RMSE é a abreviação em inglês de *Root Mean Square Error*, que é uma medida comum para verificar a **precisão** da predição, ou seja, o quanto ela está próxima da avaliação real. Uma definição formal do RMSE é apresentado neste capítulo na seção 4.3.2

Até o momento da escrita da presente dissertação a competição completa um ano e meio, sem ainda um ganhador. A seguir é caracterizado o problema a ser resolvido pela predição e na seqüência são apresentadas as regras da competição, além da sistemática de submissão e validação das predições calculadas.

4.3.1 A Predição

O propósito do algoritmo a ser desenvolvido nesta competição é fazer a predição de que nota um determinado usuário daria para um filme, baseado nas avaliações que ele fez de outros filmes e nas avaliações que outros usuários fizeram destes mesmos filmes e do filme a ser predito. Para que se tenha uma boa predição é necessária a existência de um grande número de usuários e de várias avaliações feitas pelo usuário, cuja nota deverá ser predita. Existem algoritmos que utilizam-se de outros tipos de informação, como por exemplo, que filmes um determinado usuário avaliou e quais ele não avaliou; outros ainda se utilizam de informações baseadas no conteúdo, como gênero do filme, número de Oscars, ator principal, entre outros.

Independentemente da abordagem utilizada, o que se pretende é conseguir predições de qualidade de forma a poder recomendar um filme para um usuário. Em resumo, podemos exemplificar uma predição como demonstrado na tabela 4.1. Em um conjunto de 5 usuários, queremos predizer que nota o usuário 1 daria para o filme 1 ainda não visto. Os quatro últimos usuários avaliaram todos os 5 filmes existentes.

Tabela 4.1: Exemplo Predição

	User 1	***User 2***	***User 3***	***User 4***	***User 5***
Filme 1	?	2	5	1	5
Filme 2	3	4	1	5	3
Filme 3	1	4	1	5	1
Filme 4	5	1	3	2	4
Filme 5	5	2	5	1	5

Neste exemplo, se a nota a ser predita pelo algoritmo para o filme 1 for acima de 4, então este filme será recomendado para o usuário 1.

4.3.2 As Regras

Nesta seção apresentaremos apenas as regras importantes para a compreensão da sistemática de criação do algoritmo, submissão e validação de predições. Maiores detalhes sobre a competição podem ser encontrados em [37].

A competição teve início no dia 2 de outubro de 2006 e foi aberta para qualquer pessoa, praticamente em qualquer país [4]. Para participar não é necessário ser um assinante da Netflix, sendo apenas necessário se registrar, o que pode ser feito diretamente pelo site www.netflixprize.com sem nenhum custo ou taxa. As inscrições podem ser individuais ou em grupo. Caso seja feita a inscrição de um grupo, é necessária a designação de um líder. Feita a inscrição e aceitas as regras da competição, é possível ter acesso à base de dados com todas as informações necessárias para o desenvolvimento das predições. Os dados fornecidos não podem ser utilizados senão para o desenvolvimento de um algoritmo nesta competição. Nenhuma biblioteca, função ou software que não seja de uso livre e público poderá ser utilizado no desenvolvimento do algoritmo.

Para se qualificar ao prêmio de U$1.000.000,00 é necessário submeter uma predição com uma precisão 10% melhor do que a precisão que o algoritmo utilizado no Cinematch ® alcançou na data de lançamento deste prêmio, RMSE=0,9525. Para estimular a competição, a cada ano é oferecido um prêmio parcial denominado *"Progress Prize"* no valor de U$50.000,00, que visa premiar a equipe que obteve o melhor resultado no ano. Porém, para receber o prêmio, esta equipe deverá divulgar em detalhe o algoritmo desenvolvido. O mesmo deve ocorrer para ganhar o prêmio principal, sendo que além de explicar como e por que o seu algoritmo funciona, será necessário ainda fornecer uma licença (não exclusiva) ao Netflix para a utilização do algoritmo.

Todas as predições devem ser enviadas diretamente pelo site à Netflix dentro do formato específico solicitado pela organização da competição, e deve ser respeitado o intervalo de um dia entre uma submissão e outra.

A forma definida pelos organizadores da competição para aferir a precisão de um determinado algoritmo é o cálculo do RMSE (Root-Mean Square Error) entre as variáveis preditas e as variáveis reais. O RMSE pode ser definido assim:

Definição 4.3.1 *: O RMSE de um estimador $\hat{P}$ do conjunto de predições feitas por um determinado algoritmo para um filme f feita por um usuário u - $\hat{P} = (p_{11}, p_{12}, ..., p_{1n})$ - e as avaliações reais $A^* = (a_{11}, a_{12}, ..., a_{1n})$, sendo A^* um subconjunto de A que é todo espaço de avaliações existente, é definido como:*

$$RMSE(\hat{P}) = \sqrt{\frac{\sum_{i=1}^{n}(p_{1n} - a_{1n})^2}{n}}$$

O cálculo do RMSE poderá ser feito em um subconjunto de dados de prova do algoritmo. Entretanto, a predição enviada à Netflix será feita sem saber qual é o RMSE associado, já que não são conhecidas as avaliações do subconjunto de qualificação. Neste caso quem calculará o RMSE é a própria Netflix, deixando público o RMSE alcançado.

[4] Com exceção de alguns países listados em [37]

4.3.3 A Estrutura de Dados

Os dados fornecidos pela Netflix estão subdivididos em 4 arquivos:

- **training.tar**: arquivo que contém A=100.480.507 avaliações discretizadas de 1 a 5, de F=17.770 filmes feitas por U=480.189 diferentes usuários. Este arquivo possui 17.770 arquivos txt, sendo que cada arquivo representa um filme com todas as avaliações feitas para ele. O conjunto todo possui mais de 2Gb de tamanho;
- **probe.txt**: arquivo que contém 1.408.395 pares de filme-usuário, cuja avaliação está presente no arquivo tranining.tar. Este arquivo deve ser usado para cálculo do RMSE de forma a dar uma idéia prévia da eficiência do algoritmo antes de submeter uma predição ao Netflix;
- **qualifying.txt**: arquivo que contém 2.817.131 pares de filme-usuários, para os quais não se conhece a avaliação e que deve ser, portanto, predita pelo sistema de recomendação desenvolvido. O cálculo do RMSE é feito através da submissão da predição ao Netflix, que irá calcular a precisão da predição gerando o RMSE.
- **movie.titles.txt**: arquivo que contém o nome e a data de lançamento dos 17.770 filmes em questão.

O primeiro arquivo training.tar é o arquivo que contém todas as avaliações do conjunto de usuários da Netflix e será utilizado para o treinamento dos algoritmos desenvolvidos. A estrutura interna de cada um dos seus 17.770 arquivos-textos é formada por:

```
MovieID1
UserID1,Avaliação11,Data11
UserID2,Avaliação21,Data21
```

Sendo que:

- MovieID varia de 1 a 17.770 de forma seqüencial
- UserID varia de 1 a 2.649.429, de forma não seqüencial existindo espaços. Existem 480.189 usuários distintos na base
- Avaliações foram feitas através da escolha de 1 a 5 estrelas - variação discreta de 1 a 5
- Datas possuem o formato AAAA-MM-DD

Em nenhum momento são fornecidos os dados dos usuários, para a proteção de sua privacidade e também para a segurança da competição. Para prevenir que inferências sejam feitas sobre os dados, de forma a identificar os usuários – o que poderia dar brechas para fraudes, já que tratam-se de avaliações reais –, foi feita uma série de perturbações nos dados, tanto dentro do arquivo training.tar quanto os do arquivo qualifying.txt. Estas perturbações foram feitas através da inserção de avaliações e de datas alternativas às reais, da supressão de avaliações, entre outras formas de ruídos. Entretanto, mesmo com estes ruídos inseridos dentro do conjunto de dados, o RMSE final encontrado pelo Cinematch ®sobre o conjunto de dados com perturbações não foi modificado significativamente do calculado sobre o conjunto de dados sem ruídos. O RMSE alcançado pelo Cinematch ® anunciado no início da competição foi calculado sobre o conjunto de testes com os ruídos.

Podemos organizar a informação acima descrita através de uma matriz de usuários-filmes que possui $U \times F = 8.532.958.530$ células que poderiam conter avaliações. Isto significa que se todo usuário tivesse avaliado todos os filmes teríamos mais de 8,5 bilhões de avaliações. Todavia, como temos apenas $A = 100.480.507$ avaliações, podemos perceber que 98,9% desta matriz está vazia. Na tabela 4.2 conseguimos visualizar esta estrutura de dados.

Tabela 4.2: Estrutura de Dados Netflix Prize

	User 1	***User 2***	***User 3***	***User 4***	...	***User U***
Filme 1	a_{11}	a_{12}	a_{13}	a_{14}	...	a_{1U}
Filme 2	a_{21}		a_{23}	a_{24}	...	
Filme 3	a_{31}	a_{32}	a_{33}		...	
Filme 4	a_{41}				...	
...	...	...	...	...	...	...
Filme F	a_{F1}	a_{F2}	a_{F3}	a_{F4}	...	a_{FU}

Conforme apresentado na tabela 4.2, a avaliação de uma pessoa para um determinado filme será denominada de agora em diante e exclusivamente neste trabalho de a_{fu}, onde o índice f é relativo ao filme e o índice u é relativo ao usuário. Estruturamos os filmes como linhas nesta matriz, caracterizando o primeiro índice f, pelo fato de que iremos implementar uma abordagem baseada em similaridades de filmes ao invés de similaridades baseadas em usuários. Esta diferença de abordagem permite uma precisão maior da predição como é explicado no capítulo 5. Percebemos também que pode não existir uma determinada avaliação, sendo que 98,9% desta matriz está vazia. No exemplo da tabela 4.2 a avaliação a_{44} não existe.

A estrutura de dados do arquivo probe.txt é formada por pares de filmes-usuários, cuja avaliação deverá ser predita pelo algoritmo desenvolvido e testada sua eficiência através do RMSE. Esta eficiência poderá ser calculada pelo próprio time buscando a avaliação verdadeira no arquivo training.tar. A estrutura de dados deste arquivo é a seguinte:

```
MovieID1:
```

```
UserID11,
UserID12,
...
MovieID2:
UserID21,
UserID22,
```

A estrutura de dados do arquivo qualifying.txt é idêntica a do arquivo probe.txt, porém com um número maior de pares filme-usuário e também com a diferença que para cada par é informada a data da avaliação real. E ao contrário do que ocorre nos pares do arquivo probe.txt, as avaliações não se encontram no conjunto de treinamento dos dados, e portanto, o cálculo do RMSE pode ser feito apenas pela Netflix. A estrutura de dados deste arquivo pode ser vista abaixo:

```
MovieID1:
UserID11,Data11
UserID12,Data12
...
MovieID2:
UserID21,Data21
UserID22,Data22
```

O arquivo movie.titles.txt possui os títulos e as datas de lançamento dos filmes avaliados no conjunto de treinamento. Todos os títulos estão em inglês e correspondem aos títulos utilizados no site Netflix, não necessariamente correspondendo a outros títulos utilizados em outros sites até mesmo no IMDB. [5]

```
MovieID,AnodeLançamento,Título
```

Nenhum outro dado é fornecido senão estes apresentados nesta seção. Nenhum outro dado foi utilizado para o cálculo da predição feita pelo Cinematch ® em outubro de 2006.

4.3.4 Formato da Predição

Para ser válida, cada predição deverá ser submetida através do site em formato específico e conter as predições para todos os pares definidos no subconjunto do arquivo qualifying.txt. Assim sendo, se o arquivo qualifying.txt fosse:

[5] IMDB - Internet Movie Data Base é uma organização não governamental que visa organizar as informações sobre filmes.

```
111:
3245,2005-12-19
5666,2005-12-23
6789,2005-03-14
225:
1234,2005-05-26
3456,2005-11-07
```

uma predição válida teria uma estrutura igual a:

```
111:
3.0
3.4
4.0
225:
1.5
2.0
```

Esta predição significa que o usuário 3245 avaliou o filme 111 com 3 estrelas em 19 de dezembro de 2005, enquanto que o usuário 1234 avaliou o filme 225 provavelmente um pouco acima de 1 estrela em 26 de maio de 2005. Importante salientar que apesar das notas serem discretas, as predições podem possuir valores não-inteiros com até uma casa decimal. A ordem das linhas do conjunto qualifying e a da predição devem ser as mesmas.

Os valores das predições não serão levados a público, sendo que a Netflix garantirá a confidencialidade das mesmas. Todas as predições após submissão são de propriedade da Netflix, que poderá utilizá-las de diferentes formas, contanto que não as exponha em público. Após a conquista do prêmio e conseqüente término da competição, as predições poderão se tornar públicas.

4.3.5 Qualificação e Julgamento de Algoritmos

Após uma submissão válida à Netflix, o algoritmo será qualificado através do cálculo do RMSE. Para tanto, o conjunto de pares filme-usuário presente no arquivo qualifying.txt é dividido em dois subconjuntos de uma forma randômica. Esta divisão é feita para impossibilitar que inferências sejam feitas através de pequenas alterações em duas predições consecutivas, que poderiam, através da diferença do RMSE, chegar na nota real. O RMSE encontrado para o primeiro conjunto de teste chamado “quiz set” será revelado publicamente no site. O RMSE encontrado para o segundo subconjunto denominado “test set” não será

publicado, mas será levado em conta para a qualificação do algoritmo. O RMSE reportado no site, calculado sobre o conjunto "quiz set", serve para anunciar publicamente que existe um algoritmo que potencialmente alcançou um determinado *score*, além de servir também para prover um *feedback* para os participantes.

Algoritmos que forem qualificados só serão elegíveis se forem desenvolvidos ou implementados de forma original, sem infringir nenhuma lei ou regulamento que proteja direito de terceiros; deve ser escrito em inglês e não deve necessitar de qualquer outro software ou licença.

O julgamento de um algoritmo para receber o "Progress Prize" ou "Grand Prize" será feito através da descrição do algoritmo que deverá ser entregue à Netflix em até uma semana após o cálculo do seu RMSE. Esta descrição deverá ser feita em inglês e detalhada suficientemente para que um engenheiro ou um profissional das ciências da computação possa compreender. O código fonte da aplicação que gerou a predição deve ser enviado à Netflix para que se possa refazer a predição de forma a comprovar a sua eficácia.

4.3.6 Fórum dos Competidores

Foi disponibilizado um fórum para que os competidores pudessem trocar idéias, experiências e dúvidas a respeito de seus algoritmos. Este fórum é constantemente utilizado pela maioria dos competidores, mostrando uma colaboração surpreendente em um ambiente competitivo. Todos se ajudam e procuram explicar dúvidas e compartilhar idéias. No momento da escrita desta dissertação os primeiros lugares do ranking utilizaram abordagens criadas por diferentes times, que juntas alcançaram os melhores resultados.

O fórum também é o canal entre os competidores e a Netflix para tirar dúvidas, receber informações atualizadas a respeito da competição e divulgar novos líderes do ranking.

4.3.7 Maracatu Team

Para que fosse possível receber os arquivos com a base de dados da Netflix, foi necessário criar um time para participar da competição. O nome escolhido para o time foi **Maracatu**, em analogia ao ritmo musical afro-brasileiro no qual diferentes instrumentos de percussão são orquestrados de forma a criar uma melodia. Como é grande a dificuldade de implementar algoritmos capazes de lidar com uma grande base de dados, outras pessoas integraram o grupo com a responsabilidade de garantir uma implementação otimizada dos algoritmos desenvolvidos. Participaram deste time João Lourenço Vivan Bernartt, João Bosco Pereira Filho e Alceu de Madeiros.

4.4 Conclusão

A dificuldade na escolha de um filme sugere o desenvolvimento de sistemas de recomendação que podem ser implementados através de abordagens de filtragem colaborativa. Devido à natureza do problema em questão e às condições atuais da indústria cinematográfica norte-americana, que lança mais de 600 novos filmes todo ano, aumentando cada vez mais o número de filmes a serem assistidos, sistemas de recomendações de filmes passaram a desempenhar um importante papel para empresas do ramo. Associado a isto, as novas mídias existentes na internet, baseadas em conceitos da Web 2.0, fazem mais e mais pessoas colaborarem com um sistema único, conseguindo armazenar uma grande quantidade de informações.

Dentro deste contexto, a empresa Netflix – maior locadora de filmes pela internet do mundo – lança uma competição disponibilizando toda a sua base de dados, viabilizando a pesquisa na área de sistemas de recomendação. Além de fornecer os dados – essenciais para testar algoritmos de data-mining –, a competição estabeleceu uma sistemática para validação dos algoritmos contribuindo ainda mais para a evolução dos estudos na área.

Foram apresentadas neste capítulo as características do problema de escolha de um filme, as regras da competição "*Netflix Prize*", bem como a estrutura de dados fornecida pela Netflix e que é utilizada neste trabalho. A sistemática de validação explicada neste capítulo permite a compreensão dos resultados apresentados nos próximos capítulos.