FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

FEUP

Algoritmos Colaborativos para Sistemas de Recomendação

José Graciano Almeida Ramos

Mestrado Integrado em Engenharia Informática e Computação

Orientador na FEUP: Professor Doutor Jaime S. Cardoso

Responsável no INESC-Porto: Mestre Ricardo Sousa

31 de Julho de 2010

Algoritmos Colaborativos para Sistemas de Recomendação

José Graciano Almeida Ramos

Mestrado Integrado em Engenharia Informática e Computação

Aprovado em provas públicas pelo júri:
Presidente: Doutor Rui Camacho
Vogal Externo: Doutor Manuel Ferreira
Orientador: Professor Doutor Jaime S. Cardoso

---

31 de Julho de 2010

Aos meus pais

Resumo

Os *sistemas de recomendação* têm vindo a ganhar um reconhecimento crescente devido maioritariamente ao seu uso no comércio electrónico como é o caso da *Amazon*. Estes sistemas, que consistem em metodologias também denominadas por *filtragem colaborativa*, providenciam recomendações de diferentes produtos através de informação agregada por diversas acções do utilizador do sistema. É então através deste aglomerar de conhecimento que se identifica uma grande vantagem na sua utilização dada a capacidade de generalizar e de gerar listas de recomendações para futuras aquisições que não tenham sido consideradas até então.

Dada a sua natureza ubíqua e transversal a diversas áreas tem-se recorrido assim cada vez mais ao seu uso. Um dos exemplos mais expressivos da intensidade da investigação e desenvolvimento de *algoritmos colaborativos* é o concurso *Netflix* onde pretendia-se prever o filme que um espectador iria gostar efectuando uma recomendação viável.

A presente dissertação de Mestrado intitulada "Algoritmos colaborativos para sistema de recomendação", realizada no INESC Porto, teve como principal objectivo o estudo de diversas metodologias aplicadas à filtragem colaborativa.

Com a elaboração desta dissertação pretende-se dar um contributo no desenvolvimento dos algoritmos colaborativos através da realização de um estudo comparativo das várias abordagens de filtragem colaborativa existentes e da apresentação de uma nova para o problema em estudo.

Nessa tese foi realizada o estudo e a implementação da abordagem *"Collaborative filtering with interlaced generalized linear models"* e como proposta de melhoria propôs-se uma abordagem híbrida usando informações dos utilizadores, dos itens e da matriz das avaliações para prever novas avaliações. Enquanto o primeiro modelo baseia-se apenas na informação contida na matriz das avaliações para prever novas avaliações, o segundo modelo utiliza dados demográficos dos utilizadores e dos itens e a matriz das avaliações para prever novas avaliações, apresentando desempenho comparável com outros modelos do estado da arte.

PALAVRAS CHAVES:

Aprendizagem Automática, Filtragem Colaborativa e Modelos Lineares Generalizados.

Abstract

Recommender systems have been gaining increasing recognition due mainly to its use in electronic commerce such as Amazon. These systems embedded with collaborative filtering methodologies can provide several recommendations for different products. Based on a myriad of an aggregated information regarding to the users and their iteration with the system, several suggestions can result towards acquisition of items that, possibly, never considered before. Therefore, the ability of generalization is hence a major goal of these learning schemes.

Given its ubiquitous and transverse nature, its use has been increasingly been taken. One of the most expressive mark regarding to the intensity of research and development of collaborative filtering algorithms is the Netflix prize competition. Here, the major goal is to suggest a film that a viewer will probably like.

This Master's dissertation entitled "Algorithms for collaborative recommendation system", executed at the INESC Porto, had as main objective the study of various methodologies applied to collaborative filtering. The objectives of this dissertation are to provide a contribution towards the development of collaborative algorithms by performing a comparative study of various existing approaches on collaborative filtering and a new model proposal.

This thesis carried out a study and implementation of the approach "Collaborative filtering with interlaced generalized linear models". As a proposal for improving it, we present a hybrid approach provide recommendations. While the first model is based only on the information contained on the rating, the second model uses not only this information but also demographic data of users and items. Finally, we performed a comparative study where our proposal works well when compared with the baseline model selected from the state of the art.

Agradecimentos

Ao chegar ao término da realização desta tese de Mestrado integrado não poderia deixar de expressar o meu agradecimento a pessoas que de alguma forma contribuíram para a realização da mesma.

Ao meu orientador professor Doutor Jaime S. Cardoso pelo acompanhamento e pela disponibilidade demonstrada durante o período da realização da tese.

Ao Mestre Ricardo Sousa, meu responsável no INESC - Porto, pela disponibilidade, dedicação, acompanhamento e sugestões demonstradas na realização da tese.

Aos meus pais, irmãos, namorada e familiares pelo acompanhamento e incentivo dado durante o meu percurso académico.

À professora Doutora Luísa Meireles pela sua disponibilidade.

Aos meus amigos e colegas pelas manifestações de companheirismo e encorajamento.

Mais uma vez, a todos os meus sinceros agradecimentos.

Índice

Lista de Ilustrações e Tabelas

Abreviaturas e Símbolos

INESC-Porto Instituto de Engenharia de Sistemas e Computadores do Porto.

RISMF Técnica de factorização matricial.

BRISMF Técnica de factorização matricial que consiste no melhoramento da técnica RISMF.

MIT *Massachusetts Institute of Technology*.

f.d.p Função densidade de probabilidade.

MLG Modelos Lineares Generalizados.

SVD *Singular value decomposition*.

MAE *Mean Absolute Error*.

RMSE *Root Mean Square Error*.

$\boldsymbol{\mu}$ Valor esperado.

$\boldsymbol{\sigma^2}$ Variância.

$\boldsymbol{\phi_n}$ Vector de características associado ao utilizador n.

$\boldsymbol{\omega_m}$ Vector de características associado ao item m.

$\boldsymbol{u_n}$ Utilizador índice n.

$\boldsymbol{y_m}$ Item índice m.

$\boldsymbol{\eta_{nm}}$ Expressão da apreciação do item y_m pelo utilizador u_n.

Φ Matriz das características dos utilizadores.

Ω Matriz das características dos itens.

Glossário

Norma Euclidiana
Técnica para cálculo de comprimento de vectores aplicada quando este é definido no espaço bidimensional, também denominado de espaço Euclidiano.

Factorização matricial
Técnica que consiste na decomposição de matrizes transformando-as na forma canónica.

URL
Define o endereço de um recurso localizado numa rede.

Web
Também denominada de WWW, é um sistema de documentos em hipermídia que são interligados e executados na internet.

Cadeias de Markov
Nome atribuído em homenagem ao matemático *Andrei Andreyevich Markov*, é um caso particular dos processos estocásticos com estados discretos que se baseiam no pressuposto de que os estados anteriores são irrelevantes para a previsão dos estados seguintes, desde que o actual seja conhecido. Esta cadeia é uma sequência aleatória x_1, x_2, x_3, …, x_n que pode assumir um valor dentro de um espaço de estados e podem ser visualizados através de uma máquina de estados finita.

Método de Monte Carlo
É um método estatístico utilizado em simulações estocásticas para obter aproximações numéricas de funções complexas.

Data-mining
Consiste num processo de identificação de padrões num conjunto de dados.

Métodos *spline*
Método para cálculo de interpolação entre dois ou mais pontos.

Funções de base radiais
Funções de base radiais (RBFs) são aquelas que apresentam simetria radial, ou seja, dependem apenas (para além de alguns parâmetros conhecidos) da distância $r = \|x - x_j\|$ entre o centro da função e o ponto genérico x [29].

Óptima-de-pareto
É um conceito da economia que define uma situação económica como sendo óptima se

não for possível melhorá-la, ou mais genericamente, a utilidade de um agente sem degradar a utilidade de qualquer outro agente.

Framework
Abstracção que une código comum entre vários projectos de software proporcionando uma funcionalidade genérica.

Web-service
É uma solução utilizada na integração de sistemas e na comunicação entre aplicações diferentes desenvolvidas em plataformas diferentes.

Norma Frobenius
Técnica para cálculo da norma de matrizes baseada no traço da matriz.

Matriz Hessiana
Matriz Hessiana de uma função de N variáveis é a matriz quadrada NxN das derivadas parciais de segunda ordem da função.

Apresentação do INESC-Porto

A presente dissertação foi realizada nas instalações do INESC-Porto tendo como responsável o Mestre Ricardo Sousa.

O INESC Porto - Instituto de Engenharia de Sistemas e Computadores do Porto é uma associação privada sem fins lucrativos reconhecida como instituição de utilidade pública, que adquiriu em 2002 o estatuto de Laboratório Associado. Foi constituído em 18 de Dezembro de 1998, tendo como associados fundadores o INESC, a Universidade do Porto e a Faculdade de Engenharia da Universidade do Porto.

Desenvolve actividades de investigação e desenvolvimento, consultoria, formação avançada e transferência de tecnologia nas áreas de Telecomunicações e Multimédia, Sistemas de Energia, Sistemas de Produção, Sistemas de Informação e Comunicação e Optoelectrónica.

O INESC Porto foi criado para constituir uma interface entre o mundo académico e o mundo empresarial da indústria e dos serviços, bem como a administração pública, no âmbito das Tecnologias de Informação, Telecomunicações e Electrónica, dedicando-se a actividades de investigação científica e desenvolvimento tecnológico, transferência de tecnologia, consultoria e formação avançada.

Procura pautar a sua acção por critérios de inovação, de internacionalização e de impacto no tecido económico e social, sobretudo pelo estabelecimento de um conjunto de parcerias estratégicas que garantam a sua estabilidade institucional e sustentabilidade económica[1].

[1] www.inescporto.pt/apresentacao - acedido em 23/06/2010

Capítulo 1

Introdução

1.1 Motivação

O acesso cada vez mais fácil a novas tecnologias de informação e o consequente crescimento exponencial na quantidade de dados a tratar tornou por sua vez impraticável o seu tratamento em tempo útil.

Devido a esse volume de informação quando é necessário encontrar/seleccionar informação útil é preciso um maior esforço pelas partes envolvidas no processo de tomada de decisão. Essa dificuldade levou ao surgimento dos sistemas de recomendação com o objectivo de ajudar as pessoas a lidarem com a sobrecarga de informação e obter informação em tempo útil. Estes sistemas consistem em metodologias que permitem gerar recomendações de diferentes produtos através da informação agregada da interacção entre o utilizador e o sistema.

Para as empresas o aparecimento deste novo conceito representou uma forma de dar a conhecer aos clientes novos produtos e a possibilidade de aumentarem os valores das vendas. Por outro lado, para os clientes, representou um "conselheiro" que tinha a missão de conquistar a sua confiança ao longo do tempo consoante a qualidade das recomendações apresentadas.

Os sistemas de recomendação proliferaram rapidamente devido ao posicionamento dos clientes e das empresas no mundo empresarial: por um lado, os clientes/consumidores estavam impotentes perante um paradoxo de escolha - a grande quantidade de opções disponíveis e a dificuldade em distinguir entre as ofertas. Por outro lado, os produtores/empresas estavam num pólo oposto onde precisavam de tomar decisões de investimento num ambiente lotado de produtos.

Devido ao aumento do uso dos sistemas de recomendação e das vantagens provenientes da sua utilização surgiram vários modelos de sistemas de recomendação, diferenciando-se principalmente na forma como processam os dados, sendo a filtragem colaborativa o modelo mais utilizado.

A filtragem colaborativa (FC) pode ser descrita como a tarefa de identificar (filtrar) interesses dos utilizadores através da aprendizagem de relações com outros utilizadores (colaboração) [33]. Estes modelos tomam como pressupostos a ténue variação das preferências dos utilizadores e de que comportamentos homogéneos poderão advir recomendações idênticas ou similares.

Na formulação da filtragem colaborativa, um sistema é associado a um conjunto de itens, utilizadores e vice-versa. A estes modelos está ainda associada a possibilidade de definir informação extra por utilizadores referentes a determinados itens. Esta interacção é denominada como *rating/* apreciação / avaliação e é interpretada como a expressão dos interesses dos utilizadores.

Os vários modelos existentes têm como principal objectivo analisar comportamentos dos utilizadores e encontrar padrões de comportamentos que possam ser úteis na previsão dos seus gostos sendo assim fulcral o processamento efectuado com os dados disponíveis.

Actualmente os sistemas de recomendação tornaram-se indispensáveis estando presentes em várias aplicações utilizadas no dia-a-dia. As figuras seguintes ilustram alguns exemplos da sua aplicação.

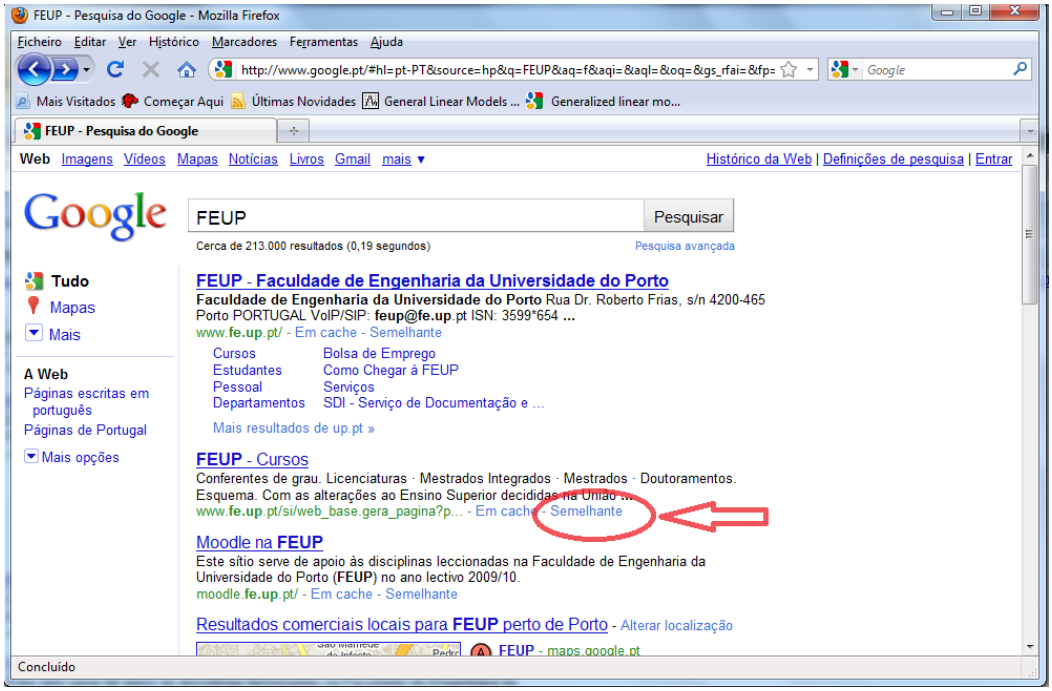

Fig. 1- Aplicação da Filtragem Colaborativa em pesquisas Web.

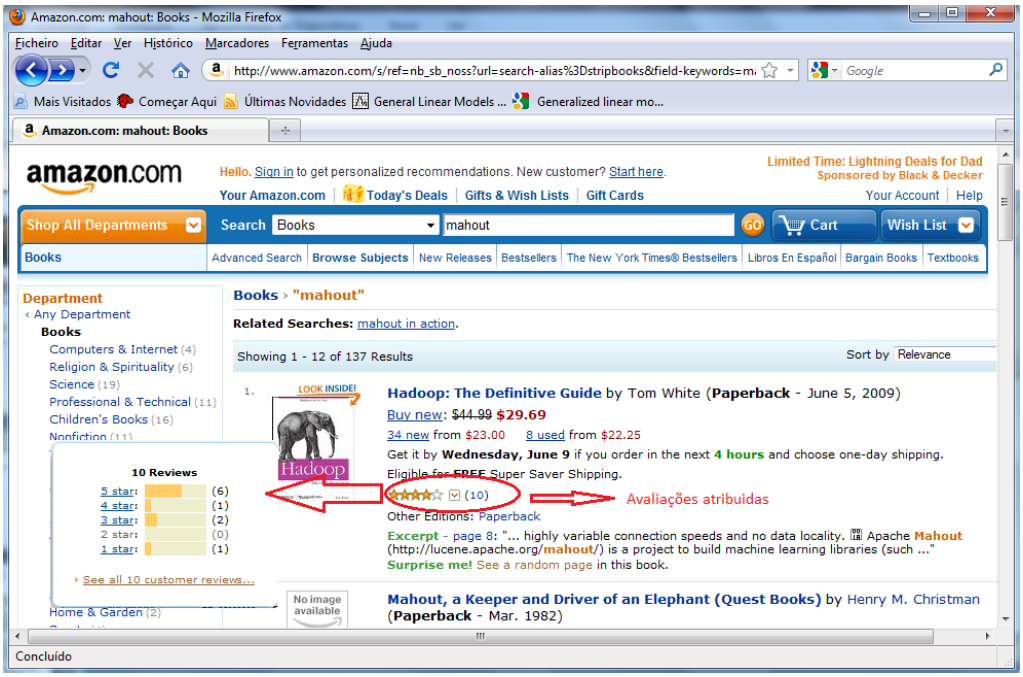

Fig. 2 - Aplicação da Filtragem Colaborativa em compras on-line permitindo avaliar produtos.

Fig. 3- Aplicação da Filtragem Colaborativa em compras on-line para sugerir novos produtos.

Fig. 4- Aplicação de Sistemas de Recomendação para ajudar o cliente a lidar com a diversidade de dados.

1.2 Objectivos

Os sistemas de recomendação têm vindo a ser cada vez mais utilizados com o objectivo de ajudar os utilizadores a lidarem com a sobrecarga de informação e a possibilidade de lhes apresentar produtos que não tivessem sido considerados. A quantidade e complexidade da informação disponível aumentam a cada dia que passa exigindo assim reajustes nos sistemas de recomendação para que possam dar resposta aos novos desafios encontrados. Para atingir este fim têm surgido vários modelos para os sistemas de recomendação que consoante o processamento efectuado dos dados apresentam melhor ou pior desempenho.

É objectivo deste trabalho de dissertação fazer um levantamento do estado da arte sobre os algoritmos colaborativos para sistemas de recomendação, identificar um modelo que seja relevante no estado da arte e proceder à sua implementação e aperfeiçoamento. Para o efeito foi implementado a abordagem *"Collaborative filtering with interlaced generalized linear models*" cujo desempenho foi avaliado no conjunto de dados da movielens [33].

1.3 Contribuições

Neste trabalho é apresentado uma nova abordagem híbrida para os algoritmos colaborativos também denominada de filtragem colaborativa baseada na média das avaliações. Centra na previsão de novas avaliações com base nos modelos lineares generalizados e em regressões lineares. Esta abordagem é considerada híbrida por combinar características dos métodos colaborativos com métodos baseados em conteúdo, a serem apresentados no Capítulo 2, a fim de evitar determinadas limitações dos mesmos.

Com a aplicação da regressão pretende-se encontrar uma função que permita descrever uma relação entre as variáveis independentes, as características dos utilizadores e dos itens, e a variável dependente avaliação atribuída por um utilizador a um determinado item, ou seja, analisar de que forma as características dos utilizadores e dos itens influenciam o processo de avaliação. Ao contrário da grande parte dos modelos identificados no estado da arte, este utiliza informação demográfica dos utilizadores e dos itens e da matriz de avaliações para prever novas avaliações apresentando melhor desempenho quando comparado com alguns modelos do estado da arte.

Apesar de o modelo filtragem colaborativa baseada na média das avaliações utilizar maior quantidade de informação em relação a *"Collaborative filtering with interlaced generalized linear models*" não implica um aumento significativo de esforço computacional na sua execução.

Foi realizado um levantamento do estado da arte dos algoritmos colaborativos e dos sistemas de recomendação apresentando um estudo comparativo de vários modelos de sistemas de recomendação.

1.4 Organização

Para melhor compreensão e exposição dos assuntos abordados, esta dissertação encontra-se estruturada em capítulos. Cada capítulo é inicializado com uma breve síntese do assunto a ser abordado.

A dissertação encontra-se estruturada nos seguintes capítulos:

O Capítulo 2 destina-se a apresentação do estado da arte. No Capítulo 3 é apresentado os fundamentos para os modelos de filtragem colaborativa analisados e apresentados no Capítulo 4. O Capítulo 5 destina-se à apresentação dos resultados e o capítulo 6 destina-se a apresentação de algumas conclusão e perspectivas de trabalho futuro.

No Capítulo 2 é apresentado os conceitos de algoritmos colaborativos e sistemas de recomendação e a relação existente entre esses conceitos. Também é apresentado exemplos da sua aplicação, a sua formulação, alguns dos seus pontos fracos e propostas de melhoria, alguns exemplos de *frameworks* de aplicação de sistemas de recomendação e estudo de possíveis tecnologias a serem utilizadas na implementação de modelos de filtragem colaborativa. No Capítulo 3, a propósito de apresentar conceitos essenciais para a percepção dos modelos de filtragem colaborativa analisados, é apresentado a família exponencial, assim como algumas funções de distribuição pertencente a essa família e os modelos lineares generalizados. No Capítulo 4 é apresentado o modelo de filtragem colaborativa "*Collaborative filtering with interlaced generalized linear models*" que foi objecto de análise nesta dissertação e "Filtragem colaborativa baseada na média das avaliações" que é um novo modelo proposto para a previsão das avaliações. O Capítulo 5 destina-se a apresentação dos resultados obtidos na implementação dos dois modelos apresentados no capítulo anterior. No Capítulo 6 são apresentadas algumas conclusões e algumas perspectivas de trabalho futuro no seguimento deste.

Capítulo 2

Estado da arte

2.1 Introdução

Os sistemas de recomendação permitem às empresas gerarem recomendações personalizadas de produtos aos seus clientes. Para gerar as listas de recomendações personalizadas estes apoiam-se nos algoritmos colaborativos que filtram as informações provenientes das interacções com os clientes. A forma como essa filtragem é realizada permite distinguir dois tipos de algoritmos colaborativos, os denominados de filtragem cognitiva ou baseada em conteúdos e filtragem colaborativa ou social.

A expressão algoritmos colaborativos também denominada de filtragem colaborativa foi criada pelos proponentes do primeiro sistema de recomendação denominado *Tapestry* com o intuito de designar um tipo de sistema específico no qual a filtragem de informação é feita com o auxílio de uma pessoa. Criada essa abordagem, o primeiro Web site a utilizá-lo em grande escala foi o *my yahoo* que o utilizou com estratégias de personalização e, a partir daí, várias outras empresas têm vindo a utilizá-los com o objectivo de apresentar novos produtos aos clientes, aumentando assim a sua fidelização.

Os algoritmos colaborativos têm vindo a adquirir muita importância devido ao seu uso no comércio electrónico, gerando recomendações aos clientes de itens que até aí ainda não tinham sido considerados. O recente aumento da investigação em novas técnicas mais eficazes e eficientes originou uma nova dinâmica como é o caso do concurso lançado pela Netflix, que tem motivado novas abordagens para tentar colmatar alguns pontos fracos dos mesmos.

O presente capítulo encontra-se dividido em dez secções, as quais apresentam de uma forma generalizada, o estado da arte.

Na segunda secção são referenciados os algoritmos colaborativos e os sistemas de recomendação. Na secção três é apresentada a aplicabilidade dos algoritmos colaborativos apresentado alguns exemplos. Na secção quatro é realizada uma formulação dos algoritmos colaborativos e, na secção cinco é apresentada a sua classificação. Na secção seis são apresentados os pontos fracos dos algoritmos colaborativos e, na secção sete, são apresentadas possíveis melhorias para estes pontos identificados. A secção oito destina-se a apresentação de exemplos de *frameworks* para sistemas de recomendação e na secção nove é apresentada a revisão tecnológica, identificando as tecnologias que podem ser utilizadas na implementação dos algoritmos colaborativos. Na secção dez é apresentada uma conclusão do levantamento do estado da arte.

2.2 Algoritmos Colaborativos e Sistemas de Recomendação

Estes sistemas emergiram como uma área de pesquisa independente em meados de 1990, quando os investigadores começaram a concentrar-se em problemas de recomendação que dependiam explicitamente de uma estrutura de avaliação.
Desde a última década têm sido realizados muitos trabalhos e estudos a fim de desenvolver novas abordagens deste tipo de sistemas. O interesse neste sector continua a aumentar constituindo uma importante área de pesquisa. No comércio electrónico utiliza informações dos interesses dos clientes para gerar uma lista de itens a serem recomendados. Desta forma ajuda o utilizador a lidar com a sobrecarga de informação fornecendo-lhe serviços e recomendações personalizadas.

2.3 Algoritmos Colaborativos e Sua Aplicabilidade

Actualmente existe uma vasta quantidade de empresas que utilizam algoritmos colaborativos para gerarem recomendações aos seus clientes, ajudando-os a escolher entre uma colectânea de produtos um subconjunto de menor dimensão que mais se adeqúe ao cliente.
Além disso, os algoritmos colaborativos ajudam os clientes a lidarem com a grande complexidade de dados disponíveis apresentando-lhes novos produtos que até então não foram tidos em conta.
Verifica-se um imenso esforço por parte de empresas e da comunidade académica no aperfeiçoamento dos algoritmos colaborativos, incentivando estudos nesta área, proporcionando assim o aparecimento de novas abordagens para o problema.
Nas subsecções seguintes serão apresentados alguns projectos onde se aplicam este género de algoritmos.

2.3.1 Netflix

O Netflix é uma empresa de aluguer de DVD, que utiliza algoritmos colaborativos no seu sistema de recomendação denominado *Cinematch* gerando recomendações personalizadas aos seus clientes com base no histórico de outros utilizadores. Lançou um concurso cujo objectivo era melhorar o seu sistema de recomendação, tendo como prémio um milhão de dólares para a melhor proposta concebida. No seguimento deste concurso disponibilizou uma base de dados que é considerada a maior base de dados disponível para avaliação dos algoritmos colaborativos [2].
O concurso já de si era bastante interessante, mas tendo em conta o prémio atribuído ao vencedor incentivou o surgimento de várias abordagens para a resolução do problema.
A análise dos resultados obtidos ao longo do concurso originou a elaboração de vários artigos científicos [2]. São ainda analisadas várias abordagens sugeridas para a resolução do mesmo que se baseiam em técnicas de factorização matricial, tais como: RISMF (técnica de factorização matricial que diminui a sobreposição dos elementos da mesma através de penalizações proporcionais ao quadrado da norma Euclidiana dos pesos de cada elemento da matriz), BRISMF (melhoramento da técnica RISMF por introdução de constantes), factorização matricial positiva e semi-positiva, redes neuronais e correcção da factorização matricial baseada na técnica do vizinho mais próximo.

A análise dos resultados obtidos ao longo do referido concurso lançado pela Netflix demonstra que as abordagens acima referidas para o aperfeiçoamento dos algoritmos colaborativos são escaláveis mesmo para sistemas de recomendações com grande quantidade de dados.
A figura seguinte ilustra recomendações geradas pelo *Cinematch* possibilitando ao cliente avaliar os novos filmes apresentados.

Fig. 5- Recomendação geradas pela *Cinematch*.

2.3.2 Referral Web

Este projecto, desenvolvido por Kautz e tal. Em 1997, tem como finalidade identificar e visualizar pessoas ligadas por actividades profissionais, denominando-as de redes sociais. A construção destas é feita com base nos dados introduzidos pelo utilizador, procurando palavras em textos já existentes na internet que mencionem as palavras introduzidas pelo mesmo. No final a rede produzida é representada sob a forma de um grafo.

2.3.3 RINGO

A RINGO foi desenvolvida pela *Massachusetts Institute of Technology (MIT)* em 1995 para recomendação personalizada de música. Explora a similaridade entre os gostos de diferentes utilizadores para recomendar itens, baseando-se no facto dos gostos das pessoas apresentarem tendências gerais e padrões entre gostos e grupos de pessoas. As recomendações são baseadas no perfil de cada utilizador, que é construído com base em descrições de preferências musicais feitas pelos mesmos.
A RINGO deu origem a novos projectos como a Last.fm e a Pandora.

2.3.3.1 Last.fm e Pandora

A Last.fm e a Pandora exploram o conceito apresentado pela RINGO e são dois serviços que proporcionam aos utilizadores ouvirem as suas músicas preferidas ao mesmo tempo que constroem uma base de dados de preferências dos utilizadores cada vez mais personalizada. A personalização desta é feita à medida que os utilizadores vão indicando se gostam ou não duma música que lhes é apresentada.

2.3.4 GroupLens

A *GroupLens* faz parte do departamento de Engenharia e Ciência da Computação da Universidade de Minnesota que se dedica ao estudo de várias áreas como sistemas de recomendação e comunidades *on-line*.

A *GroupLens* desenvolveu vários projectos na área dos algoritmos colaborativos e sistemas de recomendação como a *movielens*, *book-crossing*, *jester joke* e *Wikilens*.

A *movielens* é um site de recomendação de filmes que se baseia na informação que o utilizador fornece dos filmes para gerar recomendações personalizadas de novos filmes que sejam do agrado do utilizador. A lista de recomendações personalizadas é gerada com o auxílio dos algoritmos colaborativos.

A *wikilens* é um sistema de recomendação que permite à comunidade *GroupLens* definir tipos e categorias de itens, avaliá-los e obter recomendações.

A figura seguinte ilustra a previsão de novas avaliações nos dados da *MovieLens.*

Fig. 6- Previsão de novas avaliações.

2.3.5 Fab

O *Fab* é um sistema desenvolvido pela Universidade de *Standford*, com o objectivo de ajudar os utilizadores a filtrar a enorme quantidade de dados disponíveis na internet, recomendando documentos com base nos seus conteúdos. Este sistema combina a filtragem baseada no conteúdo e a filtragem colaborativa, com o objectivo de colmatar falhas dessas abordagens. A sua estrutura híbrida permite o reconhecimento automático de questões emergentes relevantes para vários grupos de utilizadores.

2.3.6 Collaborative Recommender Agent - CORA

É um sistema distribuído assíncrono, desenvolvido pela Universidade de *Zurique* com o objectivo de filtrar documentos *Web* habilitando os utilizadores a recomendar URL´s através de um simples clique.

Fig. 7- Interface do sistema CORA.

2.3.7 Amazon.com

A *Amazon.com* utiliza algoritmos colaborativos para personalizar a loja online ao gosto de cada cliente, mudando radicalmente o seu aspecto de acordo com os interesses do mesmo**,** por exemplo: personalizar a página com conteúdo informático para um engenheiro informático ou artigos relacionados com a gravidez para novas mães.

Na página da *Amazon* existe uma secção denominada *"Your Recommendation",* que liga o cliente a uma área onde pode filtrar as suas recomendações por linhas e áreas de produtos, avaliar os produtos recomendados, as compras efectuadas e ver a relação entre os itens. Essa recomendação de produtos é baseada no cartão de compras do cliente estabelecendo uma analogia com a disposição de produtos num supermercado, mas estando direccionada para o interesse de cada cliente.

Fig. 8- Recomendações geradas pela *Amazon.com.*

2.3.8 eBay™

O site de leilões on-line *eBay.com* possui estratégias de recomendação como o direito à resposta e comprador pessoal. O primeiro permite aos compradores e vendedores avaliarem o seu parceiro de negócio, de acordo com o grau de satisfação da compra. Por outro lado, o comprador pessoal permite aos clientes indicarem os itens que têm interesse em comprar.

2.3.9 Redes Sociais

Grande parte das redes sociais como *facebook* e *hi5* utilizam os sistemas de recomendação para apresentarem novos amigos ou aplicações aos utilizadores. Os algoritmos de recomendação utilizados em grande parte das redes sociais são baseados em filtragem colaborativa e no conceito de vizinhança, mais precisamente os k vizinhos mais próximos que são definidos através da definição de co-relação entre os utilizadores.

Fig. 9 - Aplicação dos sistemas de recomendação nas redes sociais.

2.4 Formulação dos Algoritmos Colaborativos

Na maioria das formulações, o problema de recomendação é reduzido a uma questão de estimativa de classificações de itens que não foram vistos ou avaliados pelo utilizador. Intuitivamente, esta estimativa baseia-se normalmente em avaliações dadas pelo utilizador para outros itens e em algumas outras informações. Uma vez que se pode estimar classificações para os itens ainda não avaliados, pode-se recomendar ao(s) utilizador(s) o(s) item(s) com a maior avaliação estimada(s).

Na formulação define-se:

- **U:** conjunto de utilizadores.
- **Y:** conjunto de todos os itens que podem ser recomendados.
- **T:** Função de utilidade que mede a utilidade do item y para o utilizador u, isto é **UxY → R**, onde **R** é um conjunto totalmente ordenado e para cada utilizador u pertencente a **U** pretende-se escolher o artigo y' pertencente a **Y** que maximiza a utilidade para o utilizador, mais formalmente:

$$\forall u \in U, \quad \grave{y}_u = \underset{y \in Y}{\operatorname{argmax}} T(u, y) \tag{2.1}$$

Neste tipo de sistema a utilidade de um item é normalmente representada por uma classificação/avaliação que indica como o utilizador gostou desse item em particular. Cada

elemento do espaço de utilizadores **U** pode ser definido como um perfil, que inclui várias características dos utilizadores como idade, género, etc.
O problema central dos sistemas de recomendação reside na função de utilidade **T** que normalmente não é totalmente definida no espaço **U**x**Y**, mas apenas num subconjunto. Isto significa que **T** precisa ser extrapolado ao espaço **U**x**Y**. Normalmente os utilizadores classificam/avaliam apenas alguns dos itens que conhecem, por isso o motor de recomendação deve estar apto para estimar a classificação dos itens não classificados. A extrapolação das avaliações conhecidas para as desconhecidas pode ser feita por especificação de heurísticas que definem a função de utilidade e validam empiricamente o seu desempenho, ou estimar a função de utilidade que optimiza certos critérios de desempenho. Após a estimativa das avaliações desconhecidas é seleccionado o item com maior avaliação e este é recomendado ao utilizador, alternativamente pode ser recomendado o conjunto dos N itens com a melhor avaliação ou um conjunto de utilizadores para um item.

2.5 Classificação dos Algoritmos Colaborativos

Os algoritmos colaborativos são classificados de acordo com a abordagem que usam para estimar as avaliações, existindo as seguintes categorias:

2.5.1 Métodos Baseados em Pesquisas

Estes métodos tratam o problema como uma procura de itens relacionados entre si, em que baseada numa avaliação do cliente procura outros itens que sejam do mesmo autor, artista ou director. Para utilizadores com milhares de avaliações é impensável fazer *query* baseando-se em todos os itens, sendo necessário utilizar subconjuntos dos dados reduzindo a qualidade da recomendação gerada. Para resolver este problema de escalabilidade constrói palavras-chave, categorias e autores *off-line*, mas falha ao gerar recomendações com interesses segmentados por títulos.

2.5.2 Filtragem Baseada no Conteúdo

Esta técnica foi denominada de filtragem baseada no conteúdo por gerar recomendações baseadas na análise do conteúdo dos itens e no perfil do utilizador. A informação acerca do tipo de itens pode ser obtida implicitamente ou explicitamente, através de métodos tais como, questionários, que solicitam opiniões e avaliações dos utilizadores ou através de algoritmos de aprendizagem que apreendem os gostos e interesses dos utilizadores, de acordo com as suas acções.
No caso da aplicação de métodos como questionários para obter informação dos gostos e interesses dos utilizadores sobre certos itens, estes avaliam os itens e baseado nesta avaliação o sistema procura itens que vão de encontro com o que foi classificado. Esta abordagem tem vindo a ser substituída por outra, que consiste na implementação de sistemas que aprendem com as acções dos utilizadores. Tal deve-se ao facto de ser incómodo a implementação de questionários e à existência de diversas limitações, como o facto do conteúdo dos dados provenientes da avaliação ser pouco estruturado, difícil de analisar e da pouca clareza do conteúdo das avaliações devido ao uso de sinónimos.

2.5.3 Filtragem Colaborativa

Esta abordagem foi desenvolvida para complementar pontos que estavam em aberto na filtragem baseada no conteúdo. Diferencia-se da mesma por não exigir a compreensão ou reconhecimento do conteúdo dos itens, possuindo algumas vantagens, como por exemplo, a possibilidade de apresentar aos utilizadores recomendações inesperadas. A recomendação é baseada em produtos que pessoas com gostos e preferências similares apreciaram no passado.

Nesta abordagem o utilizador é representado como um vector N dimensional, muito disperso, sendo os seus elementos positivos ou negativos consoante a avaliação dos produtos seja positiva ou negativa. Normalmente multiplica-se o vector pela frequência inversa (o inverso do número de utilizadores que compraram ou avaliaram o item). O grau de similaridade entre dois utilizadores U_i *e* U_j é calculado através da fórmula [3]:

$$\textbf{similaridade}\left(\vec{U_i}, \vec{U_j}\right) = \textbf{cos}(\vec{U_i}, \vec{U_j}) = \frac{\vec{U_i}.\vec{U_j}}{\|\vec{U_i}\| * \|\vec{U_j}\|} \qquad (2.2)$$

A aplicação deste método é computacionalmente dispendiosa porque apresenta uma complexidade O(NM), no pior caso, onde N é o número de utilizadores e M é o número de itens [3] sendo impraticável para grande quantidade de dados.

Para aumentar a sua eficiência é possível diminuir a quantidade de dados dos utilizadores e reduzir o espaço de dados dos itens. A redução de quantidade de dados dos utilizadores pode ser feita escolhendo uma amostra aleatória destes ou descartando utilizadores com poucas avaliações. A redução do espaço de dados dos itens pode ser feita eliminando alguns itens de acordo com a sua popularidade.

Essas reduções do espaço de estado diminuem consideravelmente a qualidade da recomendação gerada, visto que, limitam a recomendação a itens específicos e descartam os mais ou menos populares, diminuindo a similaridade entre os utilizadores.

Este tipo de filtragem tem vantagens relativamente à baseada em conteúdos, porque está apta a filtrar qualquer tipo de itens, como texto, música, fotos e vídeos [6].

Apesar destas vantagens relativamente à filtragem baseada em conteúdos, a filtragem colaborativa possui sérias limitações na qualidade das avaliações recomendadas, denominadas de *"Sparsity problem"* e *"cold start problem"*. O problema de dados esparsos ocorre quando o número de avaliações é insuficiente para encontrar itens ou utilizadores similares. *Cold start problem* pode ser dividida em *cold-start item* e *cold-start users* e reside na adição de novos utilizadores ou novos itens que não se encontravam previamente no sistema.

Nas situações acima referidas o sistema não está apto a gerar recomendações de boa qualidade. Para resolver estas limitações *Heung-Nam et al* [6] propuseram uma nova abordagem intitulada *collaborative tagging* que permite aos utilizadores anotar conteúdos com palavras-chave descritivas. Foram desenvolvidas duas abordagens para sistemas de recomendação baseadas em filtragem colaborativa [6]:

- Filtragem colaborativa baseada em memória;
- Filtragem colaborativa baseada em modelos;

Para melhorar o escalonamento e o desempenho em tempo real para grandes aplicações, foram desenvolvidas várias técnicas de recomendação baseada em modelos pré-existentes. Normalmente os sistemas de filtragem colaborativa baseiam-se em dois passos:

1º. Determinar o grupo vizinho, os utilizadores que têm preferências similares ao utilizador actual (em filtragem colaborativa baseada no utilizador) ou o grupo de itens similares ao item seleccionado (em filtragem colaborativa baseada em item) que pode ser definido utilizando grande variedade de métodos computacionais.

2º. Com base no grupo de vizinhos, são obtidos os valores previstos de determinados itens e, de seguida, os N itens com maior valor previsto de interesse para o utilizador são identificados. Os valores previstos estimam a probabilidade do utilizador-alvo preferir determinado item.

2.5.4 Modelo de Conjuntos (Cluster)

Nesta abordagem, a base de dados dos clientes é dividida em vários segmentos e a tarefa é transformada num problema de classificação. O objectivo do algoritmo é atribuir o utilizador a um segmento contendo os clientes mais similares a este utilizando as avaliações e compras dos clientes no segmento para gerar recomendações. Estes são normalmente criados usando técnicas de *clustering* ou algoritmos de aprendizagem não supervisionada, embora algumas aplicações determinem os segmentos manualmente.
Existem vários métodos para construir os segmentos, sendo *k-means* o mais utilizado. O algoritmo requer unicamente o conhecimento *à priori* do número (K) dos centróides existentes. Seguidamente, num processo completamente automatizado, o algoritmo iterativamente irá agregar as instâncias mais próximas dos K centróides. Além desse género de métodos que atribui um utilizador a um conjunto, existem métodos que permitem atribuir um utilizador a mais do que um conjunto.
A técnica de *cluster* tem maior escalabilidade online, do que a filtragem colaborativa porque a construção dos *clusters* é feita *off-line*. Apesar de considerar todos os clientes num segmento como sendo similares para gerar recomendações, estas são pouco relevantes.
É possível melhorar a qualidade das recomendações geradas usando numerosos segmentos, mas a classificação *on-line* do utilizador torna-se mais complexa do que encontrar utilizadores similares utilizando filtragem colaborativa, palavras-chave ou assuntos semelhantes à avaliação do cliente.

2.5.5 Filtragem Colaborativa Item-a-Item

Este algoritmo foi desenvolvido pela *Amazon*, devido à dificuldade dos algoritmos existentes em processar a sua grande quantidade de dados. Estes são processados pelo algoritmo, produzindo recomendações de boa qualidade em tempo real. Diferencia-se das outras abordagens por fazer grande parte dos cálculos que exigem maior esforço computacional *off-line*.
Calcula a similaridade entre dois itens através de (2.2) e constrói a tabela de itens similares *off-line*. Dada a tabela de itens similares o algoritmo encontra itens similares a

cada avaliação do utilizador, agrega-os e recomenda os itens mais populares ou com maior correlação. Esta computação é muito rápida, dependendo apenas do número de itens que o cliente avaliou ou comprou. A componente *on-line* do algoritmo é encontrar itens similares para as avaliações do cliente [3].

2.5.6 Abordagem Híbrida

Este tipo de abordagem combina métodos colaborativos e métodos baseados em conteúdo, com o objectivo de evitar determinadas limitações destes métodos. Estudos realizados mostram que a combinação linear de vários métodos pode levar a uma melhor solução [2].
A combinação destes métodos pode ser realizada de quatro formas distintas:

- **Combinar as recomendações obtidas separadamente**

Uma forma de construir recomendação híbrida é implementar separadamente sistemas baseados em conteúdo e sistemas colaborativos e, em seguida, aplicar um dos seguintes cenários:

1º. Combinar os *outputs;*
2º. Usar apenas uma das recomendações geradas, baseada num critério de qualidade para seleccionar a mais apropriada.

- **Adicionar características dos modelos baseados em conteúdo aos modelos colaborativos.**

Diversos sistemas de recomendação híbridos estão associados a técnicas colaborativas tradicionais mas apenas mantêm os perfis baseados em conteúdo para cada utilizador. Estes perfis, e não os itens geralmente avaliados, são usados para calcular a similaridade entre dois utilizadores. Isto acarreta alguns problemas de disparidade nas avaliações de uma abordagem colaborativa pura: normalmente não haverá muitos pares de utilizadores que terão um número significativo de produtos avaliados.
Um benefício desta abordagem é que é possível recomendar um artigo aos utilizadores, mesmo que este não seja avaliado por outros com perfis similares.
Como resultado, os utilizadores cujas avaliações estejam de acordo com alguns dos filtros das avaliações podem receber melhores recomendações.

- **Adicionar características colaborativas aos modelos baseados em conteúdo**

A abordagem mais popular nesta categoria é utilizar técnicas de redução da dimensão num grupo de perfis baseados em conteúdo. Isto resulta numa melhoria de desempenho comparado com a abordagem baseada em conteúdo puro.

- **Desenvolver um modelo único unificador de recomendação**

Esta abordagem tem sido a mais seguida nos últimos anos. *Popescul et al. e Schein et al*. propuseram um método probabilístico para combinar recomendações colaborativas e baseadas em conteúdo. Contudo, outra aproximação foi proposta onde os efeitos misturados dos modelos de regressão *Bayesianos* são utilizados para aplicar as cadeias de *Markov* e os métodos de *Monte Carlo* para a avaliação e a previsão dos parâmetros. Os parâmetros desconhecidos deste modelo são estimados a partir dos dados de avaliações

já conhecidas, usando os métodos da cadeia de *Markov* e *Monte Carlo*. Além disso, diversos artigos, como [7], [8], [9], [10], comparam empiricamente o desempenho das abordagens híbridas com os métodos colaborativos puros e com os métodos baseados em conteúdo puros demonstrando que os métodos híbridos podem fornecer recomendações mais exactas do que aproximações puras.
A tabela seguinte faz uma síntese da classificação dos sistemas de recomendação e das técnicas geralmente utilizadas em cada modelo.

Tabela 1- Classificação dos sistemas de recomendação.

Modelo de Recomendação	Técnica de Recomendação (geralmente utilizada)	
	Baseada em Heurísticas	Baseada em Modelos
Baseada em Conteúdo	Conjuntos	Classificador Bayesiano Conjuntos Árvores de decisão Redes neuronais
Filtragem Colaborativa	Vizinho mais próximo Conjuntos Teoria dos grafos	Redes Bayesianas Conjuntos Redes neuronais Regressão linear Modelos probabilísticos
Híbrida	Combinação linear das avaliações previstas Modelos de votação Incorporar um componente duma heurística como parte doutra	Incorporar um componente dum modelo como parte doutro Construir um modelo unificador

2.6 Problemas dos Algoritmos Colaborativos

O desenvolvimento dos algoritmos colaborativos encontra-se em plena progressão devido a grande diversidade e complexidade de dados e consequentemente da constante alteração dos requisitos dos sistemas de recomendação. Apesar dessa progressão e da diversidade de modelos para a sua implementação depara-se com os seguintes aspectos que continuam carecendo de aperfeiçoamento:

2.6.1 Análise Limitada de Conteúdos

Técnicas baseadas em conteúdo são limitadas por características explicitamente associadas ao objecto que estes sistemas recomendam. Além disso, se dois itens diferentes forem representados pelo mesmo grupo de características são indistinguíveis.

2.6.2 Super Especialização

Quando o sistema pode recomendar apenas artigos que vão ao encontro de um perfil de utilizador, este fica limitado a ter recomendações de artigos que são similares aos já avaliados. Para resolver este problema na filtragem de informação foi proposto o uso de algoritmos genéticos. Por vezes um item pode não ser recomendado se for similar a outros itens que o utilizador já viu, como por exemplo, diferentes artigos descrevendo o mesmo evento. O ideal seria apresentar ao utilizador um conjunto de opções e não um conjunto homogéneo de alternativas.

2.6.3 Problema do Novo Utilizador

O utilizador precisa de avaliar um número suficiente de artigos para que o sistema de recomendação possa realmente entender as suas preferências e apresentar uma recomendação viável. Para contornar este problema a maioria dos sistemas usa a técnica da recomendação híbrida combinada com a baseada em conteúdo e técnicas colaborativas.

2.6.4 Problema do Novo Item

Enquanto um novo item não for recomendado por um número substancial de utilizadores, o sistema não estará apto a recomendá-lo. Este facto pode ser colmatado com a abordagem da recomendação híbrida.

2.6.5 Problemas com Dados Esparsos

O sucesso de um algoritmo colaborativo depende da disponibilidade de uma quantidade crítica de utilizadores. Uma forma de superar o problema da disparidade de avaliações é utilizar informação do perfil quando calcula a similaridade do utilizador. Isto é, dois utilizadores podem ser considerados similares não apenas se avaliarem igualmente o mesmo produto, mas também se pertencerem ao mesmo segmento demográfico. Esta extensão de filtragem colaborativa tradicional é por vezes designada de filtragem demográfica.

2.7 Melhorias propostas para os Algoritmos Colaborativos

Na última década tem-se verificado um avanço progressivo dos algoritmos colaborativos mas apesar de todo esse avanço a geração actual deste tipo de sistemas continua a necessitar de optimizações para melhorar a eficácia e poderem então ser aplicados em qualquer situação da vida real. Essas melhorias incluem o entendimento do utilizador e dos itens, a incorporação da informação contextual no processo de recomendação, a possibilidade de suportar avaliações multi-critérios e de fornecer tipos de recomendações mais flexíveis e menos intrusivas [1].

2.7.1 Compreensão detalhada dos Utilizadores e dos Itens

Como foi mencionado em [7], [11], [12] e [13], a maioria dos métodos de recomendação produzem avaliações baseadas numa compreensão limitada dos utilizadores e dos

itens. Contudo ficam restringidos aos perfis dos utilizadores e dos itens e não tiram partido da vantagem da informação no histórico transaccional do utilizador e em outros dados disponíveis.
Além da utilização de características tradicionais do perfil, como as palavras-chave e os dados demográficos simples do utilizador [14], [15], este pode ser construído utilizando técnicas mais avançadas de construção do perfil baseadas em regras de *data mining* [16], [17], sequências [18], e assinaturas [19] que descrevem interesses de um utilizador.
A função de utilidade T_{ij}, que estima a utilidade do item i para o utilizador j, depende da aplicação e pode ser especificada pelo utilizador. Este pode usar vários métodos e várias heurísticas como o vizinho mais próximo, árvores de decisão, métodos *spline*, funções de base radial, regressões, redes neuronais e métodos de aprendizagem relacional. Apesar da existência de vários métodos para calcular a função de utilidade a maioria dos sistemas de recomendação existentes fazem a função $\boldsymbol{T_{ij}}$ dependente apenas de um subconjunto do espaço de *input* R, **U** e **Y**, referenciados na secção 2.4.

2.7.2 Extensões para Técnicas de Recomendação Baseadas em Modelos

A maioria das técnicas baseadas em modelos utiliza estimativas rigorosas de avaliações baseadas em várias técnicas estatísticas e instrumentos de aprendizagem. No entanto outras áreas da matemática e da informática, como a teoria matemática da aproximação podem contribuir para o desenvolvimento de melhores métodos de estimativa de avaliações [1]. Um exemplo de uma abordagem para definir T_{ij} é o uso de funções de base radial que possuem a vantagem de serem estudadas extensivamente na teoria de aproximação [20], [21].

2.7.3 Multi-dimensionalidade da Recomendação

A geração corrente de sistemas de recomendação opera no espaço bidimensional Utilizador x Item, ou seja, gera a recomendação baseando-se apenas nas informações do utilizador e do item. Não tem em consideração informações contextuais que podem ser cruciais em algumas aplicações [1]. Noutras situações não será suficiente gerar uma recomendação para o utilizador. O sistema de recomendação precisa de ter em conta informação contextual, como o tempo, o espaço e a companhia do utilizador quando recomenda produtos, sendo por isso importante estender os métodos de recomendação tradicionais bidimensionais para multidimensionais. O conhecimento das tarefas do utilizador poderia ser útil nessa extensão. A função de utilidade passaria a ser definida no espaço multidimensional $D_1\ X\ D_2\ X, \ldots, X\ D_n$. Apesar das conveniências desta extensão grande parte dos algoritmos bidimensionais não podem ser directamente estendidos ao caso multidimensional, como foi argumentado em [11] e [22].
Uma outra aproximação possível para produzir recomendações multidimensionais seria desdobrar o método *Bayesiano* hierárquico apresentado em [23], que pode ser estendido de bidimensional para multi-dimensional [1].

2.7.3.1 Avaliações Multi-Critérios

A maioria dos sistemas de recomendação actuais trabalha com base em critérios de avaliações únicos. No entanto em algumas aplicações é crucial incorporar avaliações multi-critérios nos métodos de recomendação.

Normalmente as soluções para os problemas de optimização multi-critério incluem [1]:

1. Encontrar solução "óptima de pareto";
2. Fazer uma combinação linear de múltiplos critérios e reduzir o problema a uma optimização de critério único;
3. Optimizar o critério mais importante e converter os outros em restrições;
4. Optimizar um critério em cada iteração, convertendo uma solução óptima em restrições.

2.7.3.2 Intrusividade

Muitos sistemas de recomendação são intrusivos, na medida em que requerem um *feed-back* explícito do utilizador e muitas vezes com um elevado nível de envolvimento do mesmo. Este pode ser formulado como um problema de optimização que tenta encontrar um número óptimo de pedidos de avaliações.

2.7.3.3 Flexibilidade

A maioria dos métodos de recomendação são inflexíveis, no sentido de serem muito restritos aos sistemas dos fornecedores e, consequentemente, suportarem apenas um conjunto pré-definido e fixo de recomendações. Como consequência, o utilizador não pode personalizar recomendações de acordo com as suas necessidades em tempo real.

2.7.3.4 Eficácia da Recomendação

Na maioria dos sistemas de recomendação da literatura, a avaliação do desempenho de algoritmos da recomendação é feita geralmente em termos de medidores de cobertura e de exactidão, mas esta medida de avaliação empírica possui certas limitações [1]. Os resultados da avaliação mostram somente quão exacto o sistema é em relação aos artigos que o utilizador decidiu avaliar, mas a habilidade do sistema de avaliar correctamente um artigo aleatório não é testada.

2.8 *Frameworks* para Sistemas de Recomendação

Neste capítulo é apresentado algumas *frameworks* utilizadas para desenvolvimento e teste de aplicações de sistemas de recomendação.

2.8.1 CofiRank

O *CofiRank*, também denominado de *cofi*, é uma *framework* de filtragem colaborativa que tem como objectivo prever preferências dos utilizadores baseando no histórico das avaliações.

O modelo é construído baseado na abordagem *Maximum Margin Matrix Factorization* expandida nos seguintes aspectos [2] :

- Faz uso das tecnologias de optimização do estado da arte tornando aplicável a grandes bases de dados;

[2] http://www.cofirank.org/- acedido em 10-06-2010

- Está apto a gerar previsões estruturadas, por exemplo, prever a ordem relativa de preferência de filmes em vez de prever uma avaliação absoluta. O modelo está adequado para prever o que o utilizador gosta ou não, propriedade importante dos sistemas de recomendação;
- É facilmente paralelizável tirando partido da máquina *multi-core* ou de conjuntos de trabalho.

2.8.2 C/Matlab Toolkit for Collaborative Filtering

C/Matlab toolkit é um conjunto de funções que implementa vários modelos de filtragem colaborativa desenvolvidas em C e Matlab. Além de implementar modelos de filtragem colaborativa disponibiliza funções para avaliá-los.[3]

2.8.3 Suggest

Suggest é uma *framework* de sistemas de recomendação que implementa a abordagens top-N baseada no conteúdo do utilizador e do item. A abordagem top-N consiste em encontrar o conjunto de N itens que serão de interesse de um determinado utilizador.
A *framework* fornece recomendações de elevada qualidade sendo aplicável a grandes bases de dados.[4]

2.8.4 Taste

Taste é uma *framework* de filtragem colaborativa em java que suporta métodos de filtragem colaborativa baseados em memória. Permite o acesso a bases de dados e disponibiliza as recomendações através de *Web-service*[5].

2.9 Revisão tecnológica

Foi realizado um estudo das possíveis tecnologias a serem utilizadas na implementação de algoritmos colaborativos e constatou-se que existe uma grande diversidade, como é o caso de *Java, C, Matlab, javaScript e PHP,* que podem ser complementadas utilizando técnicas da estatística ou algoritmos de inteligência artificial como redes neuronais e algoritmos genéticos [24]. No entanto existem alguns sistemas de recomendação implementados, como é o caso de *Rec6* e *SisRecCol,* que associam várias tecnologias. *Rec6* foi criado sob a plataforma *LAMP (Linux, Apache, Mysql e PHP)* [25], o sistema *SisRecCol* – (Sistema de Recomendação para Apoio à Colaboração) - foi implementado utilizando as linguagens de programação *PHP* e *Javascript*, a base de dados *MySql*, servidor *Web Apache* e sistema operativo *Linux* [26].
Nos sistemas de recomendação toda a informação proveniente da interacção entre o utilizador e o sistema, nomeadamente os dados dos itens, dos utilizadores e das avaliações é guardada em Matrizes. Sendo o MATLAB uma ferramenta de programação, cujo nome advêm de *Matrix Laboratory*, que tem como principal estrutura de dados a matriz, tem-se revelado muito eficiente na execução de cálculos matriciais e na implementação de sistemas de recomendação.

[3] http://www-2.cs.cmu.edu/~lebanon/IR-lab.htm - acedido em 10-06-2010
[4] http://glaros.dtc.umn.edu/gkhome/suggest/overview- acedido em 10-06-2010
[5] http://taste.sourceforge.net/ - acedido em 10-06-2010

Nesta tese será utilizado o *MATLAB* pelas razões acima referidas e por ser uma ferramenta robusta para problemas de optimização. Possui opções muito eficientes e incorpora *toolboxes* relacionadas com algoritmos identificados no estado da arte sendo possível integrá-las em aplicações desenvolvidas noutras linguagens de programação. É muito utilizado em ambiente académico e industrial e tem sido uma ferramenta padrão em muitas disciplinas científicas e técnicas de prototipagem de novos algoritmos. É ainda útil na análise de dados devido às suas vantagens de programação simples e directa e ao uso fácil de gráficos que facilitam a implementação inicial de algoritmos.

2.10 Conclusão

Neste capítulo foi apresentado o estudo efectuado sobre os algoritmos colaborativos e sistemas de recomendação onde foi realizado uma contextualização ao tema, identificando a utilização dos algoritmos colaborativos nos sistemas de recomendação, os vários modelos de sistemas de recomendação existentes, os respectivos pontos fracos e propostas de melhorias, alguns exemplos de *frameworks* para sistemas de recomendação e a revisão tecnológica apresentando um estudo de possíveis tecnologias a serem utilizadas na implementação de modelos de filtragem colaborativa.

Ao longo deste estudo ficou claro que na última década os algoritmos colaborativos têm adquirido uma importância considerável, conseguindo avanços significativos devido ao surgimento de novas abordagens para o problema. Apesar deste avanço, estes continuam a necessitar de optimizações para poderem ser aplicadas em situações da vida real. Na maioria das situações é preciso considerar informações provenientes de vários contextos, o que não é considerado pela geração actual de algoritmos colaborativos por se encontrarem centralizados apenas na informação dos itens e dos utilizadores.

Capítulo 3

Fundamentos da Filtragem colaborativa

Neste capítulo serão abordados alguns conceitos fundamentais que servirão de base para as metodologias descritas e utilizadas na presente tese. Mais concretamente, iremos apresentar na Secção 3.1 um conceito importante da teoria estatística, a família exponencial, que servirá para introduzir os modelos lineares generalizados elaborados na Secção 3.2.

3.1 Família Exponencial

Considera-se que uma variável aleatória Y pertencente à família exponencial de dispersão (ou simplesmente família exponencial) se for possível escrever a sua função densidade de probabilidade (f.d.p) na forma:

$$F(y|\theta,\phi) = \exp\left\{ \frac{Y\theta - b(\theta)}{a(\phi)} + c(y,\phi) \right\} \quad (3.1)$$

Onde θ e ϕ são parâmetros escalares, a(·), b(·) e c(· , ·) são funções reais conhecidas [34].

Existe uma grande quantidade de distribuições pertencentes à família exponencial, sendo as distribuições Normal, Gama, Poisson, Binomial e Normal Inversa identificadas como as principais distribuições pertencentes a essa família.

3.1.1 Distribuição Normal

A distribuição normal é a mais frequentemente utilizada para descrever fenómenos traduzidos em variáveis aleatórias contínuas devido à sua forma e às suas propriedades.

Define-se que sendo X uma variável aleatória resultante da soma de um grande número de efeitos provocados por causas independentes, em que o efeito de cada causa é negligenciável em relação à soma de todos os outros efeitos, então X segue uma distribuição aproximadamente normal [35]. Define-se uma distribuição normal a partir de dois parâmetros: o seu valor esperado μ que toma qualquer valor real, e a sua variância σ^2 que assume valores positivos, sendo representada por $X \rightsquigarrow N(\mu, \sigma^2)$.

A distribuição N(0,1) designada de distribuição Normal standard corresponde a uma transformação de qualquer variável normal $X \rightsquigarrow N(\mu, \sigma^2)$ para $Z \rightsquigarrow N(0,1)$. Para efectuar esta transformação considera-se:

V = a + b.X (com a e b reais) obtida por transformação linear da variável $X \rightsquigarrow N(\mu, \sigma^2)$.

O valor esperado e variância de V são definidos por:

$$\mu_v = a + b.\mu \quad (3.2)$$

$$\sigma^2{}_v = b^2.\sigma^2 \quad (3.3)$$

A variável transformada segue uma distribuição normal, ou seja:

$V \rightsquigarrow N(\mu_v, \sigma^2{}_v) \equiv N(a+b.\mu, b^2.\sigma^2)$

A distribuição normal é representada por uma curva denominada de curva normal, também conhecida como a curva em forma de sino. A história desta representação está bastante ligada a descoberta das probabilidades em matemática no século XVII que surgiram para resolver questões de aposta de jogos de azar.
A curva normal está ligada a grandes nomes como Laplace utilizando em 1783 para descrever a distribuição dos erros, Gauss que o utilizou em 1809 para analisar dados astronómicos, inclusivé actualmente é também denominada de curva de Gauss.
Embora a curva normal original seja definida apenas pela simetria, quando refere-se a curva normal, tipicamente está-se referindo a curva normal standard que é definida pela simetria e pela curtose. A vantagem da curva normal standard é que alguns parâmetros já estão definidos para qualquer escala utilizada. A média é sempre zero e a variância é sempre um.
A curtose da curva normal refere à altura do pico da curva que acontece na média da distribuição. Os nomes das curvas normais são definidos de acordo com a altura do pico, sendo chamadas de leptocúrtica, platicúrtica e mesocúrtica caso o pico seja muito elevado, achatado ou mediano respectivamente, sendo a última característica da curva normal padronizada [36].

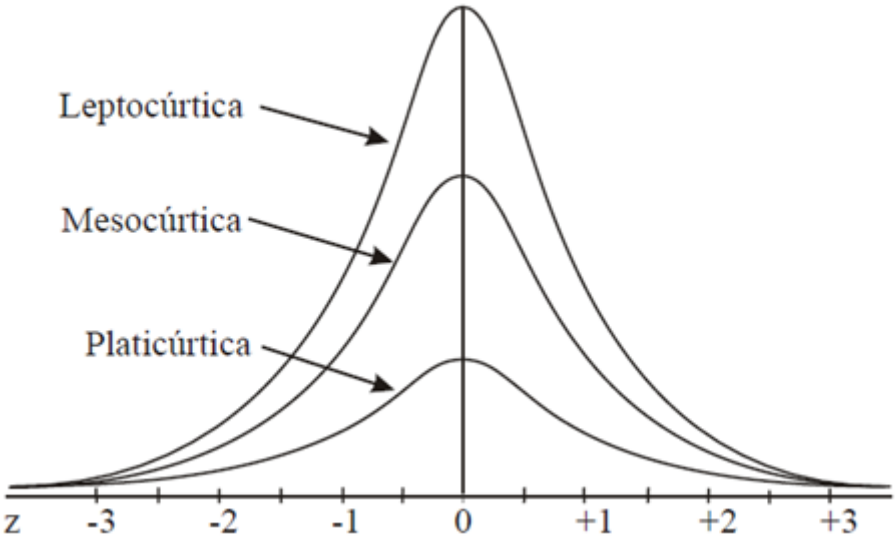

Fig. 10 Distribuições da curva normal (Fonte: [36]).

As principais características desse tipo de distribuição são:

1. A função densidade de probabilidade tem a forma de sino, simétrica em relação ao eixo x = μ e tem pontos de inflexão em x = $\mu \pm \sigma$.
2. Se a variável aleatória X tem distribuição normal X$\rightsquigarrow$N(μ, σ^2)
   Então
   E[x] = μ
   e
   V[X] = σ^2

3. Se a variável aleatória X tem distribuição normal de parâmetros $\mu\, e\, \sigma$ então Z = $\frac{X-\mu}{\sigma}$ é chamada normal estandardizada ou reduzida ou ainda normal-padrão e E[Z] = 0 e V[Z] = 1, ou seja Z ⇝N(0,1)
4. A aditividade da distribuição normal: sejam n variáveis aleatórias independentes X_i, i = 1, 2,…, n em que X_i ⇝N(μ_i,σ_i). Então a variável aleatória T=$\sum_{i=1}^{n} a_i X_i$, $a_i \in$ R, i = 1, 2,.., n, terá a seguinte distribuição

$$T = \sum_{i=1}^{n} a_i X_i \sim N\left(\sum_{i=1}^{n} a_i \mu_i\, , \sqrt{\sum_{i=1}^{n} a_i^2 \sigma_i^2}\right)$$

3.1.2 Distribuição Gama

A distribuição Gama é muito utilizada na análise de tempo de vida de equipamentos, do tempo de retorno de mercadorias com falhas, em testes de confiabilidade, em sistemas baseados em filas de esperas, na análise da carga nos servidores Web e em sistemas de gestão de risco.

Se a variável X segue uma distribuição Gama com parâmetros $\alpha\, e\, \beta$ denota-se X⇝G(α, β), e a sua função de densidade é dada por:

$$P(\chi|\alpha,\beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} \chi^{\alpha-1} e^{-\beta\chi},\ \chi > 0, \qquad (3.4)$$

Sendo os parâmetros da distribuição, que podem assumir qualquer valor positivo, definidos por: β é taxa média do processo; α o número especifico de eventos que ocorrem até que a variável χ (tamanho do segmento de tempo ou espaço) seja atingida. $\Gamma(\alpha)$é a função Gama definida por:

$$\Gamma(\alpha) = \int_0^{\infty} \chi^{\alpha-1}\, e^{-\chi}\, d\chi \quad \text{para} \quad \alpha > 0$$

Para $\alpha, \beta > 0$

$$\mu_{(x)} = {}^{\alpha}/_{\beta} \qquad (3.5)$$

$$\sigma^2{}_{(x)} = {}^{\alpha}/_{\beta^2} \qquad (3.6)$$

3.1.3 Distribuição de Poisson

A distribuição de Poisson permite descrever um conjunto de fenómenos aleatórios em que os acontecimentos se repetem no tempo ou no espaço [35].

A variável discreta número de ocorrências por unidade de tempo seguirá uma distribuição de poisson quando for possível verificar as quatro condições seguintes:

- Os números de ocorrências registadas nos intervalos da partição são independentes entre si.
- A distribuição do número de ocorrências em cada intervalo é a mesma para todos os intervalos.
- A probabilidade de se registar uma ocorrência num intervalo qualquer de dimensão $\Delta t, \Delta p_1$ é praticamente proporcional à dimensão do intervalo, ou seja,

$$\Delta p_1 \approx \lambda . \Delta t \tag{3.7}$$

- A probabilidade de registarem duas, três ou mais ocorrências num intervalo qualquer de dimensão Δt, Δp_n (n>=2) é desprezável quando comparada com a probabilidade Δp_1.

Com as condições acima mencionadas é possível estabelecer a forma da distribuição de Poisson como:

$$F(y;\mu) = \frac{\mu^y e^{-\mu}}{y!}$$
$$= \exp\left[y \ln(\mu) - \mu - \ln(y!)\right] \tag{3.8}$$

Então

$$\theta = \ln(\mu);\ \ b(\theta) = -\mu;\ \ a(\phi) = \phi;\ \ \phi = 1;\ \ c(y,\phi) = -\ln(y!)$$

O valor esperado e variância de Y são definidos por:

$$E(Y) = \frac{db(\theta)}{d\theta} = \mu \tag{3.9}$$

$$var(y) = \frac{d^2 b(\theta)}{d\theta^2} a(\phi) = \mu \tag{3.10}$$

3.1.4 Distribuição Binomial

Se a variável Y representar o número de vezes que no decurso de N experiências de Bernoulli ocorreram sucesso então Y segue uma distribuição Binomial. Define-se como distribuição de Bernoulli uma experiência aleatória com apenas duas possibilidades denominadas de "sucesso" e "insucesso" estando associadas às probabilidades p e q respectivamente. Uma distribuição Binomial possui as seguintes propriedades:

- Cada experiência corresponde apenas a um de dois resultados possíveis: "Sucesso"ou "Insucesso"
- A probabilidade de ocorrência de cada resultado mantém-se inalterada de experiência para experiência:
  p (Sucesso) = p = constante e p (insucesso) = 1-p ≡ q;
- Os resultados associados a cada experiencia são independentes.

Define-se a distribuição Binomial por:

$$f(y;\mu) = \binom{n}{y} \mu^y (1-\mu)^{n-y}$$

$$= \exp\left[y\, ln\left(\frac{\mu}{1-\mu}\right) + n \ln(1-\mu) + \ln\binom{n}{y}\right] \tag{3.11}$$

Então $\theta = ln\left(\frac{\mu}{1-\mu}\right);\ b(\theta) = n\ln(1-\mu);\ \ a(\phi) = \phi;\ \ \phi = 1;\ \ c(y,\phi) = ln\binom{n}{y}$

A Média e variância de y são definidas por:

$$\mu_{(Y)} = \frac{db(\theta)}{d\theta} = \mu \quad (3.12)$$

$$\text{var(y)} = \frac{d^2b(\theta)}{d\theta^2}\text{a}(\phi) = \mu(1-\mu) \quad (3.13)$$

3.1.5 Distribuição Normal Inversa

A distribuição Normal inversa, também denominada de distribuição de Pascal e de distribuição de Polya é uma distribuição de probabilidade discreta baseada em experiências de Bernoulli.

Considerando uma sequência de acontecimentos independentes, a distribuição de Pascal indica o número de tentativas necessárias para obter k sucessos de igual probabilidade θ ao fim de n experiências, sendo a última tentativa um sucesso.

Define-se a função de probabilidade por:

$$\text{B}(n;\mathcal{K},\theta) = \binom{n}{\mathcal{K}} \theta^{\mathcal{K}}(1-\theta)^{n-\mathcal{K}} \quad (3.14)$$

A Média e a variância são definidas por:

$$\mu = \frac{\mathcal{K}}{\theta} \quad (3.15)$$

$$\sigma^2 = \frac{\mathcal{K}(1-\theta)}{\theta^2} \quad (3.16)$$

Com a demonstração dos parâmetros de cada distribuição apresentada na apêndice A é possível verificar que a função de variância $V(\mu)$, e o parâmetro de dispersão ϕ, são definidas implicitamente a partir do momento em que é escolhida a distribuição de probabilidade. Sendo $V(\mu)$ a parte da variância da variável resposta y que depende da média e ϕ o parâmetro constante e que não depende da média para os membros da família exponencial [37]:

$\text{Var(y)} = \phi\, V(\mu)$ (3.17)

3.2 Modelos Lineares Generalizados

Os Modelos Lineares Generalizados referidos como MLG ou GLM foram introduzidos por Nelder e Wedderbun em 1972 [34], constituindo uma extensão dos modelos lineares de regressão múltipla. Têm como principal objectivo estudar as relações existentes entre variáveis, ou seja, estudar a influência que uma ou mais variáveis (explicativas) tem sobre uma variável alvo de estudo denominada variável resposta [34].

O aparecimento desta nova abordagem permitiu alterar as hipóteses admitidas nos modelos de regressão múltipla apresentados até a altura: A variável resposta do modelo passou a ser proveniente de um universo que segue uma lei da distribuição da família exponencial em vez de ter obrigatoriamente uma distribuição normal. Além disso, nos

modelos lineares de regressão múltipla a relação entre o valor médio da variável resposta e a combinação linear das variáveis explicativas é a função identidade. Nestas aquela relação pode ser estabelecida por qualquer função monótona diferenciável.
O seu uso e estudo foram atrasados devido ao acesso limitado a conteúdos bibliográficos e à complexidade do GLIM, que foi o primeiro software adequado para a aplicação desta nova abordagem. Depois de vinte anos da sua apresentação passou a ser de domínio público e actualmente grande parte dos pacotes estatísticos contem módulos adequados ao estudo dos modelos lineares Generalizados [34].
A modelação de dados através de MLG passa por três etapas [34]: formação, ajuste, selecção e validação do modelo.
Na formulação é preciso ter em consideração a escolha da distribuição da variável resposta, a escolha das co-variáveis, formulação apropriada da matriz de especificação e escolha da função de ligação. Para escolher a distribuição apropriada para a variável resposta é fundamental fazer uma análise preliminar dos dados. Por vezes é necessário transformá-los para que seja possível escolher uma família de distribuição adequada.
A escolha das co-variáveis e da formulação adequada da matriz de especificação centra-se na codificação apropriada das variáveis de forma que traduza o problema em estudo.
A escolha da função de ligação compatível com a distribuição deve ser resultado da combinação das considerações à priori do problema em estudo e da análise dos dados.
A adequabilidade da função de ligação pode ser verificada através da representação gráfica da variável dependente ajustada. Se os pontos se distribuírem aproximadamente sobre uma linha recta então a função de ligação é adequada. Se houver uma curvatura para cima ou para baixo é sinónimo de que é preciso usar uma função de ligação com potência superior ou inferior.
O ajustamento do modelo consiste em estimar os parâmetros do modelo, os coeficientes $\beta\,'s$ associados às co-variáveis e o parâmetro de dispersão ϕ.
Na selecção e validação do modelo pretende-se encontrar sub-modelos com um número moderado de parâmetros que sejam adequados para os dados e identificar discrepâncias entre os dados e os valores previstos.
A aplicação do MLG a um conjunto de dados resulta na inferência sobre o modelo. Essa inferência é baseada essencialmente na verosimilhança. Os estimadores de máxima verosimilhança são obtidos através de soluções da equação de verosimilhança

$$\sum_{i=1}^{n} \frac{(y_i - \mu_i) z_{ij}\, \partial \mu_i}{var(Y_i)\, \partial \eta_i} = 0 \quad \text{j=1, …,p} \tag{3.18}$$

cujo a sua solução não corresponde necessariamente a um maximo global da função.
Este método consiste em maximizar a função de verossimilhança de $\vec{Y}$ em ordem a β, ou seja, determinar o máximo absoluto do logaritmo da função de verossimilhança visto que o logaritmo é estritamente crescente.
Seja $L(\vec{\theta}, \phi)$ o logaritmo da função de verossimilhança para o vector $\vec{Y}$
Então

$$L(\vec{\theta}, \phi) = \sum_{i=1}^{N} \left[\frac{Y_i \theta_i - b(\theta_i)}{a_i(\phi)} + c(Y_i, \phi) \right] \tag{3.19}$$

Sendo

$$\mathrm{E}[Y_i] = \mu_i = \frac{d\, b(\theta_i)}{d\, \theta_i} \tag{3.20}$$

e

$$g(\mu_i) = \sum_{j=1}^{p} x_{ij}\, \beta_j = \eta_i \tag{3.21}$$

sendo g(.) uma função monótona e diferenciável.
Para maximar a função $L(\vec{\theta}, \phi)$ há que resolver o sistema constituido pelas equações

$$\frac{\partial\, L(\vec{\theta},\phi)}{\partial \beta_j} = 0 \quad j=1,2,\ldots,p$$

Este sistema de equações é resolvido por métodos numéricos iterativos. No caso de modelos lineares generalizados é resolvido pelo método de Fisher. A aproximação obtida por este método depende da matriz de informação de Fisher, que consiste no valor esperado da matriz Hessiana que é definida por

$$H=\left[\frac{\partial(\frac{\partial L}{\partial \beta_1}, \frac{\partial L}{\partial \beta_2}, \ldots, \frac{\partial L}{\partial \beta_p})}{\partial(\beta_1, \beta_2, \ldots, \beta_p)}\right] \tag{3.22}$$

Assim a (s+1)-ésima aproximação para o máximo de $L(\vec{\theta}, \phi)$ é dada por

$$\vec{\beta}^{(S+1)} = \vec{\beta}^{(S)} - \mathrm{E}[H]^{-1}_{\vec{\beta}^{(s)}} \left[\frac{\partial L}{\partial \vec{\beta}}\right]_{\vec{\beta}^{(s)}} \tag{3.23}$$

O método dos scores de fisher reduz ao método de Newton-Raphson. Ou seja, a i-ésima aproximação é dada por

$$\vec{\beta}^{(S+1)} = \vec{\beta}^{(S)} - [H]^{-1}_{\vec{\beta}^{(s)}} \left[\frac{\partial L}{\partial \vec{\beta}}\right]_{\vec{\beta}^{(s)}} \tag{3.24}$$

Nesta situação o algoritmo goza das propriedades dos dois métodos. O método de Newton-Raphson tem convergência rápida e é auto-correctivo e o método de *scores* de Fisher é mais adequado nas primeiras iterações para sistemas com grande número de iterações.
O cálculo de máxima verossimilhança de β processa iterativamente em dois passos:
Dado $\hat{\beta}^{(k)}$, com k a iniciar em zero, calcula-se u^k, na nova iteração calcula-se $\hat{\beta}^{(k+1)}$ e o processo é repetido iterativamente até atingir a condição de paragem

$$\frac{||\hat{\beta}^{(k+1)} - \hat{\beta}^{(k)}||}{||\ \hat{\beta}^{(k)}\ ||} \leq \epsilon \tag{3.25}$$

que é pré-definido para algum valor $\epsilon > 0$.
Regra geral, a convergência é atingida ao fim de algumas iterações. Por vezes isso não acontece devido a má estimativa inicial dos parâmetros ou devido a inexistência de estimador de máxima verossimilhança dentro dos limites dos valores admissíveis para o vector β [34].

3.2.1 Formulação dos Modelos Lineares Generalizados

Uma formulação de modelos lineares generalizados pode ser caracterizada pelos seguintes pontos:

- Variável aleatória Y = $\{y_1, y_2, \dots, y_n\}$, chamada de variável resposta ou dependente que pode ser contínua, discreta ou dicotómica com médias $\mu_1, \mu_2, \dots \mu_n$;
- A distribuição de probabilidade de y_i é um dos membros da família exponencial. O facto das distribuições admissíveis num MLG pertencerem à família exponencial e a exigência da independência entre as variáveis constituem limitações para o modelo. Apesar destas limitações existentes, o modelo tem vindo a desenvolver um papel de capital importância, sobretudo na análise estatística;
- A distribuição de todas as variáveis aleatórias Y_i é da mesma forma;
- As variáveis de regressão que são representadas pelo vector X= $(x_1, .., x_k)^T$, de k variáveis explicativas, também designadas de co-variáveis ou variáveis independentes, que explica parte da variabilidade inerente a Y, que podem ser contínuas, discretas, qualitativas de natureza ordinal ou dicotómicas.
- O modelo é constituído com um regressor linear

$$\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k \tag{3.26}$$

- A função de ligação g(μ_i) que faz a ligação entre a média e a regressão linear definindo assim a forma como os efeitos sistemáticos de $x_1, x_2, \dots, x_n$ são transmitidos para a média

$$\eta_i = g(\mu_i) = (\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k) \tag{3.27}$$

Quando $\eta = \theta$ a função de ligação é denominada canónica e garante que os valores obtidos para μ são sempre admissíveis [34].
Alguns autores como Myres e Montgomery (2002) recomendam a utilização da função de ligação canónica pelo facto de dispor de algumas propriedades interessantes como o de simplificar as estimativas de máxima verossimilhança dos parâmetros do modelo e cálculo do intervalo de confiança para a média da resposta. Outros autores afirmam não haver nenhuma razão à partida para escolher a ligação canónica e que as propriedades mencionadas acima não implicam necessariamente qualidade de ajuste do modelo não traduzindo assim numa obrigatoriedade da sua utilização e que nem sempre se obtêm os melhores resultados com a ligação canónica [34].
A tabela seguinte apresenta as ligações da família exponencial.

Tabela 2- Ligações canónicas da família exponencial.

Distribuição	Ligação Canónica
Normal	$\eta = \mu$
Poisson	$\eta = \ln \mu$
Binomial	$\eta = \ln(\pi/(1-\pi))$
Gama	$\eta = 1/\mu$
Normal Inversa	$\eta = 1/\mu^2$

Como alternativa à função de ligação canónica é possível definir uma família de funções de ligação de potência

$\eta = \mu^{\lambda}$ para $\lambda \neq 0$ e

$\eta = ln\, \mu$ para $\lambda = 0$.

Capítulo 4

Implementação

Neste Capítulo iremos descrever duas abordagens baseadas nos modelos lineares generalizados descritos no Capítulo anterior e aqui contextualizados no problema da filtragem colaborativa.

O trabalho desta tese centra-se essencialmente no artigo de Dellany e Verleysen [33] onde é proposto pelos autores a factorização da matriz de ranking, natural a um problema de filtragem colaborativa, baseando-se nos modelos lineares generalizados. A abordagem consiste essencialmente na optimização alternada das variáveis latentes respeitantes aos utilizadores e itens até que um determinado nível de precisão seja atingido.

Na Secção 4.2 apresentamos uma nova metodologia como uma extensão da proposta em [33]. Esta abordagem entra com informação mais rica tanto ao nível dos utilizadores como dos itens. Por exemplo, para a base de dados *MovieLens* a metodologia proposta considera a informação demográfica do utilizador, idade e género enquanto nos itens considera essencialmente o tipo de filme.

4.1 Collaborative Filtering with Interlaced Generalized Linear Models

Collaborative Filtering with interlaced generalized linear models apresentado por Nicolas Delannay e Michel Verleysen é uma abordagem de filtragem colaborativa que procura identificar interesses dos utilizadores através da similaridade dos seus comportamentos. Tem como propósito recomendar ao utilizador itens com maior valor previsto. Para esse fim centra-se na previsão de novas avaliações baseada na factorização da matriz das avaliações e na utilização de modelos probabilísticos para representar incertezas nas avaliações. Esse Modelo tem a vantagem de permitir utilizar diferentes configurações para representar intuições a cerca do processo de avaliação tendo a facilidade de testá-la mantendo o mesmo processo de aprendizagem.

Uma outra vantagem desse modelo é que não é preciso requisitos de memória muito elevados para aplica-lo a grandes bases de dados e apresenta desempenho comparável com as outras abordagens apresentadas no estado da arte [33].

4.1.1 Notação

Ao longo desta Secção iremos seguir a seguinte notação: os conjuntos dos itens serão denotados por Y= $\{y_m\}_{m=1}^{M}$ e o conjunto dos utilizadores registados por U =$\{U_n\}_{n=1}^{N}$. A matriz das avaliações R tem dimensão N x M. Todas as avaliações são registadas na matriz de avaliações onde cada linha corresponde a um utilizador e cada coluna corresponde a um item pertencente ao sistema. A avaliação expressa pelo utilizador u_n para o item y_m é representada por r_{nm}. As avaliações observadas são guardadas na lista D= $\{(u, y, r)_l\}_{l=1}^{L}$.

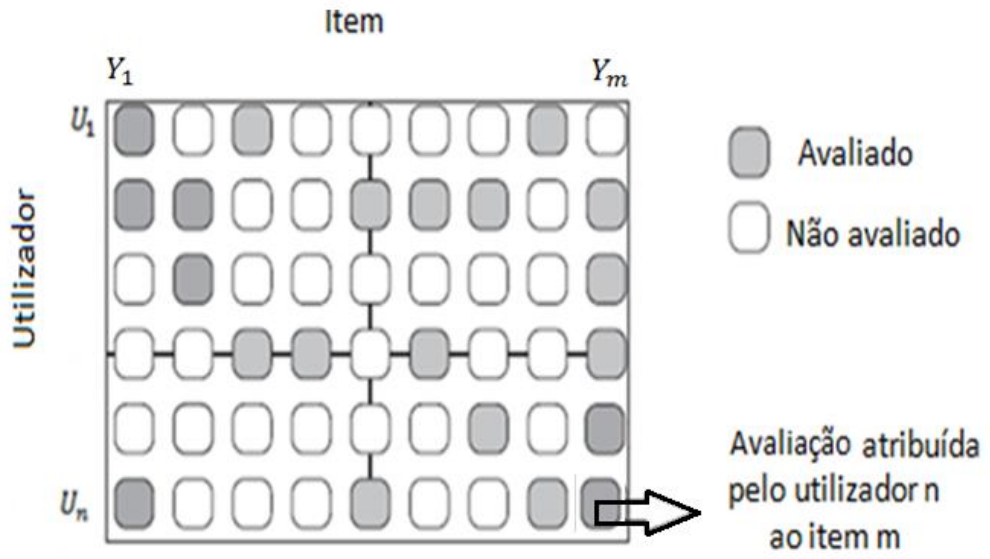

Fig. 11 Ilustração da Matriz de Avaliações (adaptado de [33]).

4.1.2 Descrição do Modelo

A representação dos comportamentos dos utilizadores em relação aos itens através duma avaliação é um passo fulcral na concepção de modelos de filtragem colaborativa [33]. Os modelos lineares entrelaçados assentam na ideia de que os utilizadores e os itens são representados como um vector de características que representam os parâmetros principais do modelo. A cada utilizador u_n é associado um vector de características $\emptyset_n \in R^k$ e simetricamente a cada item y_m é associado o vector de características $\omega_m \in R^k$. Concebida uma representação para os utilizadores e para os itens, a previsão da avaliação atribuída por um utilizador a um determinado item é representada como o produto entre os respectivos vectores de características: a apreciação do item y_m pelo utilizador u_n é avaliada como

$$\eta_{nm} = \emptyset_n^T \omega_m \tag{4.1}$$

Com esta representação o vector de características do utilizador é entendido como diferentes sensibilidades a um conjunto de aspectos que descrevem os itens e o vector de características dos itens corresponde a esses aspectos. A apreciação estimada advêm da soma de contribuições positivas e/ou negativas de todos esses aspectos.

Após a construção do modelo é preciso transformar a apreciação numa distribuição de probabilidades. Para este propósito o formalismo GLM é aplicado. A principal dificuldade da filtragem colaborativa com GLM entrelaçado é que $\emptyset$ e $\boldsymbol{\omega}$ são parâmetros a serem optimizados ao contrário do GLM standard onde apenas um dos vectores é considerado parâmetro. Por essa razão o modelo é denominado de GLM entrelaçado $\Phi\Omega$. Representando a matriz de características $\Phi = [\emptyset_1, \dots, \emptyset_N]^T$ e $\Omega = [\omega_1, \dots, \omega_M]^T$, a estimativa da média das avaliações é avaliada por

$$\overline{R} = g(\Phi\Omega^T). \tag{4.2}$$

Seguindo o procedimento usual em modelos probabilísticos, o modelo é ajustado baseado no critério de Log máxima verossimilhança. O ajuste da matriz de características é expressa por:

$$\begin{aligned}\{\widehat{\Phi},\widehat{\Omega}\} &= \underset{\{\Phi,\Omega\}}{\operatorname{argmin}}\{-\log P(R|\bar{R};\psi)\} \\ &= \underset{\{\Phi,\Omega\}}{\operatorname{argmin}}\left\{-\sum_{l}\log P(r_l|g(\phi_{n_l}^{T}\omega_{m_l});\psi)\right\}\end{aligned} \tag{4.3}$$

O somatório é efectuado apenas para as avaliações efectuadas visto que as não efectuadas não têm nenhuma contribuição para a máxima verosimilhança, reduzindo assim a ordem de complexidade quando aplicado a sistemas de recomendação com grande quantidade de dados.
As avaliações não efectuadas não são consideradas na concepção deste modelo, mas estas podem ser informativas sobre o tipo de item que o utilizador tem interesse [38].
Para que o modelo generalize bem as avaliações não efectuadas é preciso ajustar a sua flexibilidade. Para isso existem duas abordagens: seleccionar o parâmetro estrutural definindo a dimensionalidade do espaço de parâmetros, neste caso, a dimensionalidade das características K, ou ajustar a flexibilidade pelas médias dos termos de regularização. Como os termos de regularização são funções contínuas dos hiper-parâmetros, em geral fornecem maior controlo no ajuste do modelo. Utilizando uma abordagem probabilística é possível definir o termo de regularização através duma distribuição prévia nos parâmetros e dum critério de aprendizagem:

$$\{\widehat{\Phi},\widehat{\Omega}\} = \underset{\{\Phi,\Omega\}}{\operatorname{argmin}}\{-\log P(R|\bar{R};\psi) - \log P(\{\phi_n\},\{\omega_m\}|\alpha)\} \tag{4.4}$$

onde α é um conjunto de hiper-parâmetros da distribuição considerada. Uma possibilidade seria regularizar as características com distribuições Gaussianas

$$\mathrm{P}(\{\phi_n\},\{\omega_m\}|\,\alpha) = \prod_n N(\phi_n|0,\alpha^u I_k)\prod_m N(\omega_m|0,\alpha^y I_k) \tag{4.5}$$

Onde I_k é a matriz identidade K x K e α^u e α^y são parâmetros de precisão da distribuição Gaussiana.

4.1.3 Escolha da Configuração

A suposição da distribuição com ruído Gaussiano é comum em problemas de regressão. Ajustar o modelo do ruído Gaussiano é equivalente a optimizar o critério dos mínimos quadrados. Além disso, se a função de ligação utilizada for a identidade $\mu = \eta$ e se não existir os termos de regularização o GLM entrelaçado é equivalente a minimização da norma *Frobenius* $||\mathrm{R} - \Phi\Omega^{T}||_F$ para a dimensionalidade das características dadas.

Neste caso a minimização da norma Frobenius daria o mesmo resultado independentemente se fosse utilizado o método de factorização SVD ou se fosse utilizado o método da utilização de abordagens probabilísticas de aprendizagem. Porém, o uso de abordagem probabilística possui a vantagem de permitir testar diferentes representações de características e parâmetros de regularização sem alterar a técnica geral de optimização.
A escolha do ruído Gaussiano com a função de ligação identidade não é a única escolha aceitável. Existem diversas razões que levam a sugerir o uso de outros modelos de ruído: primeiro não é realmente satisfatório representar a distribuição através duma sequência limitada de avaliações com uma Gaussiana sem limites. É mais apropriado trabalhar com uma distribuição limitada como a distribuição Beta para avaliações contínuas em que os dados originais são normalizados para o intervalo [0,1], ou o uso da binomial para avaliações discretas r ϵ[0,…, r_{max}] pela fórmula:

$$\mathfrak{B}n(r|\mu;\, r_{max}) = \frac{r_{max}!}{r!(r_{max}-r)!}\left(\frac{\mu}{r_{max}}\right)^{r}\left(1-\frac{\mu}{r_{max}}\right)^{r_{max}-r} \tag{4.6}$$

Para usar esta distribuição binomial apenas será necessário escalar e normalizar as avaliações originais. Adicionalmente deve ser utilizada uma função de ligação saturada para evitar previsões fora do âmbito das avaliações. Isto pode ser feito por exemplo com a função de ligação logística:

$$\frac{\mu}{r_{max}} = \frac{1}{1+\exp(-\eta)} \tag{4.7}$$

Outra razão pela qual o modelo de ruído Gaussiano e o critério dos mínimos quadrados não é a melhor escolha é que este modelo é sensível a avaliações atípicas.
As avaliações atípicas acontecem quando utilizadores tentam induzir o modelo de filtragem colaborativa a erro atribuindo avaliações aleatórias. Particularmente quando é atribuído avaliações muito elevadas ou muito baixas aleatoriamente estas podem ter grande influência na estimativa das características mas não contribuem na identificação de padrões na matriz das avaliações. Na realidade a distribuição binomial é também sensível a avaliações atípicas. É possível usar modelos de ruído mais robustos como *t-student* ou a distribuição normal P(r|μ;v) α $\beta n(r|u, r_{max})^{v}$. Infelizmente, esta distribuição não é membro da família exponencial e a sua optimização é mais exigente. Variando a distribuição é possível introduzir considerações intuitivas no modelo. Pode ser boa ideia colocar a restrição de que as características dos utilizadores devem ser positivas, forçando-os a estarem associados ao conceito de sensibilidades a diferentes aspectos. Essa restrição pode ser imposta usando a seguinte exponencial para representar as características dos utilizadores:

$$P(\{\phi_n\}|\alpha) = \prod_n \prod_d Exp(\phi_{nd}\,|\alpha^u) \tag{4.8}$$

Onde a distribuição exponencial é definida por Exp(∅|α^u) = α^uexp(-α^u∅). Se for utilizado a distribuição Gama há maior controlo na função de distribuição. Até agora a parametrização do ruído e a distribuição eram guardadas na forma mais concisa possível

impondo uma dispersão comum ψ para todas as avaliações e os parâmetros α^u e α^y para todas as características dos utilizadores e dos itens respectivamente. No entanto, os perfis dos utilizadores apresentam diversidades: alguns utilizadores são muito previsíveis, outros nem por isso e o mesmo pode ser dito dos itens.

4.1.4 Optimização do Modelo

Existem dois níveis na optimização dos GLM entrelaçados. O nível interior que corresponde a ajustar as características Φ e Ω e o segundo nível que corresponde aos hiper-parâmetros, nomeadamente o numero de características K, o parâmetro de dispersão ψ e o parâmetro da distribuição α.

Considerando que os hiper-parâmetros são fixos, a forma mais simples de optimizar as características é actualizar um vector de características de cada vez. A optimização do erro de regularização (4.4) relativamente a ω_m e Φ_n são respectivamente:

$$\widehat{\omega}_m = \underset{\omega_m}{\operatorname{argmin}} \left\{ -\sum_{l \epsilon\, L_m^y} \log P(r_l | g(\phi_{nl}^T \omega_m), \psi) - \log P(\omega_m | \alpha^y) \right\} \tag{4.9}$$

e

$$\widehat{\phi}_n = \underset{\phi_n}{\operatorname{argmin}} \left\{ -\sum_{l \epsilon\, L_n^u} \log P\left(r_l \middle| g\left(\phi_n^T \omega_{m_l}\right), \psi\right) - \log P(\phi_n | \alpha^u) \right\} \tag{4.10}$$

onde L_m^y é o conjunto de índices associados às avaliações atribuídas ao item y_m e L_n^u é o conjunto de índices associadas às avaliações do utilizador u_n. Cada actualização do vector de características é um problema comum de regressão tipicamente com menos de algumas centenas de pares de aprendizagem $(\emptyset_{nl}, r_l)$ (Eq. (4.9)) ou (ω_{ml}, r_l) (Eq. (4.10)). Encontrar um óptimo local deste critério é rápido especialmente com distribuições da família exponencial. Pode ser encontrado aplicando o método dos mínimos quadrado reponderado ou qualquer outro algoritmo de gradiente descendente. As expressões gerais para o gradiente e a Hessiana para GLM são:

Gradiente do vector de características dos itens:

$$\nabla_{\omega_m} \mathcal{L}(\mathcal{R}_m^y) = \sum_{l \epsilon L_m^y} (r_l - g(\phi_{n_l}^T \omega_m)) \frac{\dot{g}(\phi_{n_l}^T \omega_m)}{v(g(\phi_{n_l}^T \omega_m))} \phi_{n_l} \tag{4.11}$$

Hessiana do vector de características dos itens:

$$\mathcal{H}_{\omega m}\mathcal{L}\left(\mathcal{R}_m^y\right)=\sum_{l\epsilon L_m^y}\left[\left(\frac{\dot{g}(\phi_{nl}^T\omega_m)^2}{v\left(g(\phi_{nl}^T\omega_m)\right)}-\right.\right.$$
$$\left.\left.\left(r_l-g(\phi_{nl}^T\omega_m)\right)x\ \frac{\ddot{g}(\phi_{nl}^T\omega_m)v\left(g(\phi_{nl}^T\omega_m)\right)-\dot{g}(\phi_{nl}^T\omega_m)^2\dot{v}\left(g(\phi_{nl}^T\omega_m)\right)}{v(g(\phi_{nl}^T\omega_m))^2}\right)x\ \phi_{nl}\phi_{n_l}^T\right] \quad (4.12)$$

Gradiente do vector de características dos utilizadores:

$$\nabla_{\phi_n}\mathcal{L}(\mathcal{R}_n^u)=\sum_{l\epsilon L_n^u}(r_l-\mathrm{g}(\phi_n^T\omega_{m_l}))\frac{\dot{g}(\phi_n^T\omega_{m_l})}{v(g(\phi_n^T\omega_{m_l}))}\omega_{m_l} \quad (4.13)$$

Hessiana do vector de características dos utilizadores:

$$\mathcal{H}_{\phi_n}\mathcal{L}(\mathcal{R}_n^u)=$$
$$\sum_{l\epsilon L_n^u}\left[\left(\frac{\dot{g}(\phi_n^T\omega_{m_l})^2}{v\left(g\left(\phi_n^T\omega_{m_l}\right)\right)}-\left(r_l-g\left(\phi_n^T\omega_{m_l}\right)\right)\ x\frac{\ddot{g}\left(\phi_n^T\omega_{m_l}\right)v\left(g\left(\phi_n^T\omega_{m_l}\right)\right)-\dot{g}\left(\phi_n^T\omega_{m_l}\right)^2\dot{v}(g\left(\phi_n^T\omega_{m_l}\right))}{v(g(\phi_n^T\omega_{m_l}))^2}\right)x\ \omega_{m_l}\omega_{m_l}^T\right]$$

(4.14)

Todos os vectores ω_m e $\emptyset_n$ são actualizados e é aplicado o procedimento completo até convergir.

No caso da configuração com o modelo do ruído Gaussiano e função de ligação identidade, cada actualização do vector de características (Eq. 4.9 e Eq. 4.10) corresponde à resolução dum problema de regressão dos mínimos quadrados. Resolver o problema dos mínimos quadrados para ω_m requer o cálculo e inversão da matriz $\sum_{l\epsilon L_m^y}\emptyset_{nl}\emptyset_{nl}^T$ e a resolução equivalente para ϕ_n requer o cálculo e a inversão de da matriz $\sum_{l\epsilon L_n^u}\omega_{ml}\omega_{ml}^T$. O cálculo desta matriz é de ordem O (O_mK^2) onde O_m é o número de avaliações existentes para o item y_m (o numero de avaliações em L_m^y) e O (o_nk^2) operações onde o_n é o número de avaliações efectuadas pelo utilizador u_n (o número de avaliações em L_n^u). Por outro lado a inversão desta matriz é de ordem O(K^3). Consequentemente a ordem total de complexidade da actualização completa das matrizes de características dos utilizadores e dos itens é O(2LK^2+ (N+M)K^3) onde L é o numero de avaliações observadas (a origem do factor 2 advêm da contribuição de cada avaliação para a actualização do vector de características de um item).

O procedimento é aplicável a grandes bases de dados visto que tem poucos requisitos de memória. O processo é paralelizável visto que a actualização das características de um determinado item é independente da actualização das características dos outros itens [33]. Como alternativa a este método pode-se utilizar métodos baseados em treino incremental.

Na equação da distribuição Gaussiana identifica-se dois hiper-parâmetros: α^n e α^y. Como as características dos utilizadores e dos itens interagem multiplicativamente é suficiente fixar a precisão dos itens ou dos utilizadores a um valor arbitrário e ajustar apenas a precisão do outro conjunto. Outro ponto a considerar é que o hiper-parâmetro da dispersão ψ geralmente tem dupla influência na precisão dos hiper-parâmetros duma visão não probabilística. Assim, a suposição de que uma pequena dispersão pode ser compensada com a distribuição de probabilidade (com baixa precisão de α) é aplicável apenas para ruído Gaussiano e distribuição Gaussiana. É conveniente fixar ψ com uma estimativa da dispersão do ruído. A variância da avaliação é um ponto sensível e pode ser melhorado depois da construção do modelo. Com este procedimento existe apenas um parâmetro a ser ajustado pela técnica da reamostragem. É razoavelmente conhecido que a distribuição Gaussiana é intimamente relacionada com a regularização da formulação não probabilística. Este procedimento é mais próximo da regularização pragmática do que a Framework Bayesiana.

4.1.5 Experiências

Para avaliar o desempenho do GLM entrelaçados na previsão de avaliações não efectuadas foram concebidas diferentes configurações de GLM entrelaçado.

4.1.5.1 Procedimento de Avaliação

O processo de avaliação utilizado neste modelo foi proposto em [39,40]. Os modelos são avaliados com validação cruzada em 4-fold. O conjunto dos utilizadores é dividido aleatoriamente em quatro conjuntos contendo o mesmo número dos utilizadores. Para cada *fold* no modelo de GLM entrelaçado existe $\emptyset_{treino}$ e $\emptyset_{teste}$, correspondendo à matriz contendo as características dos utilizadores no conjunto de treino (três dos quatro conjuntos de utilizadores) e ao conjunto de teste respectivamente. Feita esta divisão o modelo é construído nas avaliações dos utilizadores no conjunto de treino, retornando uma estimativa de $\emptyset_{treino}$ e Ω. Durante a fase de teste é seleccionado algumas avaliações dos utilizadores no conjunto de teste que serão utilizadas para avaliar o desempenho. A matriz de características $\emptyset_{teste}$ é optimizada com base no resto das avaliações. Como Ω não é modificado requer uma única execução sobre cada vector de características dos utilizadores de teste.
A figura seguinte ilustra a aplicação de 4-fold:

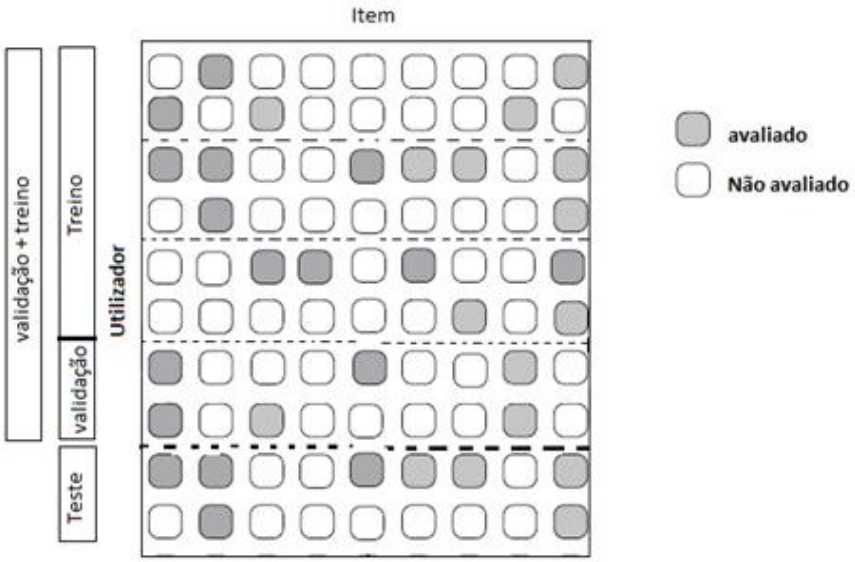

Fig. 12 Aplicação de 4-fold à Matriz de Avaliações (adaptado de [33]).

Uma vez que as características dos utilizadores de teste foram estimadas, é possível fazer previsões com as avaliações de teste. O desempenho é avaliado com base em duas funções *standards*: MAE (*Mean Absolute Error*) e RMSE (*Root Mean Square Error*) e é resumido pela média das estimativas nos 4- folds.

Para seleccionar as avaliações de teste a considerar na avaliação do desempenho foi considerado o método *given-*L_{test}. Este método consiste em considerar apenas um sub-conjunto de L_{test} avaliações por cada utilizador do conjunto de teste. Concentra na dependência entre a previsão de desempenho e o número de avaliações consideradas por utilizador.

4.1.5.2 Configuração do Modelo

Na implementação do modelo foi considerado a configuração "*Common preference*".

Nessa configuração K=2 e é imposta a restrição de que $\emptyset_{n1}$=1 e ω_{m2}=1. Não existe nenhuma restrição para $\emptyset_{n2}$ e ω_{m1}. Dessa forma não existe nenhuma interacção directa entre as características dos utilizadores e dos itens. As previsões são compostas pela média das avaliações de um item somado à tendência do utilizador em desviar dessa média. Consequentemente todos os utilizadores têm a mesma ordem de preferência sobre os itens. O modelo utiliza função de ligação identidade e ruído Gaussiano. O uso de função de ligação identidade e ruído Gaussiano permite prever avaliações fora dos limites. Estes são descartados por não terem nenhuma contribuição na avaliação do desempenho do modelo. A matriz das avaliações é dividida reservando 20% dos dados para teste e 80% para treino/validação. No conjunto de treino/validação é aplicado 4-fold.

As matrizes de características foram inicializadas com valores gerados aleatoriamente.

$$\phi_n = \begin{bmatrix} 1 & \phi_{1,2} \\ 1 & \phi_{2,2} \\ \vdots & \vdots \\ 1 & \phi_{n,2} \end{bmatrix} \qquad \omega_m = \begin{bmatrix} \omega_{1,1} & 1 \\ \omega_{2,1} & 1 \\ \vdots & \vdots \\ \omega_{m,1} & 1 \end{bmatrix} \qquad \phi_{n,2} \text{ e } \omega_{m,1} \text{ são gerados aleatoriamente}$$

Fig. 13 Representação dos vectores de características para a configuração "*Common preference*".

4.1.5.3 Base de dados

Para efeito de teste foi utilizado a base de dados da *MovieLens*.

Esta base de dados é disponibilizada pela *GroupLens* da Universidade de Minnesota. Contem 6040 utilizadores, 3900 filmes (os itens) e aproximadamente 1 milhão de avaliações discretas. As avaliações atribuídas aos filmes pelos utilizadores pertencem ao intervalo discreto [1, 5] e cada utilizador avaliou no mínimo 20 filmes.

4.2. Filtragem Colaborativa Baseada na Média das Avaliações

Filtragem colaborativa baseada na média das avaliações é um novo modelo híbrido apresentado nesta dissertação que baseia em técnicas de regressão para prever novas avaliações.

4.2.1 Descrição do Modelo

O modelo Filtragem colaborativa baseada na média das avaliações é uma nova abordagem de filtragem colaborativa para prever avaliações. Neste modelo mantêm-se a ideia chave do GLM entrelaçado que é representar os utilizadores e os itens por vector de características, associando a cada utilizador um vector de características $\phi_n \in R^k$ e a cada item o vector de características $\omega_m \in R^k$.

A apreciação dos itens expressa pelos utilizadores é estimada pelo produto dos respectivos vectores de características $\eta_{nm} = \emptyset_n^T \omega_m$.

Os vectores de características são optimizados aplicando a técnica de regressão linear regularizada. Com a aplicação da regressão linear pretende-se estudar a dependência entre as avaliações atribuídas por um utilizador a um determinado item e as características dos utilizadores e dos itens, ou seja, como é que as características dos utilizadores e dos itens influenciam no processo de avaliação.

O método consiste maioritariamente em duas fases: na primeira é analisada a influência das características do utilizador nas avaliações por ele atribuída, resultando na optimização do vector de características dos utilizadores. Na segunda fase é analisada a influência das características de um determinado item nas avaliações atribuídas a este e resulta na optimização do vector de características dos itens. Para analisar a afluência das características do utilizador nas avaliações por ele atribuídas define-se o modelo de regressão por:

X $=(x_1, x_2, \dots, x_n)^T$ a matriz de características dos utilizadores, sendo x_n o vector de características associado ao utilizador n;

$\bar{Y} = (\bar{Y}_1, \bar{Y}_2, \dots, \bar{Y}_m)^T$ o vector com a média das avaliações efectuadas pelos utilizadores, sendo $\bar{Y}_m$ a média das avaliações efectuadas pelo utilizador m.

Define-se a função linear da regressão por:

$$\mu_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_n x_{in}$$

β_j é denominado de coeficiente de regressão e é o único parâmetro desconhecido do modelo.
Em notação matricial vem

$$\mu = X\beta \tag{4.15}$$

O coeficiente de regressão é definido como [41]:

$$\hat{\beta} = (\mathrm{X^T X} + \lambda \mathrm{I})^{-1} \mathrm{X^T \overline{Y}} \tag{4.16}$$

Sendo λ o parâmetro de regularização do modelo.
É realizado o mesmo estudo relativamente às características dos itens definindo:
X =$(x_1, x_2, \dots, x_m)^T$ a matriz de características dos itens, sendo x_m o vector de características associado ao item m. $\bar{Y} = (\bar{Y}_1, \bar{Y}_2, \dots, \bar{Y}_n)^T$ o vector com a média das avaliações atribuídas aos itens, sendo $\bar{Y}_n$ a média das avaliações atribuídas ao item n.
A optimização do vector de características dos itens é efectuada por um processo análogo ao utilizado para optimizar o vector de características dos utilizadores. Após a optimização dos vectores de características é estimada a avaliação atribuída por um determinado utilizador a um item aplicando (4.1). A figura seguinte ilustra a construção do modelo:

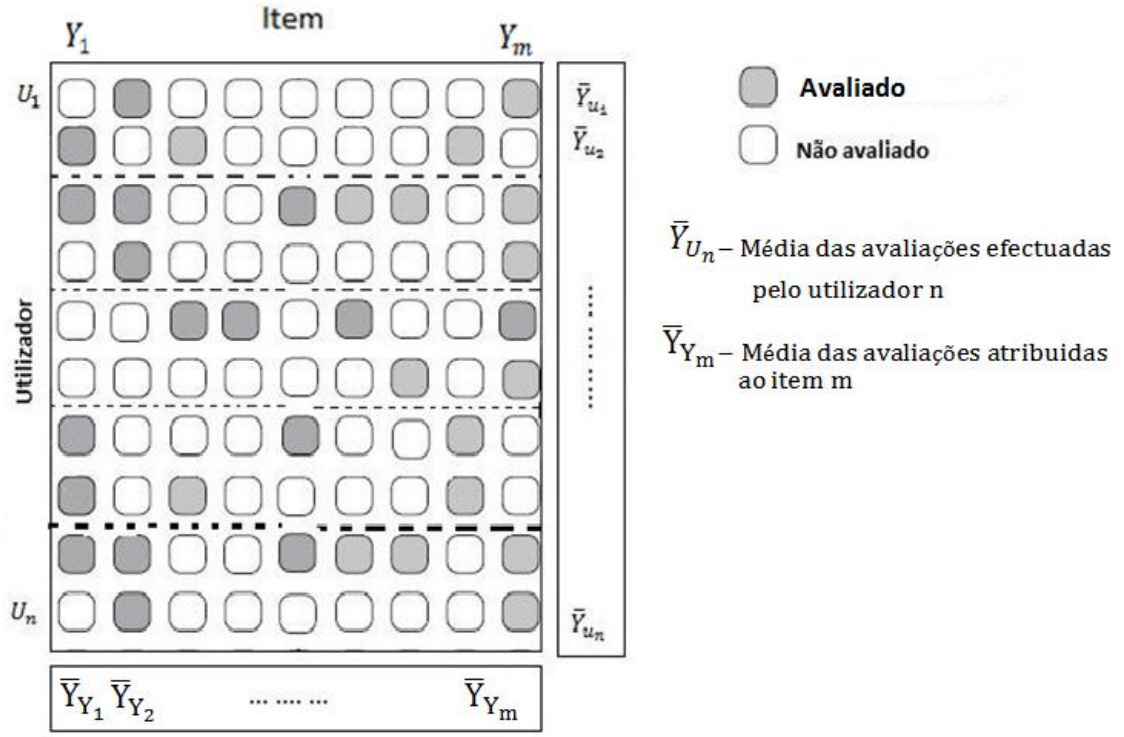

Fig. 14 Ilustração do Modelo (adaptado de [33]).

A aplicação da regressão linear para optimizar os vectores de características assume a existência de uma relação entre a variável Y, avaliação atribuída a um item, e a variável independente característica do utilizador.

4.2.2 Configuração do Modelo

Na configuração do modelo definiu-se K=2 e impôs-se a restrição de $\emptyset_{n1}$=1 e ω_{m2}=1, $\emptyset_{n2}$ e ω_{m1} são dados demográficos dos utilizadores e dos itens como idade, género (masculino, feminino) e tipo de filme. Esta abordagem foi motivada pela inicialmente proposta por Dellany e Verleysen [33] onde a inclusão desta restrição permitiria definir a não relação das características. Este modelo servirá assim como base de comparação a outros onde a relação das características irá entrar naturalmente no modelo.

A configuração deste modelo difere-se da sugerida em "*common-preference*" que inicializa os vectores de características aleatoriamente. Para os dados analisados a utilização de dados reais em vez de dados aleatórios na inicialização do modelo aprimorou a qualidade da recomendação gerada.

A figura seguinte ilustra o processo utilizado para construir os vectores de características utilizando aos dados da *MovieLens*.

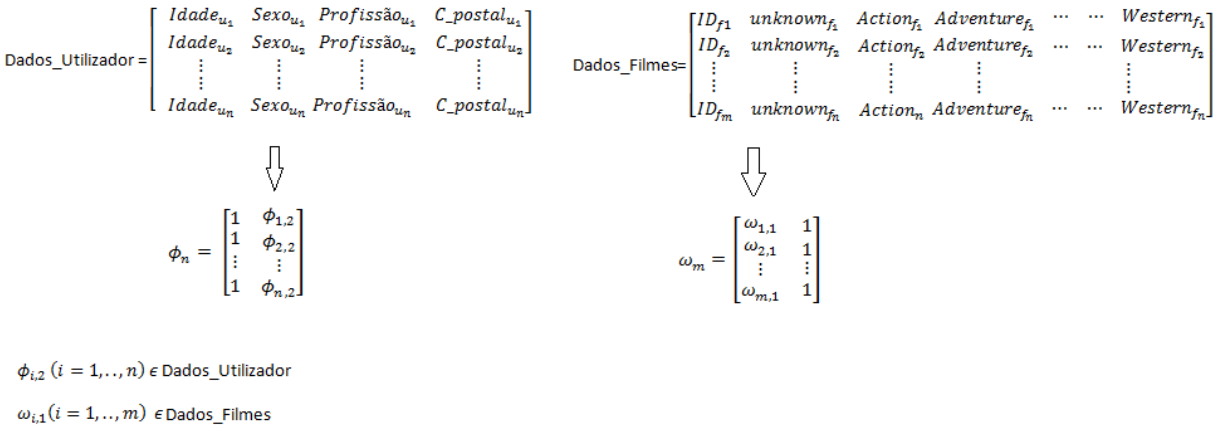

Fig. 15 Exemplo de construção dos Vectores de Características.

Para a construção dos vectores de características é aplicada validação cruzada com 4-fold, representada na Fig.16, em que é seleccionando 80% dos dados para treino/validação e 20% para teste.

Na fase de treino é processado todas as combinações possíveis de características dos utilizadores e dos itens sendo seleccionado o tuplo onde se obteve o menor erro. Seleccionado o par de características a representar os utilizadores e os itens é aplicado o procedimento "*Given L-test"* aos dados de teste, seleccionando apenas *L-test* avaliações por cada utilizador para avaliar o desempenho do modelo.

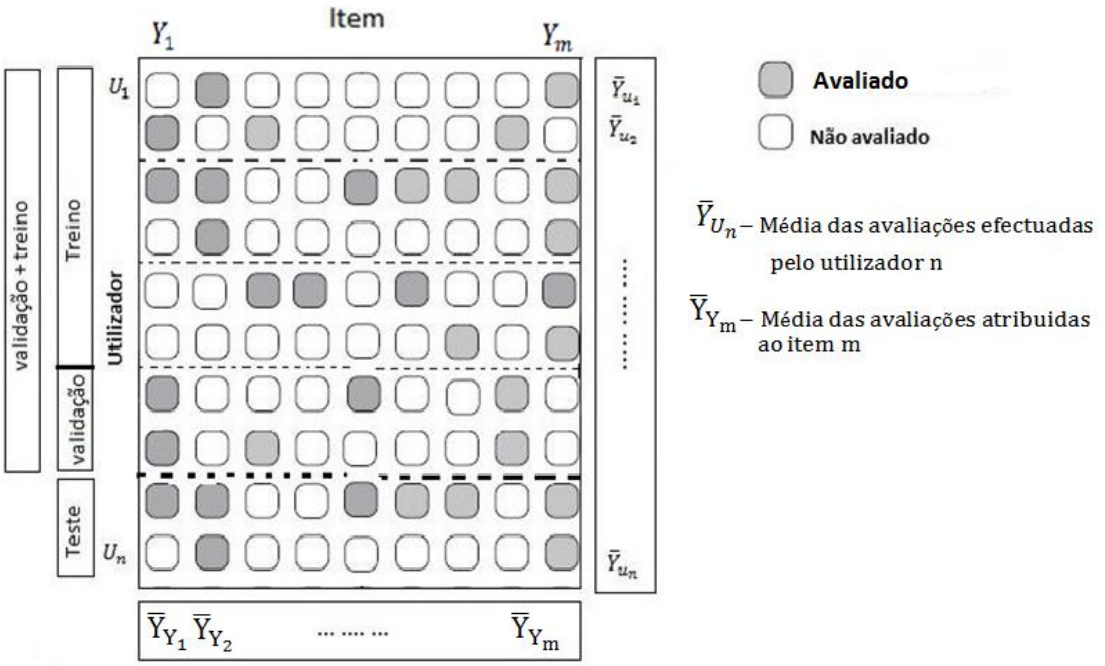

Fig. 16 Aplicação de 4-fold (adaptado de [33]).

Capítulo 5

Resultados

Para o desenvolvimento desta dissertação, procedeu-se ao estudo e implementação dum algoritmo de filtragem colaborativa identificado durante o levantamento do estado da arte. Esta abordagem incorpora a técnica dos modelos lineares generalizados para a factorização da matriz de avaliações. Relativamente a este trabalho, concentrámo-nos na implementação do modelo denominado *"common preference"* proposto pelos autores de[33]. Com base nesta abordagem, desenvolvemos uma nova metodologia híbrida com a inclusão de informação mais rica: idade e género dos utilizadores e detalhes próprios dos itens dependendo da base de dados em análise.

Os testes descritos neste capítulo foram efectuados utilizando a base de dados da MovieLens. Esta contém 6040 utilizadores, 3900 filmes e aproximadamente 1 milhão de avaliações discretas no intervalo [1,5].

A análise e comparação de sistemas de recomendação não são fáceis por estarem condicionados a diferentes aspectos. Por exemplo, diferentes algoritmos podem revelar-se piores ou melhores em diferentes bases de dados (dependendo de factores como: o número de utilizadores, itens e avaliações, a escala das avaliações e outras propriedades das bases de dados). Outra razão é que os objectivos das avaliações realizadas podem ser diferentes.

A avaliação do desempenho pode ser baseada em diferentes aspectos. Neste caso será baseado em duas das métricas de desempenho mais utilizadas em sistemas de recomendação: *Root Mean Square error* (RMSE) e *Mean Absolut Error* (MAE) que avaliam o erro entre as avaliações previstas pelo sistema e a matriz das avaliações utilizada pelo sistema para gerar a previsão. O RMSE é definido pela fórmula

$$\text{RMSE} = \sqrt{\frac{\sum(f(x_i)-y_i)^2}{n}} \tag{5.1}$$

e o MAE por

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|f(x_i) - y_i| \tag{5.2}$$

Sendo f(x_i) as avaliações previstas, y_i as avaliações presentes na matriz de avaliações e n o número de avaliações consideradas.

A análise do desempenho foi efectuada em duas etapas: a primeira consistiu em gerar a matriz com a previsão das avaliações através da aplicação dos modelos de filtragem colaborativa implementados à base de dados em estudo e a segunda na avaliação da relação entre as avaliações reais e as previstas. O desempenho foi calculado tendo em conta os utilizadores do conjunto de teste definido no subcapítulo 4.1.5.1.

Foram efectuados várias experiencias a fim de analisar a relação existente entre o número de avaliações considerado por utilizador e o desempenho do sistema. Constatou-se que o desempenho do sistema é proporcional ao número de avaliações consideradas por utilizador como é ilustrado nas figuras Fig.17 e Fig.18. Na implementação do modelo "*Collaborative filtering with interlaced generalized linear models*" foram realizadas as mesmas experiências.

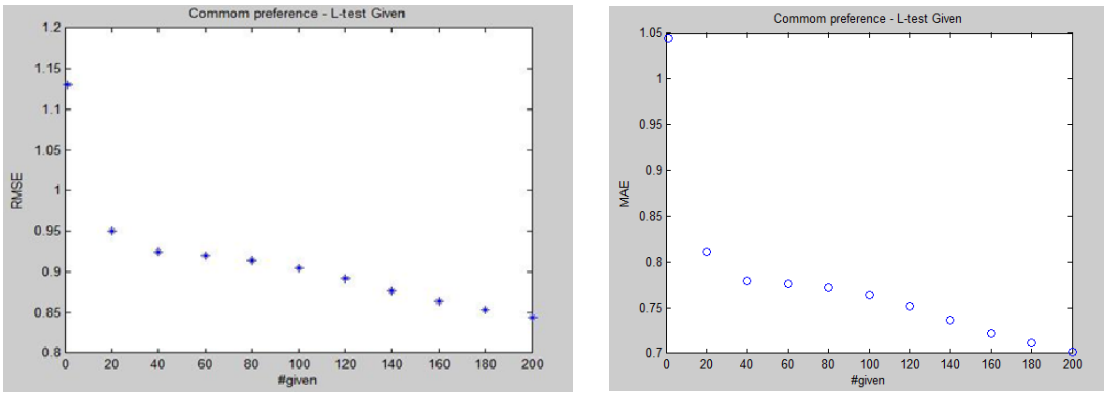

Fig. 17- Avaliação do desempenho de *Collaborative Filtering with interlaced generalized linear models* aplicado aos dados da *MovieLens.*

Relativamente à base de dados em análise o melhor valor de MAE encontrado na literatura é 0.652 [33].Os modelos analisados apresentam MAE = 0.701 e MAE = 0.652 sendo comparáveis com outros modelos do estado da arte.

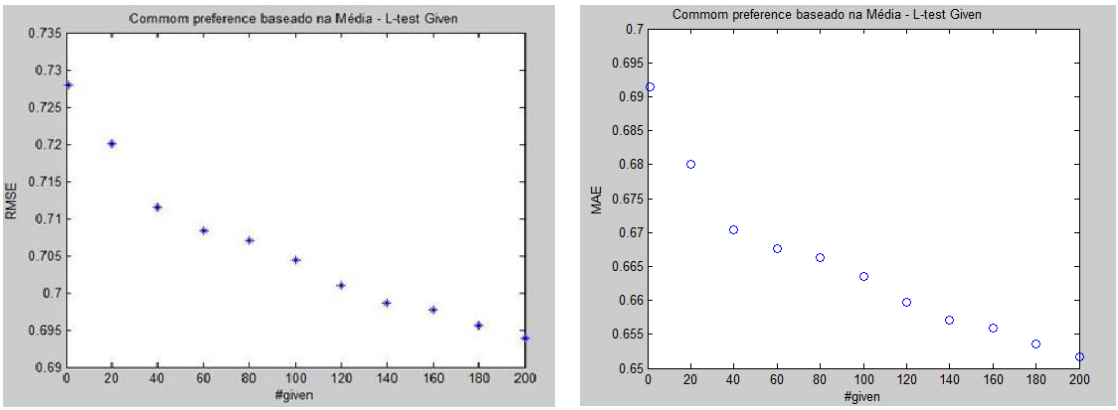

Fig. 18-Avaliação do desempenho de Filtragem colaborativa baseada na média das avaliações aplicado aos dados da MovieLens.

Foi realizado um estudo comparativo dos modelos de filtragem colaborativa implementados a fim de verificar qual das implementações representa melhor a base de dados em estudo. Na concepção dos dois modelos foram consideradas diferentes formas de introduzir restrições no modelo. Embora o modelo baseado na média das avaliações seja motivada no modelo GLM entrelaçado espera-se que as diferenças na introdução de restrições na sua concepção tenham influências no seu desempenho. A figura seguinte ilustra a representação do desempenho dos dois modelos.

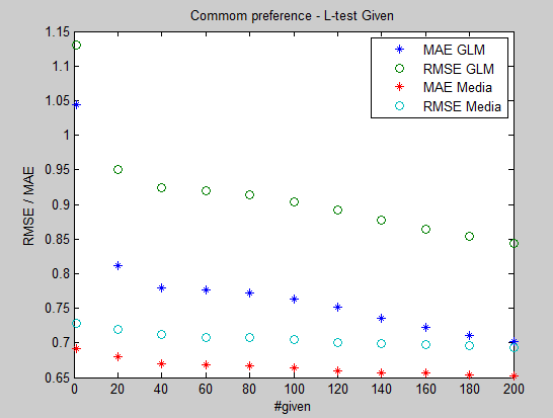

Fig. 19- Avaliação do desempenho dos modelos *Collaborative Filtering with Interlaced Generalized Linear Models* e Filtragem Colaborativa Baseada na Média das avaliações aplicada aos dados da *MovieLens.*

Da análise efectuada verifica-se que embora exista diferenças na concepção dos modelos em todos os casos o desempenho é proporcional ao número de avaliações consideradas por utilizador. Apesar desta correlação verifica-se que o modelo "Filtragem colaborativa baseada na média das avaliações" apresenta melhor desempenho relativamente a "*Collaborative filtering with interlaced generalized linear models"*.

Tabela 3- Avaliação do desempenho dos modelos Collaborative Filtering with Interlaced Generalized Linear Models e Filtragem Colaborativa Baseada na Média das avaliações aplicada aos dados da MovieLens

Given L_{test}	*Collaborative filtering with interlaced generalized linear models*		Filtragem colaborativa baseada na média das avaliações	
given	**RMSE**	**MAE**	**RMSE**	**MAE**
10	1.130	1.044	0.728	0.691
20	0.950	0.811	0.720	0.680
40	0.924	0.779	0.711	0.670
60	0.919	0.776	0.708	0.668
80	0.914	0.772	0.707	0.666
100	0.904	0.764	0.704	0.663
120	0.892	0.752	0.701	0.660
140	0.877	0.736	0.699	0.657
160	0.864	0.722	0.698	0.656
180	0.853	0.711	0.696	0.654
200	0.844	0.701	0.694	0.652

Além da análise do desempenho dos dois modelos referenciados foi seleccionado um conjunto de modelos de filtragem colaborativa do estado da arte que foram testados com a base de dados da *MovieLens* e avaliados com base no cálculo do RMSE apresentados na tabela 4.

Tabela 4- Análise do desempenho de alguns algoritmos do estado da arte.

Método		RMSE
A neural Network model for Collaborative Filtering [42]		0.98
Non-linear Matrix factorization with Gaussian processes[43]		0.874 ± 0.028
Collaborative filtering on a budget [44]		0.857 ±0.004
Matrix Factorization for Collaborative Prediction [45]	FMMMF	1.080
	Iterative SVD	1.050
	Repeated Matrix	0.950
A Guide to Singular Value Decomposition for Collaborative Filtering [48]	AVGB	0.931
	SVDNR	0.880
	SVD	0.872
	CSVD	0.870
The Long Tail of Recommender Systems and How to Leverage It [47]		0.930
Semi-Supervised Learning Methods and Memory-Based Methods [46]	User Average	1.043
	Movie Average	1.043
	Weighted Average	1.009
	Pearson Correlation	0.971
	Vector Similarity	0.973
	Default Voting	0.989
	Vector Similarity (on items)	0.955
	Default Voting (on items)	0.583
	Minimum Norm Interpolation	0.980
	Harmonic Energy Minimizing Functions	0.989
Collaborative filtering based on multi-channel diffusion[49]	Diffusion-based	0.948
	Pearson	1.026
Taste Mahout	Percentagem de utilizadores utilizada na avaliação	
	10%	0.9111
	20%	0.9173
	40%	0.909
	60%	0.896
	80%	0.899
	100%	0.899

Os valores do RMSE para a abordagem *taste mahout* foram obtidos através da implementação dos modelos de filtragem colaborativa disponibilizados pela *Framework tas-*

te. Esta dispõe de funções como *RMSRecommenderEvaluator* que permitem avaliar o desempenho dos modelos em estudo.

Os testes efectuados consistiram em considerar 80% dos dados para a construção do modelo e 20% para teste. Após a concepção deste modelo realizou-se a avaliação do desempenho tendo em conta a métrica RMSE para diferentes valores percentuais dos utilizadores do conjunto de teste.

Esta abordagem difere da *given l-test* dado que enquanto a primeira considera um subconjunto das avaliações de todos os utilizadores do conjunto de teste, esta considera um subconjunto dos utilizadores do conjunto de teste e todas as suas avaliações. Esta análise não tem como finalidade introduzir as restrições consideradas na concepção dos modelos implementados na abordagem implementada pelo *taste*, mas sim, analisar o desempenho de modelos diferentes.

Da análise dos valores de RMSE dos modelos apresentados na tabela 4 verifica-se que o modelo filtragem colaborativa baseada na média das avaliações apresenta desempenho comparável com os modelos analisados.

Capítulo 6

Conclusões e Trabalhos Futuros

Ao longo da realização desta tese foi implementado o modelo "*Collaborative filtering with interlaced generalized linear models*" validando a exposição apresentada em [33].
Como fruto dos estudos efectuados ao longo deste período foi concebido um novo modelo de filtragem colaborativa intitulado "Filtragem colaborativa baseada na média das avaliações" que assenta em técnicas de regressão para prever novas avaliações. O modelo representa os utilizadores e os itens como vectores de características que são inicializados com dados demográficos. O uso de dados demográficos na concepção do modelo reduziu a quantidade de bases de dados que poderiam ser utilizadas para efeitos de teste. Grande parte da geração actual dos modelos de algoritmos colaborativos faz uso apenas da matriz das avaliações na construção do modelo. Por isso, a maioria das bases de dados de domínio público estão muito direccionadas a este paradigma contendo apenas informação da matriz das avaliações.
Da avaliação do desempenho baseado nas métricas RMSE e MAE verificou-se que "Filtragem colaborativa baseada na média das avaliações" apresenta desempenho comparável com outros algoritmos do estado da arte.
Como perspectivas de trabalhos futuros no seguimento deste sugere-se:

- Analisar o comportamento do modelo utilizando outras funções pertencentes à família exponencial;
- Aplicar validação cruzada para escolher o valor do parâmetro λ;
- Aplicar o modelo a novas bases de dados;
- Integrar o modelo com aplicações Web.

Apêndice A

Familia Exponencial

Distribuição Normal

Para a distribuição Normal define-se:

$f(y|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\}$ com $y \in \Re$, $\mu \in \Re$ e $\sigma > 0$.

Então

$$f(y|\mu,\sigma^2) = \exp\left\{-\frac{\ln(2\pi\sigma^2)}{2} - \frac{y^2+\mu^2-2y\mu}{2\sigma^2}\right\}$$

$$= \exp\left\{\frac{y\mu-\mu^2/2}{\sigma^2} - \frac{1}{2}\left[\frac{y^2}{\sigma^2} + \ln(2\pi\sigma^2)\right]\right\},$$

ou seja, fazendo $\theta = \mu$ e $\phi = \sigma^2$

$$f(y|\theta,\phi) = \exp\left\{\frac{y\theta-\theta^2/2}{\phi} - \frac{1}{2}\left[\frac{y^2}{\phi} + \ln(2\pi\phi)\right]\right\}$$

Portanto

$$a(\phi) = \phi, \qquad b(\theta) = \frac{\theta^2}{2}, \qquad c(y,\phi) = -\frac{1}{2}\left[\frac{y^2}{\phi} + \ln(2\pi\phi)\right],$$

$E[Y] = \mu = \theta$ e $var[Y] = \phi$.

Distribuição Gama

A função densidade de probabilidade da distribuição Gama é definida por

$f(y|\alpha,n) = \frac{\alpha^n}{\Gamma(n)} e^{-\alpha y} y^{n-1}$, com $\alpha > 0, n > 0, y > 0$

então,

$$f(y|\alpha,n) = \exp\{n\ln(\alpha) - \alpha y + (n-1)\ln y - \ln(\Gamma(n))\}$$

$$= \exp\{-\alpha y + n\ln(\alpha) - (n-1)\ln(n) + (n-1)\ln(n) + (n-1)\ln(y) - \ln(\Gamma(n))\}$$

$$= \exp\left\{ny\left(-\frac{\alpha}{n}\right) + n\,ln\left(\frac{\alpha}{n}\right) + \ln(n) + (n-1)\ln(n) + (n-1)\ln(y) - \ln(\Gamma(n))\right\}$$

$$= \exp\left\{\frac{y\left(-\frac{\alpha}{n}\right) + \ln\left(\frac{\alpha}{n}\right)}{n^{-1}} + (n-1)\ln(yn) + \ln(n) - \ln(\Gamma(n))\right\}$$

e considerando

$$\theta = -\frac{\alpha}{n} \qquad e\ \phi = n$$

vem

$$f(y|\theta,\phi) = \exp\left\{\frac{y(\theta)+\ln(-\theta)}{\phi^{-1}} + (\phi-1)\ln(y\phi) + \ln(\phi) - \ln(\Gamma(\phi))\right\}$$

desta forma tem-se

$$a(\phi) = \phi^{-1}, \qquad b(\theta) = -\ln(-\theta),$$

$$c(y,\phi) = (\phi-1)\ln(y\phi) + \ln(\phi) - \ln\left(\Gamma(\phi)\right),$$

E[Y] = $-\frac{1}{\theta}$ e Var[Y] = $\frac{1}{\phi\theta^2}$.

Distribuição de Poisson

Para a distribuição de Poisson com função de probabilidade

$f(y|\lambda) = \frac{e^{-\lambda}\lambda^y}{y!}$, com y=0,1,2,2,… e $\lambda > 0$

Então

$$f(y|\lambda) = \exp\left\{\ln\left(e^{-\lambda}\right) + \ln(\lambda^y) - \ln(y!)\right\}$$

$$= \exp\{y\ln(\lambda) - \lambda - \ln(y!)\}$$

e fazendo $\theta = \ln(\lambda)$ vem

$$f(y|\theta,\phi) = \exp\{y\theta - e^{\theta} - \ln(y!)\}$$

obtem-se

$$a(\phi) = 1, \quad b(\theta) = e^{\theta}, \quad c(y,\phi) = -\ln(y!)$$

E[Y] = e^{θ} e Var[Y] = e^{θ}.

Distribuição Binomial

Se Y for tal que mY tem uma distribuição Binomial com parâmetros m e

π (Y $\sim B(m,\pi)/m$) a sua f.d.p é dada por:

$$f(y|\pi) = \binom{m}{ym}\pi^{ym}(1-\pi)^{m-ym}$$

$$= \exp\left\{ym\,ln\pi + m(1-y)\ln(1-\pi) + ln\binom{m}{ym}\right\}$$

$$= \exp\left\{m\left(y\theta - \ln\left(1+e^{\theta}\right)\right) + ln\binom{m}{ym}\right\}$$

com $y \in \left\{0, \frac{1}{m}, \frac{2}{m}, \ldots, 1\right\}$ e $\theta = \ln\left(\frac{\pi}{1-\pi}\right)$

definindo

$$\theta = \ln\left(\frac{\pi}{1-\pi}\right)$$

$$b(\theta) = \ln\left(1 + e^{\theta}\right), \qquad c(y, \phi) = ln\binom{m}{ym}$$

$$\dot{b}(\theta) = \frac{e^{\theta}}{1+e^{\theta}} = \pi, \quad \ddot{b}(\theta) = V(\mu) = \frac{e^{\theta}}{(1+e^{\theta})^2} = \pi(1-\pi)$$

$$a(\phi) = \frac{\phi}{\omega}, \ \phi = 1, \ \omega = m.$$

Bibliografia

[1] A. Tuizhlin, "Toward the Next Generation of Recommender Systems: A survey of the State-of-the-Art and Possible Extensions", IEEE Transactions on Knowledge and Data Engineering, Vol. 17 Nº 6, 2005.

[2] G. Takács, I. Pilászy, B. Németh e D. Tikk, "Scalable Collaborative Filtering Approaches for Large Recommender Systems", Journal of Machine Learning Research, 2009.

[3] G. Linden, B. Smith, e J. York, "Amazon.com Recommendations Item-to-Item Collaborative Filtering", IEEE Internet Computing, 2003.

[4] E. Reategui e S. Cazella, "Sistemas de Recomendação", congresso da Sociedade brasileira de computação, 2005.

[5] E. Schopf, C. Schepke, M. Silva, P. Silva, "Avaliação de Heurísticas de Melhoramento e da Meta heurística Busca Tabu para Solução de PRV", Centro de Electrónica e Tecnologia - Universidade Federal de Santa Maria, 2006.

[6] H. Kim, A. Ji, I. Ha, G. Jo, "Collaborative filtering based on collaborative tagging for enhancing the quality of recommendation", Electronic Commerce Research and Applications, 2009.

[7] M. Balabanovic e Y. Shoham, "Fab: Content-Based, Collaborative Recommendation," Comm. ACM, vol. 40, no. 3, pp. 66-72, 1997.

[8] P. Melville, R.J. Mooney, e R. Nagarajan, "Content-Boosted Collaborative Filtering for Improved Recommendations," Proc. 18th Nat'l Conf. Artificial Intelligence, 2002.

[9] M. Pazzani, "A Framework for Collaborative, Content-Based, and Demographic Filtering, Artificial Intelligence" Rev., pp. 393-408, 1999.

[10] I. Soboroff e C. Nicholas, "Combining Content and Collaboration in Text Filtering," Proc. Int'l Joint Conf. Artificial Intelligence Workshop: Machine Learning for Information Filtering, 1999.

[11] G. Adomavicius e A. Tuzhilin, "Multidimensional Recommender Systems: A Data Warehousing Approach," Proc. Second Int'l Workshop Electronic Commerce (WELCOM '01), 2001.

[12] J.A. Konstan, J. Riedl, A. Borchers, e J.L. Herlocker, “Recommender Systems: A GroupLens Perspective,” Proc. Recommender Systems, Papers from 1998 Workshop, Technical Report WS-98-08, 1998.

[13] L.H. Ungar e D.P. Foster, “Clustering Methods for Collaborative Filtering,” Proc. Recommender Systems, Papers from 1998 Workshop, Technical Report WS-98-08 1998.

[14] R.J. Mooney e L. Roy, “Content-Based Book Recommending Using Learning for Text Categorization,” Proc. ACM SIGIR ’99 Workshop Recommender Systems: Algorithms and Evaluation, 1999.

[15] M. Pazzani e D. Billsus, “Learning and Revising User Profiles: The Identification of Intere”, springer,1997.

[16] G. Adomavicius e A. Tuzhilin, “Expert-Driven Validation of Rule-Based User Models in Personalization Applications,” Data Mining and Knowledge Discovery, vol. 5, nos. 1 and 2, pp. 33-58, 2001.

[17] T. Fawcett e F. Provost, “Combining Data Mining and Machine Learning for Efficient User Profiling,” Proc. Second Int’l Conf. Knowledge Discovery and Data Mining (KDD-96), 1996.

[18] H. Mannila, H. Toivonen, e A.I. Verkamo, “Discovering Frequent Episodes in Sequences,” Proc. First Int’l Conf. Knowledge Discovery and Data Mining (KDD-95), 1995.

[19] C. Cortes, K. Fisher, D. Pregibon, A. Rogers e F. Smith, “Hancock: A Language for Extracting Signatures from Data Streams,” Proc. Sixth ACM SIGKDD Int’l Conf. Knowledge Discovery and Data Mining, 2000.

[20] M.D. Buhmann, “Approximation and Interpolation with Radial Functions,” Multivariate Approximation and Applications, N. Dyn, D. Leviatan, D. Levin, and A. Pinkus, eds., Cambridge Univ. Press, 2001.

[21] R. Schaback e H. Wendland, “Characterization and Construction of Radial Basis Functions,” Multivariate Approximation and Applications, N. Dyn, D. Leviatan, D. Levin, and A. Pinkus, eds., Cambridge Univ. Press, 2001.

[22] G. Adomavicius, R. Sankaranarayanan, S. Sen, e A. Tuzhilin, “Incorporating Contextual Information in Recommender Systems Using a Multidimensional Approach,” ACM Trans. Information Systems, 2005.

[23] A. Ansari, S. Essegaier, e R. Kohli, "Internet Recommendations Systems," J. Marketing Research, pp. 363-375, 2000.

[24] A. Nodari, "Os sistemas de recomendação como instrumento para atingir mercados de nicho" 2008. disponivel em: http://www.dominiopublico.gov.br/pesquisa/DetalheObraForm.do?select_action=&co_obra=117067- acedido em 27/06/2010.

[25] A. Kajimoto, R. Sousa, S. Zanetti, V. Cirone –" Sistemas de recomendação de notícias na Internet baseados em filtragem colaborativa", 2007.

[26] D. Lichtnow , J. Lima , S. Loh , R. Garin , L. Palazzo , T. Primo , A. Kampff e J. Oliveira "O Uso de Técnicas de Recomendação em um Sistema para Apoio à Aprendizagem Colaborativa", [RBIE] - Revista Brasileira de Informática na Educação, 2006.

[27] B. Sarwar, G. Karypis, J. Konstan e J. Riedl – "Application of Dimensionality Reduction in Recommender System - A Case Study ",2000.

[28] A. Ohata, J. Quintanilha- "O uso de algoritmos de clustering na mensuração da expansão urbana e detecção de alterações na Região Metropolitana de São Paulo"- Anais XII Simpósio Brasileiro de Sensoriamento Remoto, 2002.

[29] C. Tiago e V. Leitão- "Utilização de funções de base radial em problemas unidimensionais de análise estrutural", Semini 2002.

[30] D. Gohn, "A Apreciação Musical na era das Tecnologias Digitais", relatório realizado na ECA/USP, 2007.

[31] S. Caetano, J. Aires-de-Sousa, M. Daszykowski e Y. Heyden, "Prediction of enantioselectivity using chirality codes and Classification and Regression Trees", Analytica Chimica Acta, 2005.

[32] W.Hong, " Hybrid evolutionary algorithms in a SVR-based electric load forecasting model", Electrical Power and Energy Systems, March 2009.

[33] N. Delannay e M. Verleysen, "Collaborative filtering with interlaced generalized linear models",NEUROCOMPUTING, 2008.

[34] M. Turkman e G. Silva, "Modelos Lineares Generalizados - da teoria à prática", Universidade Técnica de Lisboa, 2000.

[35] R. Guimarães, J.Cabral, "Estatística", Faculdade de Engenharia da Universidade do Porto, McGraw-Hill, 1997.

[36] P. Luiz. A Curva Normal. Disponível em:

http://www.psi-ambiental.net/pdf/PasqCap03.pdf. Acedido em: 27/06/2010.

[37] A. Vieira, " Análise da Média e dispersão em experimentos factoriais não replicados para optimização de processos industriais", 2004.

[38] B.M. Marlin, R.S. Zemel, S. Roweis e M. Slaney, "Collaborative filtering and the missing at random assumption" UAI - Conference on Uncertainty in Artificial Intelligence, 2007.

[39] B. Marlin, "Collaborative filtering: a machine learning perspective", Master Thesis, 2004.

[40] J.S. Breese, D. Heckerman e C. Kadie, "Empirical analysis of predictive algorithms for collaborative filtering", 14th Conference on Uncertainty in Artificial Intelligence, 1998

[41] A. Paterek, " Improving regularized singular value decomposition for collaborative filtering" , KDD Cup and Workshop ,2007.

[42] A Neural Network Model for Collaborative Filtering, disponível em http://www.dcs.bbk.ac.uk/50years/eldaw.pdf - acedido em 24-06-2010.

[43] N.Lawrence, "Non-linear Matrix Factorization with Gaussian Processes", International Conference on Machine Learning, 2009.

[44] A.Karatzoglou, A.Smola e M. Weimer, "Collaborative Filtering on a Budget", AISTATS, 2010.

[45] A. Kleeman, N. Hendersen e S. Denuit, "Matrix Factorization for Collaborative Prediction", ICME, 2005.

[46] R. Walia, "Collaborative Filtering: A Comparison of Graph-Based Semi-Supervised Learning Methods and Memory-Based Methods", ICT,2008.

[47] Y. Park e A. Tuzhilin, "The Long Tail of Recommender Systems and How to Leverage It", ACM, 2008.

[48] C.Chao, "A Guide to Singular Value Decomposition for Collaborative Filtering",csie,2009.

[49] M.Shang, C.-Jin, T. Zhou e Y. Zhang , "Collaborative filtering based on multi-channel diffusion", Elsevier, 2009.