1
Introdução

1.1.
Motivação

Os sistemas de recomendação vêm sendo um tema de pesquisa constante desde os meados dos anos 90, quando surgiram os primeiros artigos sobre filtragem colaborativa (Adomavicius & Tuzhilin, 2005). Um sistema de recomendação tem por objetivo recomendar itens a usuários de forma a maximizar uma função de utilidade. Para isto é necessário atribuir um ranking aos itens que não foram vistos/consumidos por um determinado usuário. Por exemplo, em um sistema de recomendação de filmes a função de ranking pode ser representada através da nota que o usuário dá ao filme, e o objetivo do sistema recomendar filmes que tenham chance de obter notas altas para um usuário em específico.

Devido à abrangência desta formulação os sistemas de recomendação se aplicam aos mais diversos domínios como filmes, cds, livros, anúncios e etc. Mesmo com uma boa diversidade de aplicações, a maioria destes sistemas utilizam dados históricos do comportamento de consumo para poder recomendar novos itens e por este motivo podemos classificá-los a partir da utilização dos dados em três grupos: sistemas baseados em conteúdo, sistemas colaborativos e sistemas híbridos (Adomavicius & Tuzhilin, 2005).

Os sistemas baseados em conteúdo utilizam somente dados do perfil do usuário para realizar a recomendação, ou seja, considera somente o histórico de consumo do usuário. Já os sistemas colaborativos utilizam o perfil de outros usuários para recomendar, levando em consideração a semelhanças entre suas escolhas. Por fim os sistemas híbridos são uma combinação entre os sistemas baseados em conteúdo e os colaborativos. Este trabalho tem como objetivo avaliar o desempenho de um sistema colaborativo, no caso a Análise Probabilística de Semântica Latente, conhecida também pelo acrônimo PLSA (Hofmann, 2003).

O PLSA é um método probabilístico que identifica comunidades de usuários ou itens através de variáveis latentes, gerando grupos que possuem sobreposição,

ao contrário da maioria dos outros métodos de agrupamento. O PLSA apresenta também altos níveis de acurácia nas predições (Hofmann, 2003), porém tais níveis variam dependendo do domínio de aplicação e por este motivo analisamos o método em dois domínios: a recomendação de anúncios na web e a recomendação de filmes. Estes domínios foram escolhidos por sua grande relevância dentre as aplicações práticas de sistemas de recomendação e também pela disponibilidade de dados para experimentos.

A primeira aplicação trata da publicidade na web, que vem crescendo a uma taxa de 21% ao ano (IAB e PWC, 2008). Em 2007 o faturamento total da publicidade na web atingiu cerca de 21 bilhões de dólares, mostrando a solidez desta no mercado de publicidade. Nela existem diversas modalidades de anúncios como banners, e-mail marketing, patrocínios e outros, porém, a forma de divulgação predominante é a publicidade de busca que engloba 41% do faturamento anual. Tal modalidade consiste na veiculação de anúncios textuais correlatos às consultas submetidas por usuários de mecanismos de busca. Por exemplo, ao submeter a consulta "Flores" o mecanismo de busca retorna uma lista ordenada de anúncios correlatos à palavra onde poderíamos encontrar, por exemplo, floriculturas. É importante ressaltar que os mecanismos de cobrança por clique e o leilão de preços (Cavalcante, 2008), que são muito utilizados neste meio, tornam a ordenação dos anúncios de forma a maximizar a utilidade esperada uma tarefa não trivial, fazendo com que a veiculação de anúncios em publicidade de busca seja um tema interessante da área de sistemas de recomendação.

Já a recomendação de filmes foi experimentada através de dados do Netflix Prize (Netflix Prize, 2008) que oferece um prêmio de US$ 1.000.000,00 a quem melhorar em 10% o sistema de recomendação atualmente utilizado, o Cinematch. Estes dados consistem em uma base de 100 milhões de notas entre um e cinco estrelas dadas por de 480 mil usuários a 18 mil filmes.

1.2. Trabalhos Relacionados

Hofmann (1999) apresentou a Análise Probabilística de Semântica Latente, que consiste em um modelo estatístico de misturas gerado por uma variável não observável. O aprendizado desta variável é feito de forma a maximizar sua verossimilhança com os dados em questão utilizando para tal o algoritmo *Expectation Maximization* (Dempster et al.,1977). Foram apresentados também experimentos mostrando a melhoria conseguida pelo PLSA sobre o modelo LSA (Deerwester et al., 1990) na indexação automática de documentos. O modelo utilizado nos experimentos é apresentado em forma de Rede Bayesiana (Pearl, 1985) na figura 1.1 vislumbrando duas abordagens: a assimétrica (a) e a simétrica (b), sendo que em ambas os nós D, Z e W representam respectivamente os documentos, a variável latente e as palavras dos documentos.

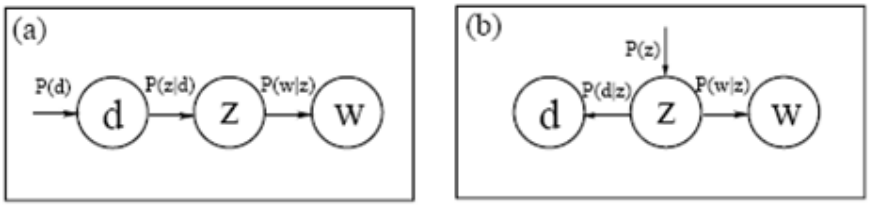

Figura 1.1 – Modelo do PLSA (Hofmann, 1999)

Schein (2002) apresentou uma abordagem hibrida do PLSA com foco em recomendar itens que ainda não foram classificados no conjunto de dados, também chamado de problema do *cold-start*. Com o intuito de atingir este objetivo são utilizados no lugar dos dados de preferência dos usuários por itens a preferência destes pelos atributos dos itens, tirando vantagem da coincidência entre os atributos já consumidos com os que ainda não foram. O exemplo apresentado é a recomendação de filmes onde são utilizados atores como atributo, recomendando os novos itens através da preferência dos usuários por estes e não pelos filmes. Este modelo é exibido na figura 1.2, onde em (a) temos as preferências entre os usuários e os filmes e em (b) entre os usuários e os atores, sendo nestes P os usuários, Z a variável latente, M os filmes e A os atores.

Hofmann (2003) apresentou uma generalização do PLSA para variáveis de resposta de valor contínuo. Esta modelagem assume que cada comunidade de usuários atribui notas aos itens segundo o comportamento de uma normal, sendo assim representada por seus parâmetros. Tal modelo foi chamado de *Gaussian Probabilistic Latent Semantic Analysis* ou *gPLSA* e foi experimentado no problema de recomendação de filmes obtendo bons resultados. Este modelo encontra-se na figura 1.3, onde U, Y, Z e V são variáveis que representam respectivamente os usuários, itens, a variável latente, e o valor das avaliações dadas pelos usuários, já Mzy, e Vzy representam os parâmetros numéricos da média e variância das gaussianas do modelo.

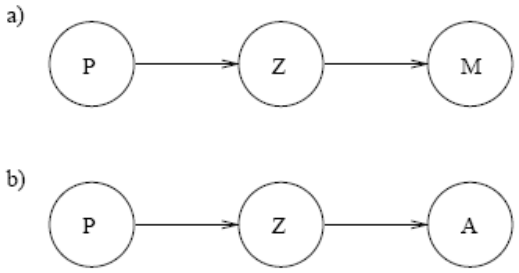

Figura 1.2 – Modelo do *PLSA* para abordagem do *cold-start*

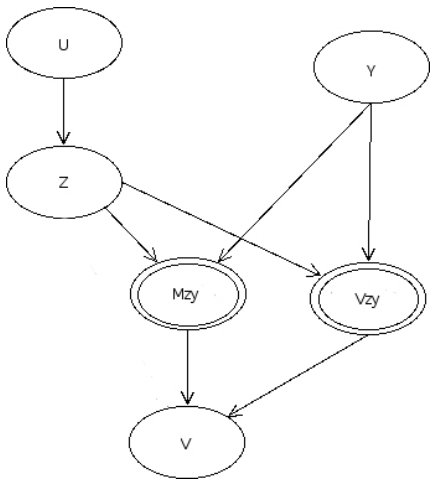

Figura 1.3 – Modelo do *gPLSA*

Si e Jin (2003) apresentaram o *Modelo Flexível de Mistura*, que é uma extensão do *PLSA* utilizando duas variáveis latentes com o intuito de formar grupos de usuários e itens. O modelo foi criado para melhorar a representatividade do *PLSA* de uma variável, onde as classes latentes representam somente a forma

com que os usuários atribuíram notas para determinados itens não sendo possível identificar os múltiplos interesses dos usuários e os diversos aspectos dos itens relacionados. O *Modelo Flexível de Misturas* é exibido na figura 1.4 onde as variáveis X, Zx, Y, Zy e R representam respectivamente os usuários, a variável latente de agrupamento de usuários, os itens, a variável latente de agrupamento dos itens e as avaliações feitas. Nos experimentos com um conjunto de dados de avaliações de filmes mostrou-se que o método supera outros cinco modelos de filtragem colaborativa no mesmo conjunto de dados.

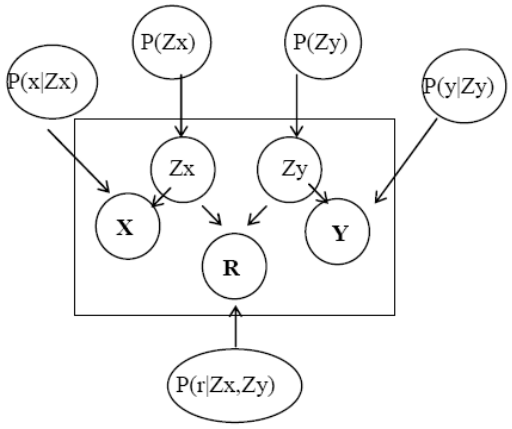

Figura 1.4 – Modelo Flexível de Misturas

Lin et al. (2007) apresentaram uma patente intitulada *Scalable Probabilistic Latent Semantic Analysis* ou *sPLSA* com o intuito de melhorar o desempenho computacional do algoritmo. Neste é gerado um modelo preliminar onde o objetivo é predizer a variável latente em questão e não as avaliações dadas pelos usuários. Para tal é executado o EM sobre o modelo preliminar e os dados gerados são extrapolados para o modelo original. Após isto é definido um valor mínimo para a probabilidade de um usuário pertencer a uma classe latente, caso na inicialização de uma probabilidade esta seja inferior ao limite estabelecido ela é eliminada do cálculo não sendo computada pelo algoritmo. Desta forma a quantidade de cálculos executados é diminuída acelerando o processamento e elimina-se o ruído gerado pela consideração de probabilidades pequenas para as predições aumentando a precisão do método.

Attardi, Esuli e Simi (2004) apresentam uma modelagem de recuperação de documentos chamada *Best Bets,* onde os documentos são recuperados através do perfil do usuário utilizando filtragem de informação, ao contrário das

metodologias tradicionais que utilizam a recuperação de informação. Neste são identificados os pares consulta / documento que casam no sistema, sendo esta informação armazenada a priori. O perfil do usuário é representado como um documento de vida curta contendo a consulta submetida, seu histórico de navegação e informações adicionais que possam identificar seu perfil. Quando o usuário efetua uma busca é feito o casamento entre o perfil e as consultas armazenadas, recuperando os documentos associados de forma ordenada. Tal processo de busca pode ser observado na figura 1.5.

É mostrada também uma modelagem do Google AdWords™ como um sistema *Best Bets*, onde no lugar de documentos são recuperados anúncios de interesse do usuário. Neste, o critério de casamento entre consultas e palavras-chaves é dado através de propriedades do anúncio, como por exemplo, o orçamento diário (quanto o usuário deseja gastar com o anúncio por dia). Para realizar a ordenação dos anúncios são utilizadas quatro variáveis, o custo por clique (CPC), a taxa de cliques dos anúncios (CTR – número de cliques dividido pelo número de impressões), o orçamento diário e a quantidade do orçamento diário já gasto. Além disto, a modelagem *Best Bets* exibe uma forma de recuperar documentos (ou anúncios) eficientemente, permitindo também a atualização incremental dos parâmetros e uma ordenação dinâmica.

Chickering e Heckerman (2007) apresentam uma patente de modelagem do problema de alocação de anúncios como um programa linear, onde a função objetivo é a utilidade, podendo englobar, por exemplo, métricas de vendas, lucro, ou projeção da marca. Esta pretende maximizar a utilidade dada às oportunidades de apresentação de anúncios (um determinado espaço de anúncio em uma determinada solicitação de página), podendo também agrupar as oportunidades de apresentação para diminuir o número de variáveis do problema. Já as restrições podem incluir capacidades máximas de oportunidades de apresentação, bem como quotas de impressões e utilidades mínimas para grupos de um ou mais anúncios.

Cavalcante (2008) apresentou uma abordagem para predição da taxa de cliques de um anúncio (CTR) utilizando filtragem colaborativa. A predição desta métrica foi proposta com base na influência considerável que o CTR tem na qualidade da recomendação de anúncios para a publicidade de busca. O modelo proposto foi o algoritmo de fatoração de matrizes *SVD (Singular Vector*

Decomposition) utilizando em uma base de dados sinteticamente desenvolvida ainda não experimentada por outros métodos até então.

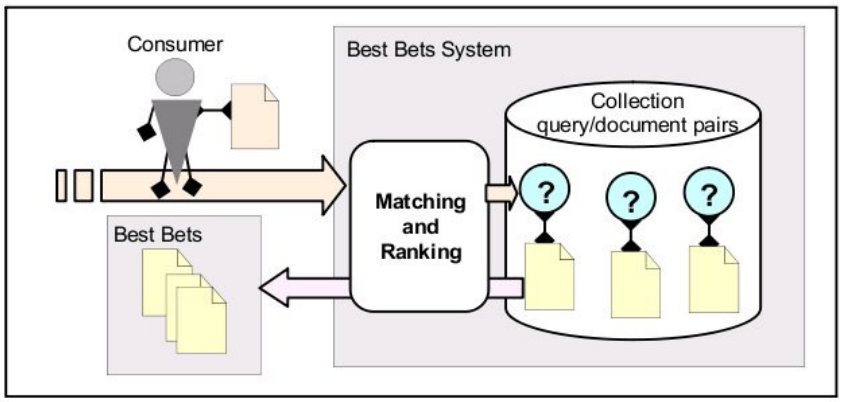

Figura 1.5 – Processo de busca do sistema Best Bets

Bennett et al. (2007) apresentam dados do progresso do Netflix Prize com o objetivo de identificar qual a abordagem obtém melhor desempenho para o problema e quais delas superam o sistema atualmente utilizado, o Cinematch, cujo erro quadrático médio é de 0.9514. Para tal, os dados das submissões ao prêmio foram organizados em um gráfico e foram utilizados os comentários do fórum do Netflix Prize para identificar as metodologias utilizadas. Os melhores resultados são as abordagens baseadas no SVD (Singular Vector Decomposition) e em comitês de modelos (Bell et al., 2007) que apresentam até 9,15% de melhoria sobre o cinematch (Netflix Prize, 2008). A figura 1.6 mostra o gráfico gerado a partir das 13 mil propostas submetidas pelos dois mil times ativos no prêmio.

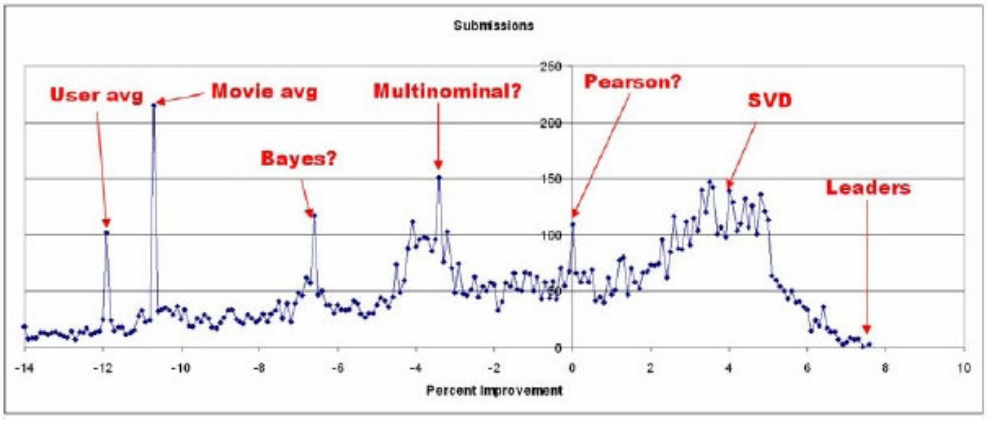

Figura 1.6 – Progresso das submissões ao netflix prize com os prováveis métodos extraídos do fórum do prêmio

1.3. Objetivo

O objetivo deste trabalho é avaliar o desempenho da Análise Probabilística de Semântica Latente nos problemas de recomendação de anúncios na web e recomendação de filmes.

Este trabalho também contribui para o projeto LearnAds, um *framework* de recomendação de anúncios baseado em Aprendizado de Máquina, e para a construção de um módulo de reutilizável do algoritmo PLSA.

1.4. Organização do Trabalho

Este trabalho está organizado da seguinte forma: o capítulo 2 descreve o funcionamento da Análise Probabilística de Semântica Latente, apresentando sua base matemática e um exemplo de funcionamento do algoritmo. O capítulo 3 apresenta a implementação do *PLSA* de forma a tornar flexível e reutilizável o código, além de apresentar a arquitetura do framework *LearnAds* juntamente com o módulo desenvolvido integrá-lo com o *PLSA*. O capítulo 4 expõe os experimentos realizados e por fim, no capítulo 5, são descritas as considerações finais.