1
Introdução

1.1.
Motivação

Devido à violenta quantidade de informações e sua disponibilidade facilitada pelo uso da Internet, as pessoas se deparam com uma imensa diversidade de opções. São inúmeros os produtos ou serviços oferecidos na Web e, se somarmos isto ao fato de que a maioria dos indivíduos não detém experiência para realizar uma escolha adequada, então obtemos justificativas suficientes para a existência dos Sistemas de Recomendação.

Figura 1: Motivação para os Sistemas de Recomendação.

Dentre as maneiras para implementar Sistemas de Recomendação, destacamos a Filtragem Colaborativa, técnica cujos primeiros trabalhos surgiram na década de 1990 (Hill et al., 1995; Resnick et al., 1994; Shardanand & Maes, 1995). A principal idéia da Filtragem Colaborativa é sugerir, para um dado indivíduo, os itens preferidos de outros indivíduos com gostos similares. A Figura 2 caricatura o conceito de recomendação a partir da sugestão alheia.

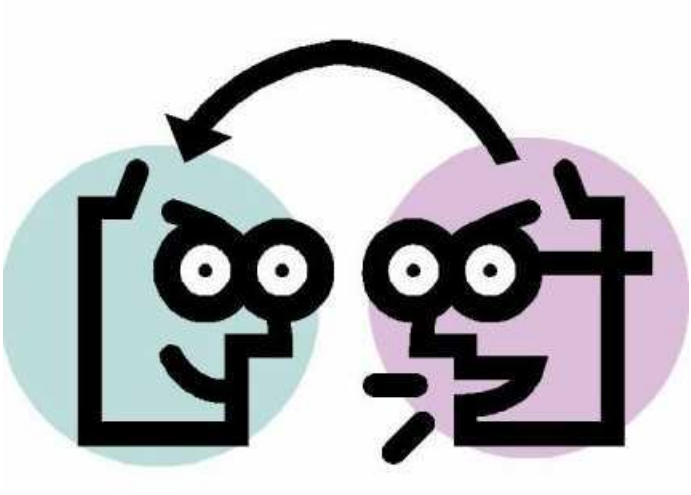

Figura 2: Recomendação colaborativa a partir da experiência de outrem.

Apesar de ordinariamente positivos, os resultados obtidos pelos modelos de Filtragem Colaborativa são passíveis de otimização através de um algoritmo denominado *Boosting* (Dietterich, 2000). Neste trabalho, realizamos experimentos com dois casos particulares de recomendação avaliando a eficiência do *Boosting* quando aplicado ao modelo de Filtragem Colaborativa.

1.1.1. Desafio da Recomendação de Anúncios na Internet

Eleita pela magnitude das receitas que movimenta, cerca de 21.1 bilhões de dólares no ano de 2007 (IAB Brasil & Pricewaterhouse Coopers, 2008), a recomendação de anúncios na Internet é um dos casos particulares de Sistemas de Recomendação analisados neste trabalho.

Trata-se de um serviço no qual os anunciantes pagam para exibir seus anúncios nas ferramentas de busca[1] (Cavalcante & Milidiú, 2008). Sua implementação mais popular é baseada no casamento entre as palavras-chave associadas aos anúncios e palavras-chave extraídas da busca do usuário (Weideman & Haig-Smith, 2002). Posteriormente, os anúncios pertinentes são ordenados e exibidos juntamente com os documentos retornados pela busca.

O lucro da recomendação de anúncios advém do clique dos usuários. Visando proporcionar credibilidade ao serviço e ampliar sua receita, os administradores das ferramentas de busca se defrontam com o desafio de expor constantemente um conteúdo atrativo.

A Figura 3 exemplifica um sistema de recomendação de anúncios. Os anúncios recomendados ao usuário que buscou por “emprego” estão destacados em marrom, a consulta está evidenciada em vermelho, na parte superior, e os seus resultados se encontram no retângulo preto.

[1] Google (http://www.google.com – ultimo acesso em julho de 2008), Yahoo! (http://www.yahoo.com – último acesso em julho de 2008), Live Search (http://www.live.com – último acesso em julho de 2008), entre outros.

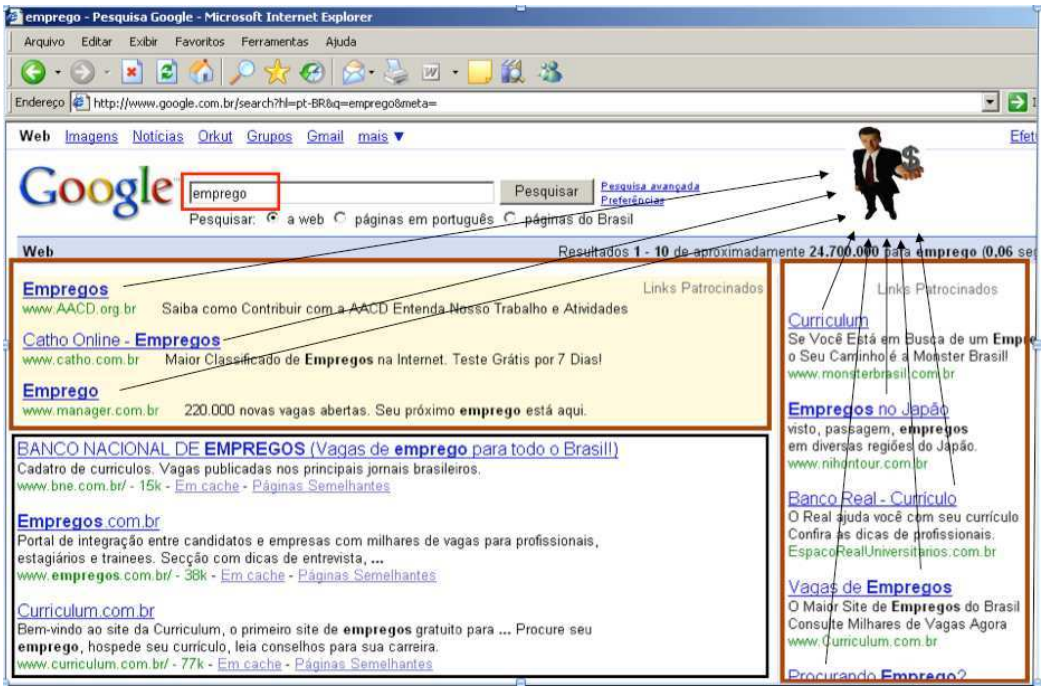

Figura 3: Lucro obtido através do clique nos anúncios.

1.1.2. Desafio da Recomendação de Filmes

A Netflix[2] é uma empresa dedicada a "conectar pessoas aos filmes que amam". Trata-se de uma locadora de DVDs que entrega, gratuitamente, os filmes através dos correios. Não existe o conceito de "entrega atrasada", ou seja, os clientes permanecem com o filme durante o tempo que desejarem. Entretanto, somente recebem novos filmes quando devolvem os que estão sob seu poder. A Figura 4 esquematiza o seu funcionamento.

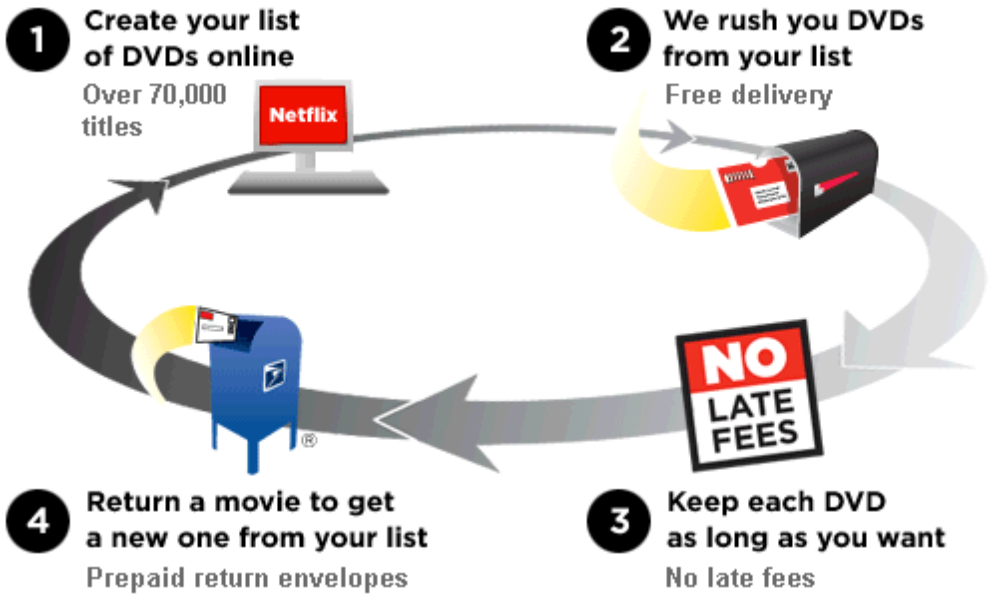

Figura 4: Logística de aluguel da Netflix.

Com o intuito de oferecer um diferencial aos seus usuários, a Netflix desenvolveu o Cinematch®, Sistema de Recomendação cujo trabalho é predizer quanto um cliente irá gostar de um filme que ainda não assistiu.

[2] Disponível em: http://www.netflix.com (último acesso em maio de 2008).

Como o desempenho do Cinematch® está aquém do esperado, a Netflix propôs um inusitado desafio: 1 milhão de dólares pelo desenvolvimento de um novo sistema que supere em 10% o atual. Em virtude da sua extrema dificuldade, são oferecidos prêmios de progresso no valor de 50 mil dólares para incentivar os competidores. As equipes contempladas por quaisquer prêmios devem compartilhar sua solução com a comunidade. A competição teve seu início em 2 de outubro de 2006 e irá se estender até 2 de outubro de 2011[3].

[3] Disponível em: http://www.netflix.com/rules (último acesso em julho de 2008).

1.2. Trabalhos Relacionados

Quando iniciamos nossa pesquisa sobre Sistemas de Recomendação no primeiro semestre de 2007, a proposta foi construir um *framework* para auxiliar nos experimentos. Como não possuíamos experiência com Sistemas de Recomendação, implementamos a recomendação de consultas descrita em Baeza-Yates et al. (2004). Trata-se de um algoritmo que, dada uma consulta submetida a uma ferramenta de busca, recomenda consultas a ela relacionadas. Para isto, consultas são automaticamente agrupadas de acordo com as informações contidas no histórico das ferramentas de busca (*query log*). O processo de agrupamento consiste na representação de cada consulta por um vetor de pesos e termos (*Term-weightVector*) das URLs clicadas a partir dessa consulta. Apesar da alta possibilidade de consultas semanticamente similares não compartilharem termos, elas com certeza compartilham termos nos documentos clicados pelo usuário. Desta forma, ao invés de compará-las diretamente, comparamos documentos clicados e por elas retornados.

Após a supracitada experiência com Sistemas de Recomendação, realizamos, no segundo semestre de 2007, um *brainstorm* de artigos para levantar o que estava sendo pesquisado na área. Encontramos um interessante estudo de Cristo & Ribeiro-Neto (2006) que analisa o uso de técnicas para otimizar Sistemas de Recomendação no estilo do Google AdSense®[4]. Já Mossri et al. (2007) utiliza programação genética para gerar boas funções de ordenação de anúncios no ato da recomendação, melhorando estratégias clássicas para este problema, como o *Term Frequency – Inverse Document Frequency* (TF-IDF) (Salton & McGill, 1983). Chickering & Heckerman (2007), por sua vez, apresentam uma invenção constituída de métodos aplicáveis à publicidade direcionada. A proposta determina aonde apresentar anúncios nas páginas e define o problema de recomendação como um programa linear.

Munidos de todo este bom referencial, voltamos nosso interesse para o aprendizado combinado (*ensemble learning*), em decorrência do seu sucesso na

[4] Disponível em: https://adsense.google.com (último acesso em julho de 2008).

competição Netflix. Os times ocupantes dos melhores lugares no *ranking*[5] revelaram a necessidade de utilizar preditores combinados para obter maior acurácia.

Em 2007, a equipe da AT&T[6] ganhou 50 mil dólares pelo seu desempenho na disputa. Segundo Bell et al. (2007), a solução contém uma combinação de 107 preditores. Dentre os diversos algoritmos de aprendizado combinado, optamos pelo *Boosting* para ser o objeto de pesquisa deste trabalho.

[5] Disponível em http://www.netflixprize.com/leaderboard (último acesso em julho de 2008).

[6] Disponível em: http://www.research.att.com (último acesso em julho de 2008).

1.3. Objetivo

Nosso objetivo é validar e testar o algoritmo de *Boosting*, quando empregado ao modelo de Filtragem Colaborativa utilizado no Sistema de Recomendação aqui abordado. Para tanto, realizamos experimentos com recomendação de anúncios e filmes. As figuras 5 e 6 ilustram a situação atual e a situação desejada.

Figura 5: Ponto de partida do trabalho.

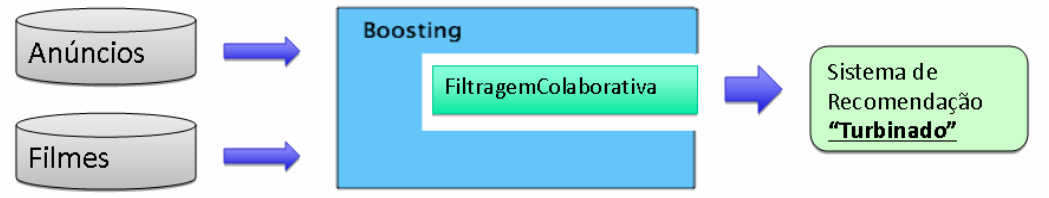

Figura 6: Objetivo do trabalho.

1.4. Organização do Trabalho

A dissertação está organizada da seguinte forma: o capítulo 2 aborda o problema de classificação e sua otimização através do *Boosting*. O capítulo 3introduz o modelo de Filtragem Colaborativa utilizado e propõe uma adaptação ao algoritmo de *Boosting* original. No capítulo 4 apresentamos os nossos resultados experimentais. Finalmente, no capítulo 5, tecemos as considerações finais.