1
Introdução

1.1.
Motivação

O surgimento da World Wide Web representou uma nova oportunidade de publicidade disponível para qualquer empresa: a possibilidade de exposição global para uma grande audiência a um custo extremamente pequeno. Na verdade, durante a década de 1990, muitas organizações estavam dispostas a fazer altos investimentos em publicidade na Web, aparentemente sem preocupação com o retorno desses investimentos (Weideman & Haig-Smith, 2002). Em virtude disso, a Web se tornou a mídia de crescimento mais rápido durante seus primeiros cinco anos, de acordo com o Interactive Advertising Bureau (IAB & PricewaterhouseCoopers, 2006).

Essa situação mudou radicalmente na década seguinte, quando a falência de muitas empresas Web resultou na queda da oferta de capital barato. Isso levou a uma grande preocupação quanto à confiabilidade de tais empresas como parceiros publicitários e, conseqüentemente, a uma considerável redução nos investimentos em publicidade *online* (Weideman, 2004; Weideman & Haig-Smith, 2002). Tal redução causou consecutivas quedas nas receitas dessas empresas no mercado dos EUA, a partir do primeiro quadrimestre de 2001. Essa tendência de perdas, no entanto, reverteu-se no fim de 2002 (IAB & PricewaterhouseCoopers, 2006).

Para entender melhor as razões da recuperação da publicidade *online*, temos que analisar o desempenho dos diferentes formatos de publicidade na Web ao longo do tempo. A Tabela 1 mostra as receitas geradas por sete formas distintas de publicidade na Web: *banners*/logotipos, patrocínios, e-mail, classificados/leilões, multimídia, busca, e *merchandise* (IAB & PricewaterhouseCoopers, 2006).

Tabela 1: Porcentagem das receitas geradas por tipo de anúncio na Web. Fonte: IAB, 1998-2005.

Formatos de publicidade	**1998**	**1999**	**2000**	**2001**	**2002**	**2003**	**2004**	**2005**
Banners/logotipos	56	56	48	36	29	21	20	20
Patrocínios	33	27	28	26	18	10	9	5
E-mail	-	2	3	3	4	3	3	2
Classificados/leilões	-	-	7	16	15	17	17	18
Multimídia	5	4	6	5	10	10	8	8
Busca	-	-	1	4	15	35	40	40
Merchandise	-	-	4	2	1	1	2	6
Outros	6	11	3	8	8	3	2	1
Total	100	100	100	100	100	100	100	100

Como podemos ver na Tabela 1, houve importantes mudanças na popularidade das várias formas de publicidade na Web. Por exemplo, os *banners*/logotipos sofreram um declínio gradual de 56% em 1998 para 20% em 2005, e um declínio similar é observado também no uso de patrocínios. Por outro lado, a **publicidade de busca** cresceu de 1% em 2000 para 40% em 2005, se tornando a principal forma de publicidade na Web. Portanto, a recuperação da publicidade na Web coincidiu com o crescente uso da publicidade de busca, o que, inclusive, não é um fenômeno transitório, já que tanto anunciantes quanto criadores de conteúdo declararam seus planos de aumentar os investimentos em publicidade de busca (Krol, 2005; Shields, 2005). De fato, segundo projeções da Forrester Research, até 2010 apenas a publicidade de busca representará um mercado de US$ 11,2 bilhões somente nos EUA (Maddox, 2005). Como conseqüência, surgiu toda uma nova indústria oferecendo serviços relacionados à publicidade de busca, em parte através de engenharia reversa dos algoritmos de ordenação das máquinas de busca (Feng et al., 2006).

Nos métodos de publicidade de busca, uma empresa anunciante paga por uma posição de destaque em listas de anúncios. Dentre tais métodos, o mais

popular é a técnica não-intrusiva de **publicidade baseada em palavras-chave** (Weideman & Haig-Smith, 2002). Nessa técnica, palavras-chave extraídas da consulta do usuário são casadas com palavras-chave associadas aos anúncios pelos anunciantes. Em seguida, os anúncios são ordenados, levando em consideração também a quantia que cada anunciante está disposto a pagar, de forma a escolher quais deles serão exibidos na página de resultados, juntamente com as respostas para a busca realizada pelo usuário.

O sucesso da publicidade baseada em palavras-chave motivou as máquinas de busca a oferecerem seus serviços de anúncios em diferentes contextos (Cristo, 2006). Por exemplo, anúncios relevantes poderiam ser exibidos aos usuários diretamente em páginas de portais de informação, com o objetivo de aproveitar suas necessidades imediatas de informação no momento da navegação. O problema de casar anúncios com a página Web que está sendo navegada, também conhecido como **publicidade baseada em conteúdo** (Lee, 2003), é diferente da publicidade baseada em palavras-chave porque, ao invés de lidar com palavras-chave fornecidas pelo usuário, é utilizado o conteúdo da página Web para decidir quais anúncios exibir.

É importante notar que as estratégias de publicidade de busca implicam em alguns riscos para as máquinas de busca. Por exemplo, existe a possibilidade de se causar um impacto negativo em sua credibilidade que, no longo prazo, pode reduzir sua participação de mercado (Bhargava & Feng, 2002). Portanto, os anúncios devem ser exibidos apenas para os usuários que se interessem por eles, no que se chama de **publicidade direcionada**.

A fim de minimizar a probabilidade de exibição de anúncios que não interessem aos usuários, as máquinas de busca investem na qualidade de seus sistemas de recomendação de anúncios, aumentando assim sua credibilidade e reforçando uma atitude positiva do usuário com relação aos anunciantes (Wang et al., 2002). Além disso, estes investimentos podem se refletir em maiores taxas de cliques, levando a um acréscimo nas receitas da máquina de busca e dos anunciantes, com ganhos para todas as partes (Bhargava & Feng, 2002).

1.2. Trabalhos Relacionados

Baeza-Yates et al. (2004) apresentam um algoritmo que, dada uma consulta submetida a uma máquina de busca, recomenda consultas a ela relacionadas. Para tanto, são criados grupos de consultas relacionadas que são automaticamente agrupadas com base em informações contidas nos *logs* da máquina de busca. Nesse processo de agrupamento, cada consulta é representada por um vetor de pesos de termos, obtido através da agregação dos vetores de pesos de termos das URL's clicadas a partir dessa consulta. Consultas semanticamente similares podem não compartilhar termos, mas provavelmente compartilham termos nos documentos selecionados pelo usuário. Dessa forma, o método proposto evita os problemas de comparação e agrupamento de coleções de vetores esparsos, nas quais é difícil encontrar consultas semanticamente similares, um problema que surge em trabalhos anteriores sobre agrupamento de consultas. Nesse método, as consultas sugeridas são ordenadas de acordo com dois critérios de relevância: a similaridade das consultas sugeridas com a consulta submetida à máquina de busca; e o suporte, que mede o quanto as respostas da consulta atraíram a atenção dos usuários, pois é interessante recomendar consultas que sejam úteis para muitos usuários na sua busca por informação.

Já Zhang & Nasraoui (2006) apresentam um método mais simples e intuitivo para recomendação de consultas em máquinas de busca. A fim de obter uma solução mais abrangente, os autores combinam dois métodos. Por um lado, interpretam o comportamento de buscas consecutivas do usuário como um processo de refinamento de consultas, o qual deve formar a base para o processo de refinamento de consultas da máquina de busca. Nesse âmbito, a seqüência de consultas é modelada como um grafo direcionado, no qual os vértices representam as consultas e a similaridade entre elas é inversamente proporcional à sua distância. Por outro lado, esse método é combinado com um método de similaridade baseado em conteúdo, no qual são comparados os vetores de pesos de termos das próprias consultas, a fim de compensar a pequena quantidade de consultas realizadas em uma mesma sessão.

No contexto da geração de palavras-chave para a publicidade em máquinas de busca, Joshi & Motwani (2006) propõem uma estratégia para a escolha de

palavras-chave que sejam, ao mesmo tempo, relevantes e não-óbvias, pois elas são economicamente mais viáveis. Os autores afirmam que apostar em muitas palavras-chave não-óbvias reduz o custo com publicidade, mantendo, apesar disso, o mesmo volume de cliques obtido com palavras-chave caras. Nesse âmbito, o principal desafio é gerar uma grande quantidade de palavras não-óbvias que sejam relevantes. Para isso, eles desenvolveram o TermsNet, uma abordagem que auxilia máquinas de busca a capturarem as relações semânticas entre termos, através de um grafo direcionado. Para gerar um conjunto de palavras-chave não-óbvias relacionadas a um termo, basta observar seus vizinhos no grafo.

Yih et al. (2006) relatam que uma parte substancial das receitas que financiam serviços gratuitos na Web advém da exibição de publicidade contextualizada baseada em palavras-chave extraídas automaticamente de páginas de conteúdo. Partindo disso, os autores propõem que essa extração seja realizada por um classificador automático que tenta predizer, dentre um conjunto de palavras-chave candidatas contidas em uma página, quais são boas alternativas para seu casamento com anúncios. O sistema desenvolvido aprende a classificar palavras-chave usando um variado conjunto de atributos, tais como a freqüência dos termos de cada palavra-chave candidata, a freqüência inversa do documento, a presença em metadados, e a freqüência com que os termos ocorrem nos *logs* de consultas de uma máquina de busca, além de atributos lingüísticos bastante estudados na área de Processamento de Linguagem Natural. Baseado no treinamento realizado com um conjunto de páginas de exemplo, etiquetadas manualmente com palavras-chave "relevantes", o sistema é capaz de extrair palavras-chave de páginas que não foram vistas anteriormente. Através de experimentos, os autores mostram que o sistema desenvolvido é substancialmente melhor que alguns sistemas de referência, tais como o tradicional modelo *Term Frequency – Inverse Document Frequency* (TF-IDF) (Salton & McGill, 1983), atingindo, em uma das métricas utilizadas, um resultado equiparável ao desempenho humano.

Cristo (2006) analisa o uso de técnicas de Recuperação de Informação para melhorar o desempenho de sistemas de recomendação de anúncios no casamento destes com o conteúdo de páginas da Web. Inicialmente, o autor investiga como utilizar adequadamente, no processo de recomendação de anúncios, diferentes evidências já disponíveis para empresas que operam sistemas de publicidade

baseada em palavras-chave. Em seguida, estuda métodos de classificação automática de páginas da Web, obtendo seus melhores resultados quando une estratégias de classificação baseadas em apontadores com as tradicionais, baseadas unicamente na análise de texto. Finalmente, utiliza os melhores classificadores obtidos como fonte de informação conceitual, e conclui que a combinação de métodos baseados em casamento sintático com os baseados em casamento conceitual apresenta melhor desempenho que aqueles baseados unicamente em casamento sintático.

Por sua vez, Chickering & Heckerman (2007) apresentam uma invenção constituída de métodos aplicáveis à publicidade direcionada. O sistema desenvolvido determina aonde apresentar anúncios, de forma a maximizar uma utilidade esperada, sujeita a uma ou mais restrições. Essas restrições podem incluir capacidades máximas de oportunidades de apresentação, bem como quotas de impressões e utilidades mínimas para grupos de um ou mais anúncios. Apesar da métrica de utilidade tradicional em publicidade na Web ser o *Click-Through Rate* (CTR), os autores adotam uma definição de utilidade mais ampla, podendo englobar, por exemplo, métricas de vendas, lucro, ou projeção da marca. O problema de maximização da utilidade sujeita a restrições é definido como um programa linear. Dessa forma, os autores apresentam algumas técnicas de agrupamento automático das oportunidades de apresentação de anúncios a fim de reduzir a quantidade de incógnitas desse conhecido problema.

Attardi et al. (2004) apresentam o *Best Bets*, um sistema de recuperação que generaliza a Filtragem de Informação, na qual as consultas são recuperadas a partir de documentos – o contrário do que é realizado na Recuperação de Informação. A Figura 1 descreve o modelo *Best Bets*, no qual o usuário ou suas necessidades de informação são representados, indiretamente, como um documento que coleta alguns aspectos sobre o usuário ou suas necessidades de informação, tal qual o histórico de navegação na Web ou uma consulta a uma máquina de busca. O material relativo ao conteúdo é armazenado em uma coleção, na qual cada item é composto pelas unidades de conteúdo propriamente ditas e pelas consultas que servem de critério de seleção para seus alvos planejados. Para cada requisição do usuário, todas as consultas que casam são selecionadas, e o conteúdo a elas associado é ordenado através de uma métrica e exibido para o usuário.

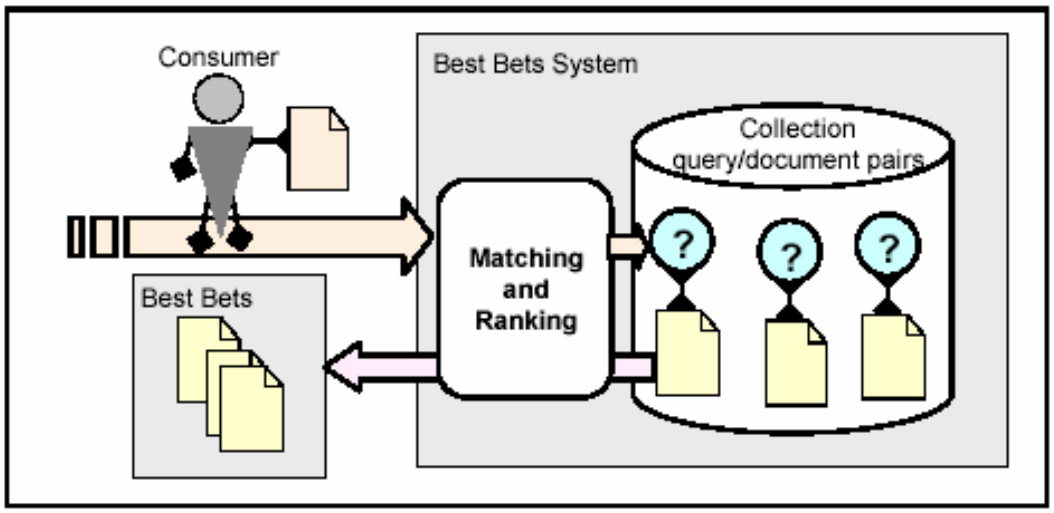

Figura 1: Modelo Best Bets (retirado de Attardi et al., 2004).

O *Best Bets* é bastante apropriado para aplicações nas quais a seleção de alvos para conteúdos específicos baseia-se em critérios definidos pelos próprios provedores desses conteúdos. É exatamente isso o que ocorre no caso da publicidade direcionada, na qual o conteúdo específico é o anúncio, o provedor deste conteúdo é o anunciante, e os critérios para seleção de alvos são as palavras-chave que o anunciante associa ao seu anúncio. Nesse contexto, os autores modelam o Google AdWords™[1] como um sistema *Best Bets*, e discutem como tratar questões de desempenho em sua implementação, mais especificamente: busca eficiente, atualizações incrementais e ordenação dinâmica.

Resumindo, os trabalhos desenvolvidos em (Chickering & Heckerman, 2007; Cristo, 2006; Yih et al., 2006; Baeza-Yates et al., 2004) mostram a relevância das técnicas de Aprendizado de Máquina na resolução de problemas relacionados à publicidade direcionada. Enquanto Chickering & Heckerman (2007) utilizam o agrupamento automático de oportunidades de apresentação para reduzir a quantidade de incógnitas do programa linear que definem, Baeza-Yates et al. (2004) agrupam consultas similares como parte de seu método de recomendação de consultas relacionadas a uma consulta de entrada. Já Cristo (2006) propõe estratégias de classificação automática de páginas baseadas em

[1] Disponível em: https://adwords.google.com/ (último acesso em março de 2008).

apontadores, utilizando diferentes algoritmos de aprendizado supervisionado, cujo resultado serve como fonte de informação conceitual para o casamento de páginas com anúncios. Finalmente, Yih et al. (2006) tratam a extração de palavras-chave de páginas de conteúdo como um problema de classificação automática dos termos contidos nessas páginas.

1.3. Objetivo

O objetivo deste trabalho é aplicação de técnicas de filtragem colaborativa baseada em fatoração de matrizes ao problema de predição do CTR em publicidade direcionada.

Adicionalmente, o trabalho contribui para o projeto e implementação do LearnAds, um *framework* de recomendação de anúncios baseado em Aprendizado de Máquina.

1.4. Organização do Trabalho

Este trabalho está organizado da seguinte forma: o capítulo 2 descreve o desenvolvimento da publicidade de busca, apresentando os modelos de leilão de palavras-chave utilizados, e destacando a importância da predição do CTR. O capítulo 3 apresenta os sistemas de recomendação baseados em filtragem colaborativa, indicando como utilizá-los na predição do CTR, e descrevendo um algoritmo de filtragem colaborativa baseado na técnica de fatoração de matrizes. O capítulo 4 apresenta o LearnAds, um *framework* de recomendação de anúncios baseado em Aprendizado de Máquina, que é um dos produtos deste trabalho, e que foi utilizado como base para a realização dos experimentos descritos no capítulo 5. Por fim, no capítulo 6, são apresentadas as considerações finais.