1
Links Patrocinados

1.1
Introdução

Links patrocinados são aqueles que aparecem em destaque nos resultados de uma pesquisa na Internet; em geral, no alto ou à direita da página, como na Figura 1.1.

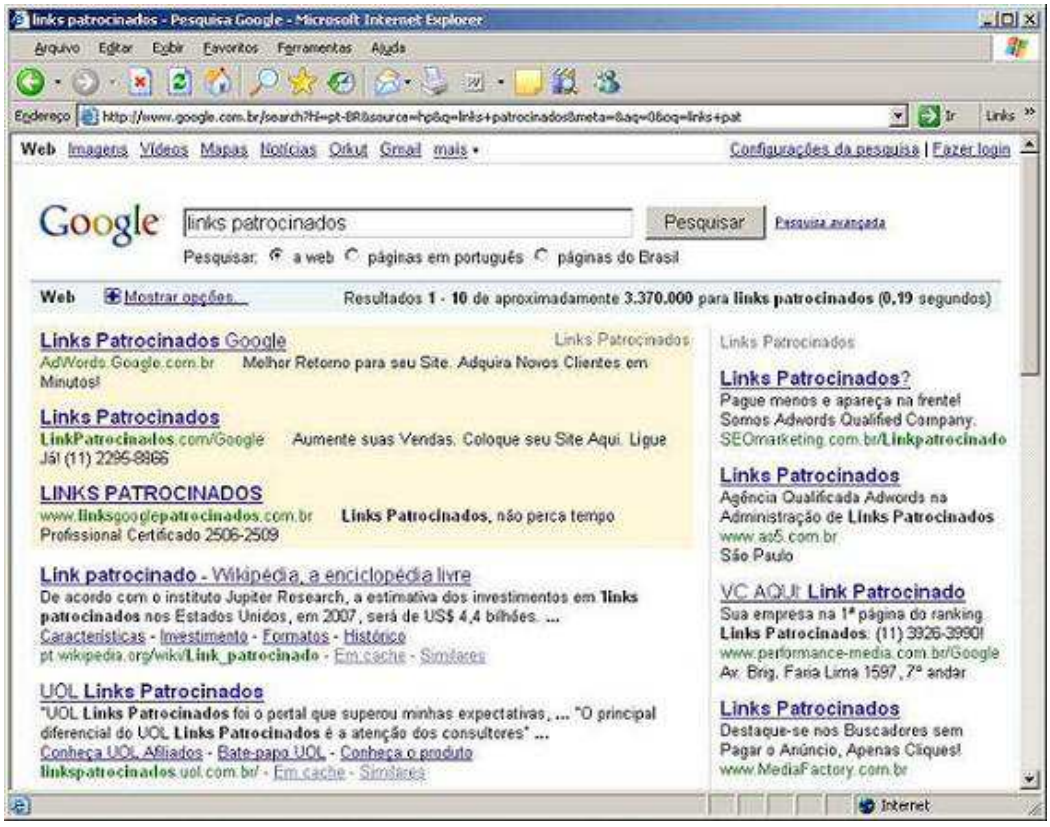

Figura 1.1: Exemplo de links patrocinados no Google

As consultas de quem navega pela Internet, apesar de curtas, trazem embutidas diversas informações sobre os interesses de quem a está realizando (MSVV07). As empresas, que implementam máquinas de busca, ao perceberem este fato, revolucionaram a forma de anunciar na Internet, mudando do modelo de anúncios fixos (*banners*) por página, para anúncios específicos por consulta.

PUC-Rio - Certificação Digital Nº 0711297/CA

Esta mudança atraiu pequenos e grandes anunciantes que conseguem através de links direcionados para suas próprias páginas captar a atenção de seu público alvo.

A empresa Overture, que hoje pertence a Yahoo!, é considerada pioneira neste modelo de negócio (NRTV07), pois foi a primeira a ordenar os resultados das consultas dos usuários, priorizando a exibição de links dos anunciantes que faziam ofertas por estas consultas. Atualmente todas as grandes empresas que oferecem máquinas de busca na Internet como Google, Yahoo! e MSN possuem este serviço. Segundo (Bat05), os links patrocinados já se tornaram a principal fonte de receita destas empresas, movimentando bilhões de dólares na Internet.

A seleção de quais anunciantes aparecerão em destaque no resultado de uma consulta tem início quando anunciantes participam de leilões e fazem ofertas por determinadas palavras. Cada anunciante determina também um orçamento máximo por dia. De acordo com estas ofertas, orçamentos e a importância da palavra nos resultados da consulta, são selecionados os anunciantes para aparecerem em destaque a cada nova consulta.

Neste modelo de negócio, um dos grandes desafios das empresas, que implementam as máquinas de busca, é selecionar os anunciantes que serão exibidos para cada consulta de modo a maximizar sua receita em um determinado período. Este problema de otimização online foi discutido primeiramente em (MSVV05) e posteriormente em (BJN07), (MNS07) e (MSVV07), entre outros. Estudar este problema e avaliar experimentalmente algumas possíveis soluções é a motivação principal desta dissertação.

Existem outros problemas importantes neste modelo de negócio como a condução dos leilões na Internet e a detecção de consultas falsas que tenham por objetivo estourar o orçamento dos anunciantes. Estes problemas são objetos de estudo de diversos artigos técnicos, mas não fazem parte do escopo desta dissertação.

1.2 Descrição do Problema

Considere que n anunciantes, $a_1, a_2, ..., a_n$, contratam o serviço de links patrocinados com um orçamento B_i por anunciante. Dentro de um conjunto Q de consultas possíveis, cada anunciante i faz por uma consulta j uma oferta b_{ij}. Durante um dia de processamento chega uma sequência de consultas $q_1, q_2, ..., q_m$ e para cada uma deve ser alocado um anunciante. O objetivo da empresa que oferece o serviço é maximizar sua receita diária respeitando os orçamentos de cada anunciante. Existem ainda outros objetivos, como garantir um número mínimo de exibições por anunciante, mas não serão considerados

PUC-Rio - Certificação Digital Nº 0711297/CA

para efeito deste estudo.

Uma das variações comuns a este problema é estender o número de anunciantes por consulta a um limite de T espaços para exibição (*slots*). Neste caso, existem duas diferentes abordagens: em (MSVV07), considera-se que as ofertas b_{ij} são iguais para todos os espaços e, em (BJN07), que são feitas ofertas b_{ijk} diferenciadas para cada espaço k de acordo com a posição que este espaço ocupa na página a ser exibida.

O cálculo da solução ótima no fim de um dia é um problema NP-completo, mas podemos atingir uma $1 + \epsilon$-aproximação por programação linear (BHJMQS07) se consideramos que as ofertas são bem pequenas em relação aos orçamentos, o que é bastante aceitável nos casos reais. Isto nos permite calcular a solução ótima no fim de um dia e avaliarmos o desempenho real do algoritmo online utilizado na seleção dos links.

Em nossa dissertação substituímos a restrição dos orçamentos B_i por um limite de exibições por anunciante L_i, independente da consulta. Esta mudança pode ser importante em problemas similares. Em (MG07), por exemplo, é proposta uma nova linguagem para os leilões que permite definir limite de exibições do anunciante por palavra chave. Esta alteração também permite o cálculo da solução ótima por fluxo em redes com capacidades inteiras, o que garante que a solução ótima da programação linear será exata e não apenas uma aproximação.

PUC-Rio - Certificação Digital Nº 0711297/CA

1.3 Trabalhos Relacionados

Nesta seção descrevemos alguns trabalhos relacionados ao problema dos links patrocinados. Os algoritmos apresentados em (MSVV07) e (BJN07), que serão descritos brevemente nesta seção, foram avaliados experimentalmente e são detalhados mais adiante nesta dissertação.

1.3.1 Algoritmos de Matching

O problema descrito pode ser considerado uma generalização do *online matching* discutido em (KVV90). As ofertas por cada palavra seriam 0 ou 1 e os orçamentos sempre 1. No artigo, é apresentado o algoritmo aleatório chamado RANKING, que gera uma permutação aleatória dos anunciantes para desempates e alcança uma competitividade de $1 - 1/e$. Os autores também provam que nenhum algoritmo aleatório online pode atingir uma

competitividade melhor.

Para um outro caso especial do problema, o *online b-matching*, (KP00) propõe um algoritmo determinístico chamado BALANCE. Neste caso, as ofertas são 0 ou 1 e o orçamento é B para todo anunciante. Este algoritmo seleciona o anunciante com o maior orçamento disponível no momento da consulta e sua competitividade também é $1-1/e$, quando B tende a infinito.

1.3.2 Algoritmos para Seleção de Links

O problema da seleção de links patrocinados é efetivamente discutido em (MSVV05). A motivação vem do interesse da Google em melhorar sua receita neste modelo de negócio e por isso, no artigo, o problema é tratado como *AdWords*, nome deste mecanismo na Google. O algoritmo proposto no artigo favorece os anunciantes com maiores ofertas, mas também aqueles que possuem mais orçamento disponível. Para isto foi derivada uma função de *trade-off* que utiliza os dois parâmetros. A cada nova consulta, o algoritmo seleciona o anunciante que maximiza o valor da função. Desde que as ofertas sejam bem pequenas em relação ao orçamento, este algoritmo atinge uma competitividade de $1 - 1/e$.

Para encontrar esta função e provar a competitividade do algoritmo, os autores utilizaram uma técnica que chamaram de *trade-off revealing*. Primeiramente o artigo prova a competitividade do algoritmo BALANCE, quando os orçamentos são iguais, modelando o problema através de programação linear e encontrando a função ideal para maximizar a receita, pela técnica de *factor revealing* (JMMSV03). Quando os orçamentos são diferentes, como no caso do *AdWords*, na modelagem por programação linear (PL) surge uma função do lado direito da restrição, que podemos considerar como uma família de problemas de PL ou uma PL para cada instância do problema. A função que maximiza o conjunto de problemas é encontrada através de *trade-off revealing*.

Em (BJN07) é apresentado um algoritmo com a mesma competitividade encontrada em (MSVV05), mas com uma prova bem mais simples. O método utilizado é a aproximação da solução ótima por programação linear através da relação primal-dual. No artigo, o problema online é definido como se fosse offline e já tivéssemos todas as consultas. Considerando as ofertas pequenas em relação aos orçamentos, é montado um modelo de programação linear (dual) que aproxima a solução inteira, permitindo que uma consulta seja alocada de

PUC-Rio - Certificação Digital Nº 0711297/CA

forma fracionada a mais de um anunciante. Na implementação do algoritmo entretanto, cada consulta é inteiramente atribuída a um anunciante apenas, respeitando-se os orçamentos.

Para atingir uma competitividade de $1 - 1/c$, o algoritmo faz com que, a cada alocação de consulta, a variação do custo primal seja no máximo $1 + 1/(c - 1)$ a variação do lucro dual, e maximizando o valor de c garante que as duas soluções permaneçam viáveis. Definindo $R_{max} = max\{b_{i,j}/B_i\}$ e fazendo $c = (1 + R_{max})^{1/R_{max}}$, temos um algoritmo $(1 - 1/c)(1 - R_{max})$-competitivo. Quando R_{max} tende a zero, o algoritmo tende a $1 - 1/e$.

O algoritmo guloso onde cada nova consulta seria atribuída ao anunciante com a maior oferta para esta consulta foi apresentado em (MSVV05) como tendo uma competitividade de 1/2 numa análise de pior caso. Entretanto, considerando que as consultas pudessem ter sua ordem alterada de forma aleatória, em (GM08) está provado que a competitividade deste algoritmo passa para $1 - 1/e$.

No restante da dissertação vamos nos referir a estes algoritmos como:AdWords, Primal-Dual e Guloso, respectivamente.

1.3.3 Sobre o Uso de Informações Estocásticas

Segundo (MNS07), as empresas que implementam as máquinas de busca possuem um grande histórico de informações que permite uma boa estimativa para a expectativa de consultas em um dado período. Também acena para o risco de confiar demais no passado e ser surpreendido por eventos inesperados.

Baseado nestas duas considerações, o artigo combina o algoritmo AdWords (MSVV05) com um oráculo para obter um melhor resultado final, sem perder a garantia da competitividade em momentos de variações imprevistas nos dados de entrada. O algoritmo contem ainda um parâmetro α que pode ser alterado a qualquer momento para dar maior ou menor credibilidade ao oráculo. O artigo não discute, entretanto, como construir este oráculo.

1.4 Contribuições

Nesta dissertação avaliamos experimentalmente os algoritmos para seleção de links patrocinados propostos na literatura em comparação com a solução ótima offline. Embora estes algoritmos tenham sido propostos para o problema definido com limite de orçamento por anunciante, verificamos

PUC-Rio - Certificação Digital Nº 0711297/CA

através de simulações com dados sintéticos que a competitividade se manteve superior a $1 - 1/e$, mesmo com a mudança da restrição para um limite de exibições por anunciante.

Outro resultado importante foi a identificação dos cenários em que o algoritmo Guloso tem um desempenho melhor e em quais é mais indicado um algoritmo que tem como estratégia preservar os limites dos anunciantes, como os algoritmos AdWords e Primal-Dual.

Propomos ainda dois algoritmos baseados em informações estocásticas. Considerando a expectativa de que a frequência de consultas seja parecida em dias subsequentes, estes algoritmos utilizam a frequência histórica de cada tipo de consulta para prever o que ocorrerá durante um novo dia.

O primeiro, que chamamos de Preditivo, é determinístico e apresentou excelente desempenho quando temos mais consultas que o limite total de exibições dos anunciantes. O segundo, que chamamos Aleatório e utiliza o histórico de consultas como base para sorteio, apresentou ótimo resultado na situação inversa, isto é, quando temos menos consultas que capacidade total dos anunciantes.

Em todos os cenários experimentais, um dos dois algorimos que utilizam predição mostrou-se superior ou igual aos algoritmos que não utilizam este recurso. Em nossas simulações, um algoritmo híbrido, que utilizasse a estratégia de um ou outro, de acordo com a expectativa total de consultas, seria sempre o de melhor desempenho.

1.5 Organização do Texto

Por se tratar de um problema online, no Capítulo 2 abordamos as dificuldades em resolver este tipo de problema e os métodos mais frequentemente utilizados em sua análise e solução.

No Capítulo 3 definimos o problema substituindo a restrição do orçamento por um limite de exibições por anunciante e descrevemos os algoritmos offline usados para cálculo da solução ótima. Por fim discutimos sobre a competitividade de um algoritmo online para este modelo que utiliza o limite de exibições.

A descrição dos algoritmos adaptados da literatura técnica e dos algoritmos propostos, que utilizam informações estocásticas para previsão na seleção dos links, está no Capítulo 4.

PUC-Rio - Certificação Digital Nº 0711297/CA

A avaliação experimental dos algoritmos online utilizando dados sintéticos é apresentada no Capítulo 5. Neste capítulo descrevemos também o ambiente de simulação, a geração dos dados de entrada, a medição do desempenho e a evolução gráfica diária dos algoritmos implementados.

Concluímos a dissertação com uma discussão sobre o desempenho geral dos algoritmos nos experimentos e em que casos é melhor a utilização de cada um destes algoritmos. Falamos um pouco sobre a experiência com o uso de predição nos algoritmos e apontamos a possibilidade de trabalhos futuros nesta área.

PUC-Rio - Certificação Digital Nº 0711297/CA