Principais variáveis na ordenação de anúncios

André Henrique Serafim Casimiro

Texto apresentado
ao
Instituto de Matemática e Estatística
da
Universidade de São Paulo
para
obtenção do título
de
Mestre em Ciências

Programa: Ciência da Computação

Orientador: Prof. Dr. João Eduardo Ferreira

Coorientador: Dr. Marcos Eduardo Bolleli Broinizi

São Paulo, abril de 2016

Principais variáveis
na ordenação de anúncios

Esta é a versão original da dissertação elaborada pelo candidato André Henrique Serafim Casimiro, tal como submetida à Comissão Julgadora.

Agradecimentos

Agradeço primeiramente a Jesus, minha Rocha, que tem me guiado em cada momento desde o início de minha vida, e não foi diferente do início ao fim deste trabalho de mestrado.

Agradeço à minha esposa Diana, por seu suporte e incentivo tão essenciais para finalização deste trabalho. Desde que estamos juntos tenho visto minha vida florescer em todos os aspectos, inclusive o acadêmico. Te amo.

Agradeço a meus pais pelo amor, carinho e dedicação que sempre recebi. Agradeço também pela educação que com tanto esforço me proporcionaram. Agradeço à minha irmã pelo companheirismo e amizade tão preciosos.

Agradeço ao meu orientador Prof. Dr. João Eduardo Ferreira pela direção, aconselhamento e flexibilidade no desenvolvimento deste trabalho, por ser um mentor mas também um amigo, e por sua perseverança incansável na busca de soluções para problemas reais, os quais nem sempre domina. Agradeço também ao meu co-orientador Dr. Marcos Broinizi pelas horas de dedicação e mentoria despendidas, pela confiança acadêmica e profissional depositada em mim, além da amizade tantas vezes demonstrada em conversas e em conselhos de vida.

Resumo

CASIMIRO, A. H. S. **Principais variáveis na ordenação de anúncios**. 2016. 80 f. Dissertação - Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2015.

A rápida popularização do acesso a Internet desde a década de 90 atraiu cada vez mais pessoas para a utilização da rede. Com o aumento deste público, naturalmente surgiu o interesse pela exibição de material publicitário nas páginas hospedadas. Semelhantemente ao que já acontecia em outras mídias (e.g. jornais, revistas, TV, rádio), a exibição de publicidade consolidou-se como uma das principais fontes de renda dos publicadores de conteúdo. Em certo sentido, foi a publicidade que financiou grande parte do desenvolvimento da Internet, pois permitiu que conteúdo e serviços cada vez mais sofisticados fossem oferecidos gratuitamente para os usuários. Desde seu início, a publicidade computacional tem acompanhado o crescimento da própria Internet, e hoje é um mercado multibilionário. Um dos grandes fatores para tamanho sucesso foi a personalização dos anúncios exibidos com base no conteúdo da página e dos interesses do usuário, o que criou um ecossistema favorável não somente aos interesses de anunciantes e publicadores, mas também, e principalmente, do usuário. São inúmeros os modelos de publicidade computacional, variando desde o formato visual até o modelo de cobrança. Sabe-se, no entanto, que quanto melhor a contextualização do anúncio com os interesses do usuário maiores são as chances atrair sua atenção e, portanto, de gerar receita. A escolha dos anúncios a serem exibidos é modelada por um problema de ordenação por relevância que contempla inúmeras variáveis de contexto, por exemplo: a semelhança com a página, o desempenho histórico do anúncio, o valor de receita a ser gerado, o modelo de pagamento escolhido, os dados do usuário, seu padrão de navegação, etc. Muitos trabalhos científicos já foram publicados nesta área explorando separadamente cada um das variáveis envolvidas, mas há ainda uma lacuna de entendimento sobre como elas se comportam quando combinadas. Esta lacuna precisa ser preenchida, pois muitas vezes a definição da regra de ordenação é feita por especialistas de domínio, que precisam entender como combinar as variáveis de modo a obter um melhor desempenho dos sistemas de publicidade. Nosso trabalho estuda o problema de ordenação de anúncios e propõe uma análise dessas variáveis de ordenação baseada em técnicas estatísticas aplicadas: teste A/B e análise de componentes principais. Ela fornece ao especialista de domínio insumos essenciais para compreensão dos pontos fortes e fracos de seu sistema de publicidade, indicando caminhos para evolução do mesmo. Para aferir a validade de nossa proposta, realizamos um experimento em ambiente real e apresentamos as conclusões obtidas. Apresentamos também algumas das análises extras realizadas como desdobramentos de nossa proposta inicial.

Palavras-chave: publicidade computacional, análise de componentes principais, dados de usuários, aposta, contexto.

Abstract

CASIMIRO, A. H. S. **Ad ranking main variables**. 2016. 80 f. Dissertação - Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2016.

The rapid popularization of the Internet access since the 90s attracted increasingly more people to use the network. Due to this increased public, an interest naturally grew by the display of advertising material on hosted web pages. Similar to what was already happening in other media (e.g. newspapers, magazines, TV, radio), display ads established itself as a major source of income for online content providers. Somehow, it was the advertising that funded much of the Internet development, enabling providers to build increasingly sophisticated content and services free of charge to the end user. Since its inception, computational advertising followed the Internet's growth itself, and is now a multibillion-dollar market. One of the major factors for such success was the customization of displayed ads based on page content and user interests, which created a favorable ecosystem not only to the interests of advertisers and publishers, but also and mainly, to the user. There are countless computational advertising designs, ranging from the visual format to the billing model. It's a known fact, however, that the better the matching between the ad's content and user's interest greater the odds of cathing his attention and, therefore, making profit. Choosing which ads to show is modeled by a relevance ranking problem including many context variables, for instance: page textual similarity, ad historical performance, revenue generation, payment model, user data, his browsing patterns, etc. Many papers have already been published in this field exploring each one of the involved variables separately, but there is still a lack of understanding of how they behave when combined. This needs to be overcome, for in many many cases the ranking formula is handcrafted by domain experts, who needs to understand the better way to combine all the different variables so as to obtain the most performance of theirs advertising systems. Our work studies the ad ranking problem and proposes an analysis on the ranking variables based on applied statistical techniques, such as A/B testing and principal component analysis. The analysis provides the domain expert with valuable insights for understanding strengths and weaknesses of the advertising system, pointing paths to evolve it. We assess the proposal's validity running online experiments on a real environment and present obtained conclusions. We also present some extra analysis executed as outspread of our initial proposal.

Keywords: digital advertising, principal component analysis, user data, bid, context.

Sumário

Lista de Abreviaturas ix

Lista de Figuras xi

Lista de Tabelas xiii

1 Introdução 1

1.1 Problema . . . 4
1.2 Objetivo . . . 5
1.3 Proposta . . . 6
1.3.1 Etapa 1 - Análise isolada por Teste A/B . . . 6
1.3.2 Etapa 2 - Análise combinada por PCA . . . 6
1.4 Organização do Trabalho . . . 6

2 Fundamentos 9

2.1 Publicidade Computacional . . . 9
2.1.1 Contextos de publicidade . . . 9
2.1.2 Formato visual . . . 13
2.1.3 Modelos de pagamento . . . 15
2.1.4 Modelo de ordenação teórico . . . 16
2.1.5 Anatomia de um anúncio . . . 17
2.2 Sistema de exibição de anúncios . . . 20
2.2.1 Visão geral . . . 20
2.3 Reordenação de anúncios . . . 23
2.3.1 Métricas de desempenho . . . 24
2.3.2 Função objetivo . . . 25
2.4 Teste A/B . . . 27
2.5 Análise de Componentes Principais . . . 28
2.5.1 Definição . . . 28
2.5.2 Procedimento . . . 28
2.5.3 Discussão . . . 29

3 Trabalhos Relacionados 31

4 Proposta e Análises **37**
4.1 Escopo do trabalho . . . 37
4.2 Análise de variáveis em ordenação de anúncios . . . 38
4.2.1 Etapa 1 - Teste A/B . . . 38
4.2.2 Etapa 2 - Análise de componente principais . . . 39
4.3 Experimento 1: validação da análise . . . 40
4.3.1 Validação da Etapa 1 . . . 41
4.3.2 Validação da Etapa 2 . . . 43
4.4 Relevância da proposta . . . 46
4.5 Experimento 2: combinação de variáveis . . . 46
4.6 Experimento 3: desempenho ao longo do tempo . . . 48

5 Conclusão **55**
5.1 Contribuições . . . 55
5.2 Publicações . . . 56
5.3 Trabalhos futuros . . . 57

A Implementações das análises **59**
A.1 Implementação da análise de desempenho . . . 59
A.2 Implementação da análise de PCA . . . 63

Lista de Abreviaturas

CPM	Custo por milhar de impressão, custo por mil visualizações (*Cost per mille*)
CPC	Custo por clique (*Cost per click*)
CPA	Custo por ação (*Cost per action*)
CTR	Taxa de clique (*Click-through rate*)
ROI	Retorno sobre o investimento (*Return over investment*)
SA	Sistema de exibição anúncios

Lista de Figuras

1.1 Gasto com publicidade digital no mundo, 2010-2016 bilhões de dólares, % de alteração, % do gasto em relação a outras mídias. 2
1.2 Distribuição do mercado publicitário por formatos, 2006-2013 (% da receita total) Busca, Banner, Dispositivos móveis, Classificados, Vídeo, Geração de Leads, Mídia Rica. 3

2.1 Exemplo do modelo tradicional de publicidade online em que um blog sobre casamento faz publicidade de seus parceiros. 10
2.2 Exemplo de busca patrocinada, onde vendedores de celulares apostam em termos da consulta "galaxy S2". 12
2.3 Exemplo de publicidade contextualizada, site sobre aulas particulares exibe anúncio sobre cursos de graduação de um anunciante da rede de anúncios *Google AdSense*. . . 12
2.4 Funil das etapas de conversão de usuários e a relação com cada modelo de rentabilização. 16
2.5 Estrutura hierárquica das informações de um anúncio. 18
2.6 O funcionamento de um sistema de anúncios. 20
2.7 Processo de uma requisição. O navegador requisita uma página web e os anúncios advindos do SA são inseridos via Javascript. 23
2.8 Componentes principais em um conjunto de pontos 30

3.1 Visualização 3D das variáveis CTR, CTX e BID normalizadas 36

4.1 Desempenho diário das estratégias no teste A/B. 42
4.2 Valor diário da função objetivo para os parâmetros $\alpha = \beta = \gamma = \frac{1}{3}$ e $\delta = 1$ 42
4.3 Valor diário da função objetivo para os parâmetros $\alpha = \beta = \gamma = \frac{1}{3}$ e $\delta = 0$ 43
4.4 *Loading plots* das componentes principais 45
4.5 Desempenho diário das estratégias no experimento 2. 47
4.6 Valor diário da função objetivo para os parâmetros $\alpha = \beta = \gamma = \frac{1}{3}$ e $\delta = 1$ no experimento 2. 48
4.7 Valor diário da função objetivo para os parâmetros $\alpha = \beta = \gamma = \frac{1}{3}$ e $\delta = 0$ no experimento 2. 48
4.8 Desempenho diário das estratégias no experimento 3. 50
4.9 Desempenho diário das estratégias stCTR e stLCTR no experimento 3. 51
4.10 Valor diário da função objetivo para os parâmetros $\alpha = \beta = \gamma = \frac{1}{3}$ e $\delta = 1$ no experimento 3. 52

4.11 Valor diário da função objetivo para os parâmetros $\alpha = \beta = \gamma = \frac{1}{3}$ e $\delta = 0$ no experimento 3. 52
4.12 Valores obfuscados diários de impressões, cliques e receita das estratégias vencedoras do experimento 3. 54

Lista de Tabelas

2.1 Exemplos de diferentes formatos visuais de anúncios. 14
2.2 Entidades que compõe os dados de um anúncio. 19
2.3 Lista de atividades representadas na figura 2.6 21
2.4 Variáveis utilizadas pelo sistema estudado para reordenação de anúncios. 24
2.5 Métricas representativas do interesse de cada participante do sistema. 25
2.6 Terminologia usada em testes A/B 27

3.1 Tabela com a distribuição dos artigos utilizados em nossa pesquisa. 31

4.1 Matriz de transofrmação dos vetores da base original na nova base. 43
4.2 Conclusões obtidas na validação de nossa proposta 44
4.3 Estratégias de ordenação do segundo experimento. 46

Capítulo 1

Introdução

No mundo atual estamos cercados por anúncios publicitários, dos mais criativos, aos mais tradicionais, eles estão por toda parte e são essenciais para o estímulo do consumo e, portanto, da economia. Na Internet não é diferente, eles estão presentes em sites de busca, portais de informação, notícias, redes sociais e até mesmo blogs pessoais. A exibição de publicidade online é um mercado multibilionário que estimulou o desenvolvimento da Internet desde sua concepção. Muitas empresas da Internet, inclusive grandes como *Google* e *Yahoo!*, tem como principal fonte de receita a exibição de publicidade online. Segundo Shanaham e Kurra, desde seu início em 1994, a indústria de propaganda online manteve um crescimento médio de dois dígitos ao ano, se tornando uma indústria de $65 bilhões de dólares em 2009 [SK11]. Como podemos observar na figura 1.1 [1], extraída do artigo de julho de 2014 do *www.emarketer.com* [Ema14], os gastos mundiais com publicidade online superaram $120 bilhões de dólares em 2013 e a estimativa de crescimento é de cerca de $20 bilhões em 2014 e mais $20 bilhões em 2015, chegando a representar quase 30% de todo dinheiro gasto com publicidade em 2016. Tamanha expressividade explica como é possível que empresas ofereçam serviços e aplicativos online lucrando somente com publicidade.

O sucesso de um produto ou serviço depende muito de sua estratégia de divulgação, e não é difícil encontrar projetos em que se gasta mais dinheiro com a divulgação do que no próprio desenvolvimento do produto. Muitos são os canais de divulgação disponíveis, tais como, jornal, revista, TV, internet, outdoor, entre muitos outros; sendo cada um mais apropriado para o tipo de público que se deseja atingir. As equipes de divulgação são responsáveis por definir campanhas que sejam eficientes em termos dos recursos disponíveis. Elas fazem a segmentação de clientes, definem seu público-alvo, criam peças publicitárias que capturem a atenção e escolhem o canal de divulgação mais apropriado para a audiência que querem atingir. São muitas variáveis que determinam o sucesso ou fracasso de uma campanha publicitária e nesse cenário a publicidade online tem se posicionado estrategicamente de forma a minimizar os riscos e aumentar a performance de cada centavo investido. Os anunciantes tem optado cada vez mais pela cobrança baseada em performance (CPC e CPA), e pagam apenas quando há evidências de que o usuário se interessou por seu anúncio (e.g. efetuar um clique ou realizar uma compra).

Segundo o relatório de 2013 da IAB [IAB14], em 2013 nos EUA, pela primeira vez os gastos com publicidade na Internet superaram os gastos com publicidade em TV aberta. As tendências para a área ainda são de forte crescimento, com destaque nos últimos anos para a publicidade

[1] Inclui publicidade apresentada em computadores de mesa e laptops bem como telefones móveis e tablets, e inclui todos os diversos formatos de publicidade nessas plataformas; excluindo publicidade de mensagens SMS, MMS e P2P

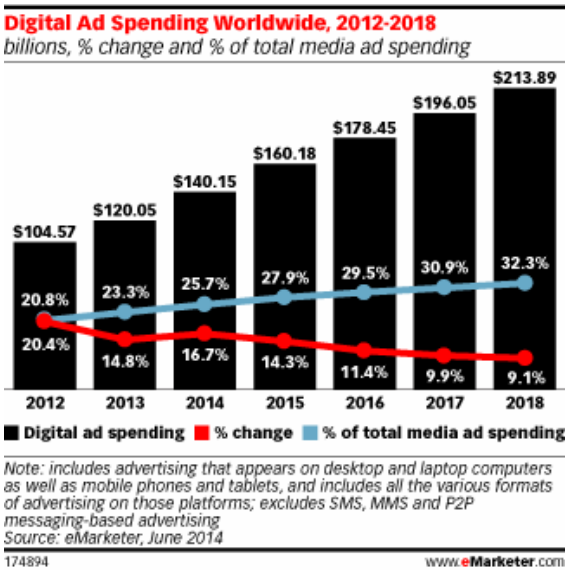


Figura 1.1: *Gasto com publicidade digital no mundo, 2010-2016*
bilhões de dólares, % de alteração, % do gasto em relação a outras mídias.

em dispositivos móveis, como mostra a figura 1.2 [2]. Essa alta taxa de crescimento, e o grande aumento de dispositivos que acessam a Internet, irão consolidar os sistemas de publicidade online como o principal veículo de publicidade mundial, e o mais pervasivo, uma vez que ele atinge toda a crescente população conectada à rede pelos mais diversos tipos de dispositivos. Essa universalização da Internet, independente dos dispositivos, dá aos anunciantes a possibilidade de concentrar seus recursos na divulgação pela Internet, tendo certeza de que estarão atingindo grande parte de seu público alvo.

Sistemas de publicidade online mais recentes permitem que os anunciantes criem suas campanhas de divulgação e segmentem a exibição somente a um público-alvo específico. Os anunciantes não precisam determinar as páginas em que ela será exibida, isso é feito automaticamente pelo sistema de veiculação de anúncios, que se utiliza de técnicas baseadas em dados coletados do contexto de navegação do usuário para exibir publicidade direcionada.

A escolha da melhor publicidade a ser exibida em uma página é estudada por uma área de pesquisa científica. Shanahan e Kurra definem o termo ***Publicidade Digital*** no capítulo *Digital Advertising: An Information Scientist's Perspective* [SK11]. Josifovski e Broder, por sua vez, referenciam a área como ***Publicidade Computacional*** [BJ11], e é este termo que também utilizaremos. Segundo esses autores, Publicidade Computacional é composta pela intersecção de disciplinas de processamento de texto, recuperação de informação, modelagem estatística, aprendizado computacional, teoria dos jogos e de leilões, ordenação, otimização, microeconomia e sistemas de recomendação. Publicidade Computacional tem atraido muita atenção, e o número de trabalhos publicados sobre o assunto teve um crescimento muito expressivo nos últimos anos. Ela surgiu, como diversas outras áreas de pesquisa, da união dos conhecimentos tradicionais ao poder computacional (e.g. Biologia Computacional, Física Computacional, Linguística Computacional, etc).

Podemos definir publicidade como uma mensagem destinada a um público-alvo que tem como objetivo convencer o receptor acerca do objeto anunciado, seja para alterar positivamente sua

[2] As definições dos formatos podem ter mudado ao longo do período retratado, tanto devido ao processo de pesquisa quanto a interpretação dos respondentes.

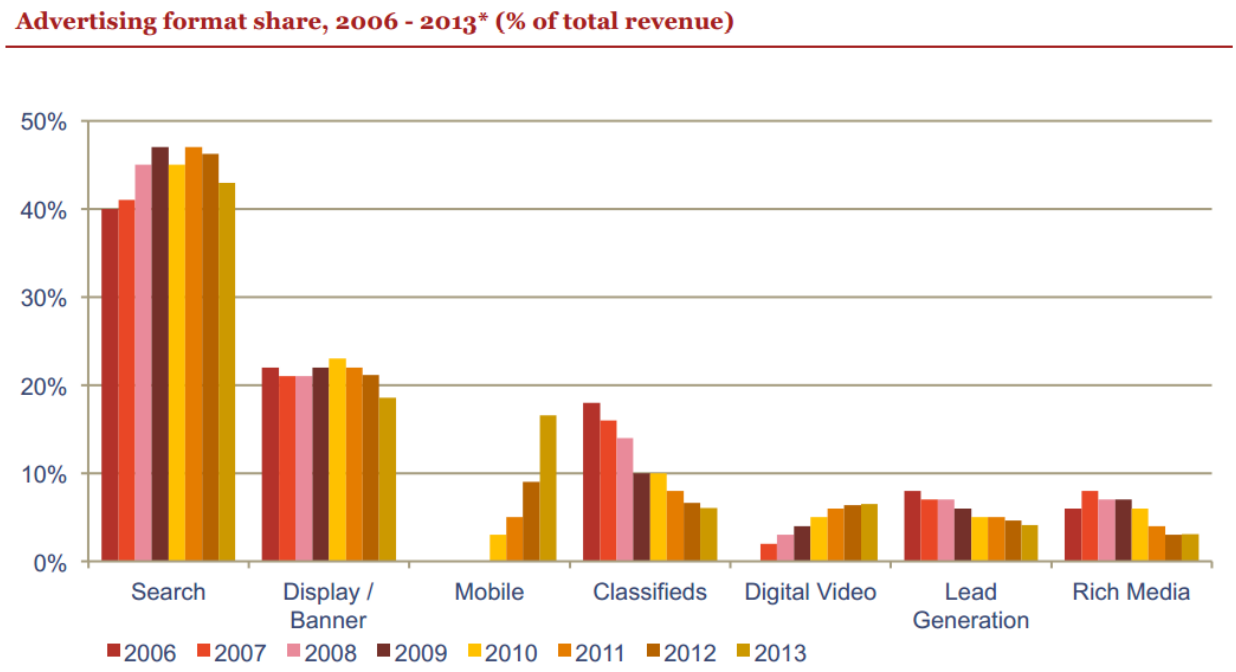


Figura 1.2: *Distribuição do mercado publicitário por formatos, 2006-2013 (% da receita total) Busca, Banner, Dispositivos móveis, Classificados, Vídeo, Geração de Leads, Mídia Rica.*

percepção acerca deste (e.g. opinião sobre uma marca), ou influenciá-lo a tomar uma ação (e.g. comprar um produto). A publicidade feita em meios digitais tem o mesmo objetivo, mas utiliza-se de técnicas computacionais para aumentar os ganhos do anunciante, do usuário e do publicador (veiculador do anúncio) e ela visa resolver o seguinte problema:

Dado um usuário em um determinado contexto, encontrar o anúncio que melhor se adeque dentre o conjunto dos anúncios disponíveis.

Em que explicitamos os seguintes conceitos:

- ***usuário:*** a pessoa que está navegando em uma página que terá a exibição de anúncios publicitários, bem como seus interesses e informações sobre perfil de navegação;
- ***contexto:*** a página acessada pelo usuário, que compreende o título, descrição, palavras-chave, texto, categorias, entre outros;
- ***conjunto de anúncios disponíveis:*** todos os anúncios cadastrados sistemicamente pelos anunciantes, disponíveis para exibição no momento da requisição feita pelo usuário;
- ***melhor adequação de anúncio:*** qualidade subjetiva que descreve o quanto um anúncio satisfaz os interesses do anunciante, do usuário e do publicador.

Sistemas de veiculação de publicidade online dependem da participação de anunciantes, usuários e ao menos um publicador; eles devem, portanto, ser vantajosos para cada uma dessas partes. O interesse de cada participante é diferente, sendo, possivelmente, conflitantes. Mais explicitamente, temos:

- o ***usuário*** acessa a Internet com a intenção de obter informações, esperando ter uma boa experiência pessoal e tem potencial de adquirir algo;
- o ***anunciante*** disponibiliza informação sobre um bem ou serviço que deseja oferecer a um usuário e possui recursos para investir em publicidade, esperando receber acessos e vendas (ou conversões) advindos dessa publicidade;
- o ***publicador*** exerce papel de intermediário, possui alta audiência e deseja aumentar a receita exibindo publicidade e a audiência atraindo usuários.

Nesse contexto de múltiplos interesses a escolha do melhor anúncio determina fortemente quais os interesses listados acima serão mais bem atendidos. O desafio, então, é manter o sistema em equilíbrio de modo que todos os participantes continuem interessados. Neste trabalho estudamos o problema de escolha do melhor anúncio disponível e propomos uma análise que suporta a tomada de decisões concernentes a estratégia que o sistema deve utilizar para escolher os melhores anúncios para exibição.

1.1 Problema

A veiculação de anúncios na velocidade e escala que a Internet requer exige a criação de uma grande infrestrutura de serviços computacionais, e é comum que sistemas de anúncios estejam sujeitos a restrições da ordem de bilhões de requisições por hora e milhões de usuários únicos por mês. Para satisfazê-las eles são construídos semelhantes a sistemas de busca, realizando ações como *crawling* da página destino, indexação dos anúncios cadastrados, recuperação, ordenação e exibição dos anúncios recuperados.

A abordagem tradicional para escolher os melhores anúncios infere uma ordenação nos anúncios disponíveis e seleciona os primeiros, considerados mais relevantes. Por exemplo, se uma página possui 3 espaços disponíveis para publicidade, os 3 primeiros da lista de anúncios ordenada por relevância serão selecionados para exibição. A escolha do anúncio que melhor se adequa ao contexto do usuário (ou simplesmente melhor anúncio) é crucial para o desempenho de sistemas de veiculação de publicidade. Considere, por exemplo, uma ordenação que favorecesse o interesse financeiro do publicador. Ela aumentaria o lucro mas, a longo prazo, provavelmente menos usuários se interessariam pelos anúncios veiculados, uma vez que os mesmos seriam pouco relacionados ao conteúdo da página; e isso, por sua vez, implicaria em degradação da efetividade do sistema. É extremamente importante a definição de uma ordenação que equilibre os interesses, e maximize o ganho dos participantes. Tal equilíbrio se torna ainda mais desafiador ao considerar o fato de que a quantidade de espaço publicitário em cada página comporta poucos anúncios, enquanto existem milhares disponíveis para exibição.

A ordenação de anúncios é feita por meio de variáveis relevantes que representam os interesses dos participantes. Cada uma dessas variáveis é uma medida objetiva extraída de cada anúncio que, quando combinados, resultam em uma estratégia de ordenação de anúncios. Neste trabalho estudaremos quatro dessas variáveis e seu impacto no desempenho geral do sistema. São eles: a aposta, que é o valor máximo pago por um anunciante por um clique em seu anúncio (BID), a taxa histórica de clique do anúncio (CTR, ou *Click Through Rate*), o *score* de relevância com o contexto

da página (CTX) e o *score* de relevância com o usuário (BTU, ou *behavioral targeted user*). Podemos agora reescrever a definição dos interesses de cada participante em termos dessas variáveis:

- o ***usuário*** acessa a Internet com a intenção de obter informações (CTX), esperando ter uma boa experiência pessoal (BTU) e tem potencial de adquirir algo (CTR);
- o ***anunciante*** disponibiliza informação sobre um bem ou serviço que deseja oferecer a um usuário e possui recursos para investir em publicidade, esperando receber acessos e vendas (ou conversões) advindos dessa publicidade (CTR);
- o ***publicador*** exerce papel de intermediário, possui alta audiência e deseja aumentar a receita exibindo publicidade (BID) e também a audiência atraindo usuários (CTR).

Essas variáveis são números reais que são combinados em fórmulas utilizadas para a ordenação dos anúncios. Uma das primeiras fórmulas utilizadas, por exemplo, foi:

$$rank = BID * CTR$$

As possibilidades de combinações das variáveis em fórmulas matemáticas são infinitas, e a escolha da fórmula de ordenação utilizada é geralmente feita de maneira empírica por especialistas do domínio. Vimos que estabelecer uma boa ordenação de anúncios pode determinar o sucesso ou fracasso de todo o sistema de anúncios. Como, então, podemos ter certeza que uma alteração na fórmula de ordenação trará ganhos para os participantes do sistema? Como os especialistas podem avaliar os resultados dessas mudanças? Será que eles podem prever os resultados que uma alteração de fórmula terá? Com essas perguntas em mente, chegamos então ao problema que nosso trabalho busca resolver:

Como fazer alterações na função de ordenação de anúncios baseado em dados reais a fim de melhorar o desempenho do sistema de publicidade computacional?

Neste trabalho propomos uma avaliação do desempenho de funções de ordenação e o validamos no ambiente de produção do UOL Cliques, o sistema de anúncios de um dos maiores portais da internet brasileira.

1.2 Objetivo

Neste trabalho estudamos a infraestrutura necessária para veiculação de anúncios e discutimos não apenas o aspecto técnico mas também aspectos relacionados a área comercial, como modelos de pagamento e o impacto da ordenação de anúncios nos interesses dos participantes. Nosso objetivo principal, entretanto, é solucionar o problema descrito na seção 1.1. Muitos trabalhos já foram feitos em busca de se obter melhores desempenhos quanto às funções de ordenação utilizadas em sistemas de publicidade computacional. Entretanto, geralmente eles consideram apenas melhorias individuais nos aspectos de ordenação. Por exemplo os trabalhos de Lei Wang et al. [WYZ11] e de Andrei Broder et al. [BFJR07] propõem melhorias na contextualização dos anúncios com a página, já os trabalhos de Jian Tang et al. [TLY+11] e de Sarah Tyler et al. [TPGJ11] descrevem maneiras de se capturar os interesses do usuário e utilizá-los na escolha dos anúncios. Tais abordagens têm ampla aceitação

em conferências e os ganhos de desempenho relatados chegam a 80%. Nosso estudo, no entanto, não se concentra em nenhuma das variáveis de maneira específica, mas em como combiná-las da melhor maneira possível, ou seja, que atenda aos interesses do anunciante, do usuário e do publicador.

Na seção 4.2 descrevemos a análise das variáveis de ordenação de anúncios que esclarece a contribuição de cada uma no desempenho geral do sistema, bem como as interações entre eles. Em seguida, na seção 4.3, mostramos os resultados obtidos na aplicação de nossa análise em um sistema real de publicidade computacional.

1.3 Proposta

Visando a solução de nosso problema inicial, que é auxiliar especialistas na definição e modificação das estratégias (fórmulas) de ordenação dos anúncios, apresentamos uma análise das variáveis de ordenação de anúncios composta por duas etapas:

1. medir o impacto *isolado* de cada variável nos interesses de cada participante;
2. medir expressividade e correlação das variáveis *combinadas* entre si.

Estas duas medições fornecem insumos importantes para que os especialistas de domínio que gerenciam os sistemas de publicidade possam projetar a evolução do sistema.

1.3.1 Etapa 1 - Análise isolada por Teste A/B

A primeira etapa é a realização de um teste A/B que configura uma estratégia de ordenação para cada variável disponível. As fórmulas são configuradas de modo que a ordenação dos anúncios seja feita exclusivamente por uma única variável. Além dessas acrescentamos uma estratégia de ordenação aleatória para servir de base de comparação para as outras.

No caso do sistema que utilizamos para validação, por exemplo, a análise das quatro variáveis disponíveis (BID, CTX, CTR e BTU) exigiu a configuração das estratégias stBID, stCTX, stCTR, stBTU e stRND. Onde as 4 primeiras representam ordenações exclusivas pelas variáveis que as descrevem e a última, stRND, executa a ordenação aleatória.

1.3.2 Etapa 2 - Análise combinada por PCA

A segunda etapa é a análise de componentes principais (PCA, *Principal Component Analysis*) sobre as variáveis de ordenação. Consideramos os valores de cada uma dessa variáveis (números reais) como uma dimensão de um espaço vetorial e aplicamos a técnica de PCA para identificar as que mais influenciam a ordenação, bem como os relacionamentos entre elas.

1.4 Organização do Trabalho

Neste capítulo apresentamos uma descrição de todo nosso estudo em publicidade computacional. No Capítulo 2 são apresentados os fundamentos referentes a publicidade computacional: as variações de veiculação estão na seção 2.1, o funcionamento de um sistema de publicidade em 2.2, detalhes sobre a implementação utilizada para estudo da reordenação de anúncios em 2.3 e, por fim, os fundamentos de nossa análise em 2.4 e 2.5. No Capítulo 3 posicionamos nosso estudo ante as

diferentes frentes de pesquisa em publicidade computacional. No Capítulo 4 apresentamos nossa proposta para análise de componentes principais em ordenação de anúncios e a sua validação por meio de um experimento em um ambiente real 4.3, bem como em 4.5 e 4.6 apresentamos resultados e análises complementares que enriquecem nossas conclusões acerca das variáveis utilizadas na ordenação de anúncios. Por fim, no Capítulo 5 apresentamos nossas conclusões e considerações finais, as publicações obtidas e as possibilidades de trabalhos futuros.

Capítulo 2

Fundamentos

O objetivo principal de nosso trabalho é fornecer ao especialista de domínio de publicidade computacional uma análise das variáveis de ordenação, de modo que ele possa modificar a função de reordenação de anúncios de acordo com seus objetivos. Nossa análise foi concebida para o escopo de sistemas que veiculam anúncios de texto contextualizados e que tenham uma arquitetura semelhante a sistemas de busca. Descrevemos os fundamentos deste escopo nas primeiras seções deste capítulo (2.1 e 2.2), onde apresentamos os principais conceitos da área de publicidade computacional e da arquitetura dos sistemas de anúncios. Na seção 2.3 apresentamos com maior detalhe a fase de reordenação de anúncios, as variáveis envolvidas e as métricas de interesse de cada participante do sistema. Por fim, nas seções 2.4 e 2.5 apresentamos os fundamentos inerentes a aplicação das duas etapas da análise, a saber Teste AB na primeira etapa e o uso de PCA na segunda.

2.1 Publicidade Computacional

A história da publicidade computacional nos remete a própria história da Internet. Com o crescimento da Internet, se tornando parte integrante do dia a dia de muitas culturas, houve também maior interesse na veiculação de anúncios online. Assim, a grande quantidade de recursos investidos possibilitou um acelerado crescimento da área, que desenvolveu diversos modelos de veiculação de anúncios. Tais modelos variam desde o contexto de página em que serão exibidos, até os formatos visuais e o modelo de pagamento.

2.1.1 Contextos de publicidade

A evolução dos contextos de exibição de publicidade ocorreu naturalmente com o aparecimento de novas tecnologias e levou a exploração de novas oportunidades de publicidade. Essas oportunidades dizem respeito principalmente ao tipo de atividade, ou página, em que o usuário está quando visualiza o anúncio. A categorização que veremos a seguir é a principal diferença entre os modelos de publicidade online e surgiu com o passar dos anos. A pesquisa de Shatnawi et al. [SM12a] faz uma categorização semelhante e sumariza os principais desafios técnicos para cada modelo.

O primeiro modelo de publicidade na Internet surgiu em meados da década de 90 baseado nos formatos tradicionais de contratos de publicidade utilizados em outras mídias, como televisão e revistas. Nele, um anunciante paga para ter seu anúncio exibido em um espaço da página por um determinado período de tempo ou até um número máximo de exibições, ambos previamente

combinados. Esta modalidade é simples e segue uma negociação direta entre as partes, ela ainda é grandemente utilizada em portais de conteúdo e em blogs de nichos. A figura 2.1 mostra um blog sobre casamento que faz publicidade de parceiros. Chamaremos este modelo de ***exibição tradicional***. Em inglês ele é conhecido por *banner ads*, como no artigo de Ankit sobre a história da publicidade online [Obe13].

Figura 2.1: *Exemplo do modelo tradicional de publicidade online em que um blog sobre casamento faz publicidade de seus parceiros.*

Em 1998, Bill Gross, da empresa de busca Goto.com (que mais tarde passou a se chamar Overture e foi adquirida pela Yahoo por 1,63 bilhões de dólares), inventou o sistema *Paid Placement Model* (modelo de posicionamento pago). A principal novidade era que os anunciantes concorriam em um leilão para obter melhor posicionamento, quem pagasse mais pela exibição de um anúncio aparecia melhor posicionado juntos aos resultados da busca. Em termos das variáveis de ordenação, isto equivale a uma função de ordenação apenas pelo BID (descrita em 1.1).

O grande divisor de águas da publicidade computacional veio em 2000, quando a recém fundada Google, conhecida pela melhor experiência de busca adicionou ao modelo de Bill Gross o conceito de qualidade de um anúncio. Os anúncios eram posicionados não apenas pela aposta dos anunciantes (BID), mas também por sua relevância quanto ao assunto buscado pelo usuário. Conforme publicação em seu próprio blog, em fevereiro de 2002, essa variável nada mais era do que o CTR do anúncio [Goo02], hoje, entretanto, a qualidade certamente inclui diversos outros fatores. Este segundo modelo é conhecido como busca patrocinada (*sponsored search*) e desde então tem sido um grande sucesso. Ele continua evoluindo e é utilizado até hoje, a figura 2.2 mostra um exemplo em que vendedores de celulares apostam em uma consulta de interesse.

O momento que um usuário faz uma consulta em uma máquina de busca é muito propício para a exibição de publicidade pois ele certamente está procurando por algo, e há chances de que aquilo

que ele procura seja exatamente o que anunciante divulga. As páginas de busca são bons pontos de partida para navegação na Internet, mas o tempo que o usuário gasta navegando nelas é pequeno se comparado ao tempo gasto consumindo conteúdo na página alvo de sua busca. A publicidade, como em outras mídias, não aparece somente quando se está em busca de algo. Geralmente ela tem uma postura mais agressiva, tenta atrair atenção e despertar interesse pelo objeto anunciado. Sendo assim, não demorou muito até que se percebesse que qualquer página da internet poderia exibir anúncios relacionados ao seu conteúdo, e não somente a página de resultados de busca.

Para suprir essa demanda de exibição de publicidade surgiu o que chamamos de publicidade contextualizada (*contextual advertising*), que é a exibição de anúncios relacionados em qualquer página da Internet. Uma solução em escala para esse problema surgiu em março de 2003, quando a Google lançou a primeira rede de anúncios (*ad network*), o *Google AdSense* [Goo03]. As redes de anúncios agregam em uma única plataforma anunciantes e publicadores. Qualquer anunciante pode cadastrar um anúncio e qualquer publicador pode integrar seu site para exibição de publicidade. Quando um usuário requisita uma página do site do publicador, os anúncios são requisitados ao servidor de anúncios e inseridos na área reservada a publicidade. Este modelo de publicidade foi uma grande contribuição para o ambiente da internet, pois permitiu a criação de conteúdos e serviços totalmente gratuitos aos usuários, mantidos somente por publicidade. A figura 2.3 mostra o exemplo de um site que oferece um serviço gratuito para encontrar e divulgar aulas particulares e que obtém receita de publicidade computacional contextualizada.

A publicidade contextualizada é uma generalização do problema de busca patrocinada, no sentido que extende o desafio de contextualização dos anúncios a qualquer página da Internet, não somente a resultados de buscas. Além do significativo aumento de tráfego, é também mais difícil estimar com precisão se determinado anúncio é de interesse do usuário, visto que em busca patrocinada temos uma informação precisa do interesse do usuário naquele instante: a consulta que acabou de realizar, mas em publicidade contextualizada não. Na verdade, tal informação é inferida a partir do contexto (daí o nome publicidade contextualizada), como o conteúdo da página e o comportamento do usuário. As variáveis CTX e BTU representam uma forma de capturar a relevância de um anúncio com relação ao conteúdo da página e o comportamento do usuário, respectivamente.

Uma outra modalidade de publicidade que é bastante comum, são os ***classificados***, que ocorrem em sites com propósito de divulgação de anúncios em geral. Como exemplos brasileiros poderíamos citar Mercado Livre, OLX e Bom Negócio.

Os 4 tipos de anúncios acima destacados são os mais comuns e, de acordo com o relatório de 2012 da IAB [IAB13], juntos são responsáveis por mais de 90% da receita anual de publicidade online. Cada uma dessas modalidade tem suas particularidades e desafios. Do ponto de vista científico, há maior campo de exploração para a busca patrocinada e a publicidade contextualizada, e por isso eles tem recebido maior atenção, e contribuição da academia.

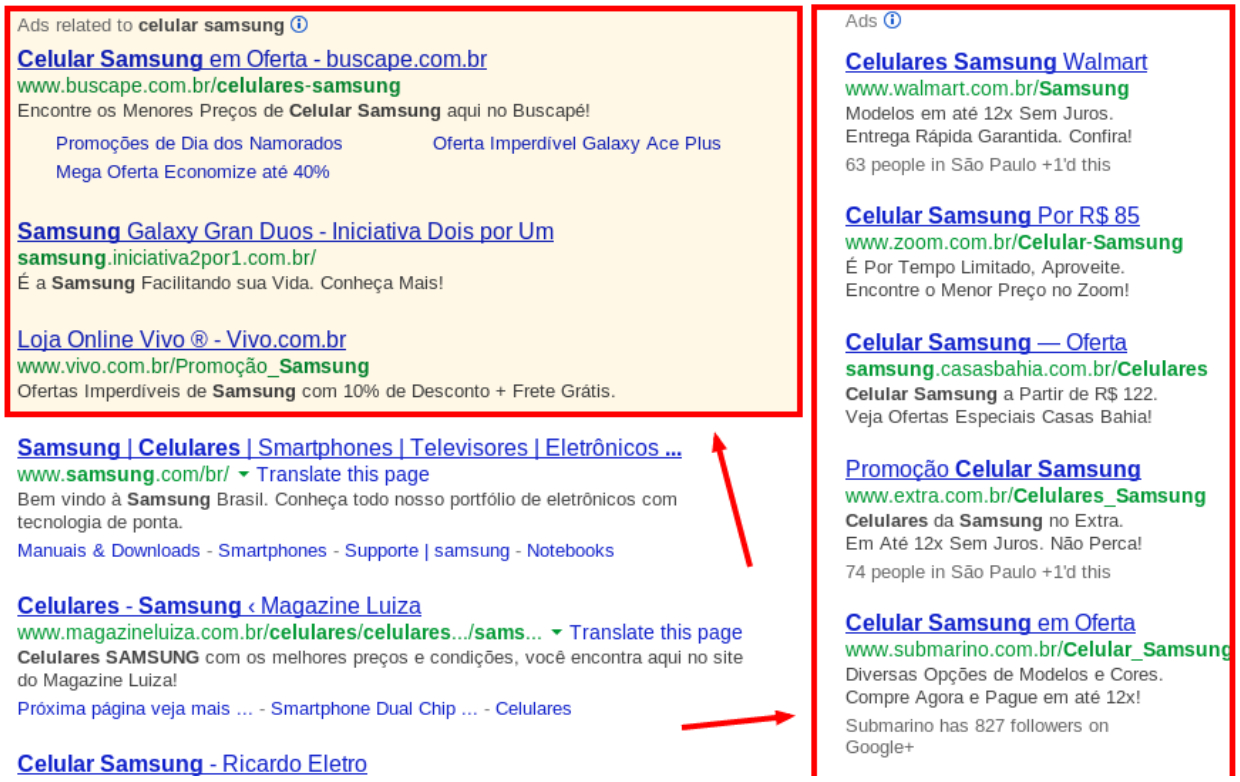


Figura 2.2: *Exemplo de busca patrocinada, onde vendedores de celulares apostam em termos da consulta "galaxy S2".*

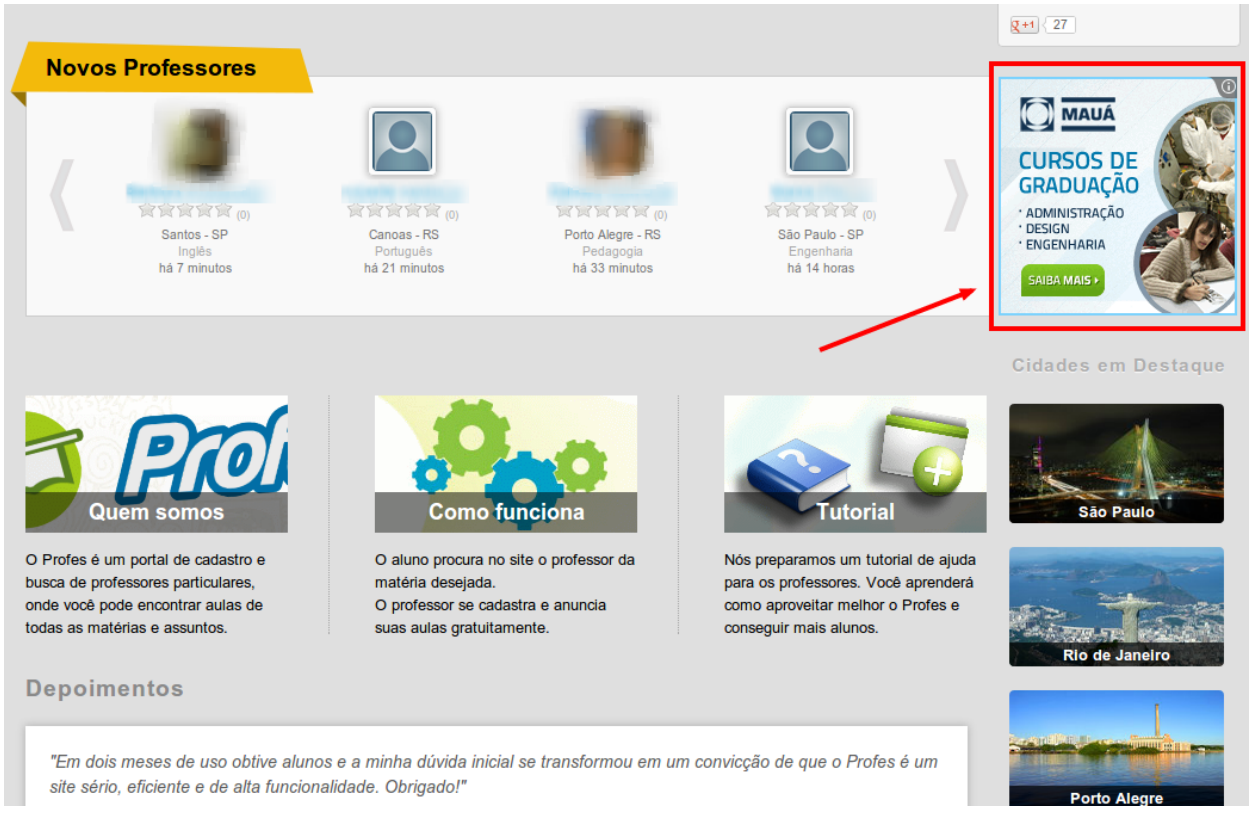


Figura 2.3: *Exemplo de publicidade contextualizada, site sobre aulas particulares exibe anúncio sobre cursos de graduação de um anunciante da rede de anúncios* Google AdSense.

2.1.2 Formato visual

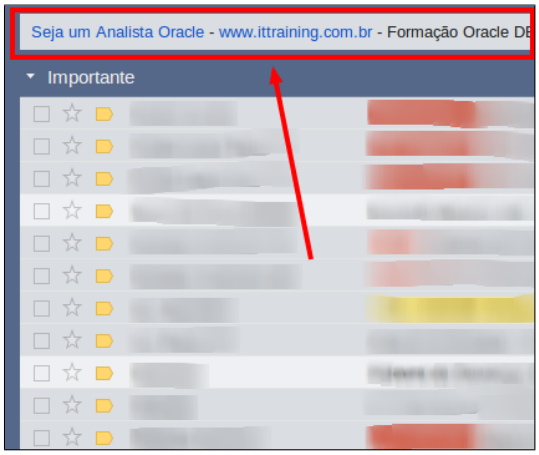
Um grande desafio da publicidade online é conseguir captar a atenção do usuário que visita a página pois, exceto no caso de classificados 2.1.1, publicidade raramente é o alvo principal de interesse do usuário. Dessa forma, uma boa apresentação visual é essencial para atrair atenção do usuário, e pode garantir que o mesmo opte por visualizar a oferta de um anunciante ao invés de seu concorrente. Para atingir este objetivo, é preciso que o anúncio seja exibido em um formato visual coerente com o local de exibição. Para cada ambiente virtual que se deseja exibir anúncios existem maneiras diferentes e mais apropriadas de se fazê-lo; e em cada caso, há trabalhos realizados para melhorar o entendimento e a performance comercial dos mesmos. [AZZ+12, CZA+12]

A seguir vamos enunciar uma lista dos principais formatos visuais existentes na internet, descrevendo sucintamente cada um deles. A tabela 2.1 contém imagens de exemplos dos formatos visuais mais populares.

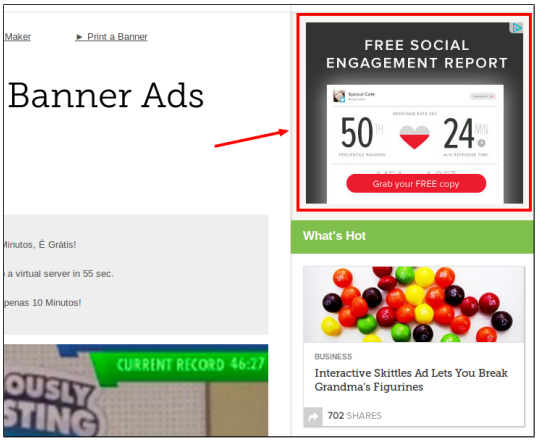
- **Texto**: composto por um título, uma ou duas linhas de descrição e um link de destino. Exibido ao lado de resultados de busca, clientes de e-mail, fórum, etc...
- **Imagem**: composto por uma imagem fixa e palavras-chave (não exibidas). Pode aparecer em qualquer página, é comum ser posicionado em uma barra lateral a direita.
- **Texto e imagem**: exatamente como um anúncio de texto mas com a possibilidade de adicionar uma imagem pequena. Pode ser exibido em muitos lugares diferentes.
- **Flash**: uma aplicação em flash simples. Geralmente faz uso de movimentação, interação com o mouse e até sons.
- **Vídeo**: antes da exibição de um vídeo é exibido um vídeo curto com a publicidade de algum anunciante.
- **Texto sobre vídeo**: anúncios de textos são exibidos em caixas flutuantes sobre a imagem enquanto o vídeo passa.
- **Texto em celular**: anúncios de texto embutidos em aplicativos para smartphones.
- **Imagem em celular**: anúncios de imagem embutidos em aplicativos para smartphones.
- **Tela de redirecionamento**: quando o usuário clica em um link é levado para uma tela em que publicidade é exibida. Depois de alguns segundos é então redirecionado para o destino do link.
- **Pop-up externo**: uma janela secundária do navegador é aberta contendo publicidade. É um dos formatos mais antigos e, por ser demasiadamente intrusivo, é irritante ao usuário; caindo em desuso.
- **Pop-up interno**: janela flutuante embutida no corpo do código da página.

Estes formatos listados não são os únicos, sendo comum também combinações destes formatos dependendo da página de exibição. Grandes redes anúncios, como o *Google AdSense*, possuem variados formatos de exibição de anúncios e dependendo da configuração e do contexto, são capazes de

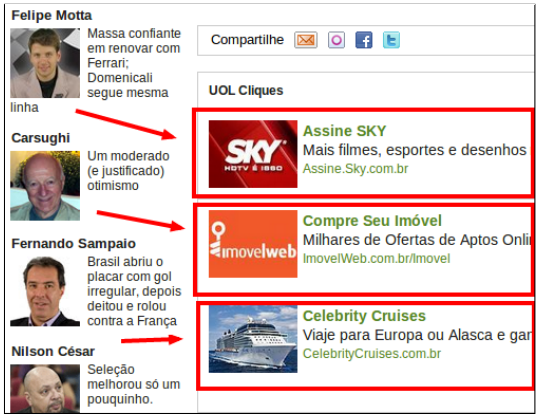
Texto



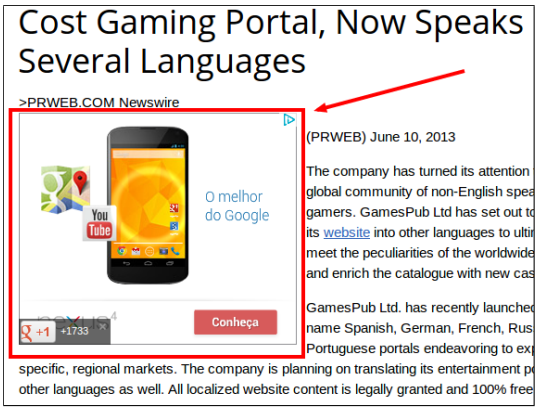
Imagem



Texto e imagem



Flash



Vídeo

Texto em celular

Tabela 2.1: *Exemplos de diferentes formatos visuais de anúncios.*

exibir o mesmo anúncio em um formato ou outro. Isso abre novas oportunidades para a publicidade, como por exemplo reforçar uma marca ou apelo a um produto exibindo o mesmo anúncio para o mesmo usuário em contextos diferentes.

2.1.3 Modelos de pagamento

Na publicidade online existem 3 modelos de pagamentos principais: CPM, CPC e CPA, sendo que cada um deles está associado a uma etapa do processo de conversão de um usuário. Uma *conversão* é o termo usado para uma ação do usuário que tem algum benefício para o anunciante e que acontece após o clique no anúncio (e.g. comprar um produto anunciado) [BHR10]. Uma *conversão* é uma sequência de eventos, a saber:

1. *Impressão*: o anúncio é exibido na tela.
2. *Clique*: o usuário vê o anúncio, se interessa, e clica no mesmo.
3. *Conversão*: direcionado para o site do anunciante, ali ele decide tomar uma ação, por exemplo comprar um produto ou preencher um cadastro.

A quantidade de vezes em que cada etapa ocorre é bem menor do que a sua anterior. Ou seja, o número de cliques em anúncios é bem menor do que o número de impressões, bem como o número de conversões é bem menor do que o número de cliques. Por conta disso, dá-se o nome de *funil de conversão* a este processo de convencimento de um usuário [BHR10].

O modelo de pagamento CPM cobra um valor do anunciante a cada mil impressões do anúncio. As impressões podem ser feitas sob demanda, ou garantidas em um determinado período de tempo, abordagem conhecida como *entrega garantida*[SLY12, RH12]. O custo por impressão é baixo (da ordem de décimos de centavos de reais) entretanto não há qualquer garantia de que o anúncio foi sequer notado por qualquer pessoa. No modelo CPC o anunciante paga ao publicador somente quando o anúncio exibido recebe um clique. Isto significa que o anunciante só irá gastar com usuários que já demonstraram algum interesse em seu produto ou serviço, diminuindo o risco do investimento em publicidade. Este compartilhamento de risco entre publicador e anunciante eleva o custo do anunciante a um nível intermediário (da ordem de centavos a unidades de reais). Por fim, no modelo CPA o anunciante paga ao publicador cada vez que uma ação de seu interesse é realizada pelo usuário em seu site (e.g. comprar um produto, ou cadastrar-se na newsletter). Neste modelo o anunciante tem garantia de retorno de seu investimento com publicidade, logo o custo tende a ser alto para o anunciante (da ordem de unidades de reais), visto que todo risco está colocado sobre o publicador. Vale notar que a implantação de CPA exige alguns cuidados por parte do publicador, uma vez que o anunciante precisa incorporar em seu site uma chamada de notificação ao sistema de anúncios toda vez que o usuário conclui a ação desejada. A figura 2.4 ilustra o funil de conversão, os respectivos modelos de pagamento associados e a relação custo×risco do publicador em cada um.

Dependendo da estratégia de marketing e do orçamento disponível, um modelo pode ser mais adequado do que o outro. Por exemplo, um anunciante que busca aumentar o tráfego de seu site gostaria de pagar somente quando seu anúncio obtivesse um clique, enquanto que uma campanha de fixação de uma marca pode estar interessada somente que os usuários vejam sua marca, sem necessidade de um clique. É também possível usar combinações dos modelos se isso for interessante para a estratégia de marketing definida.

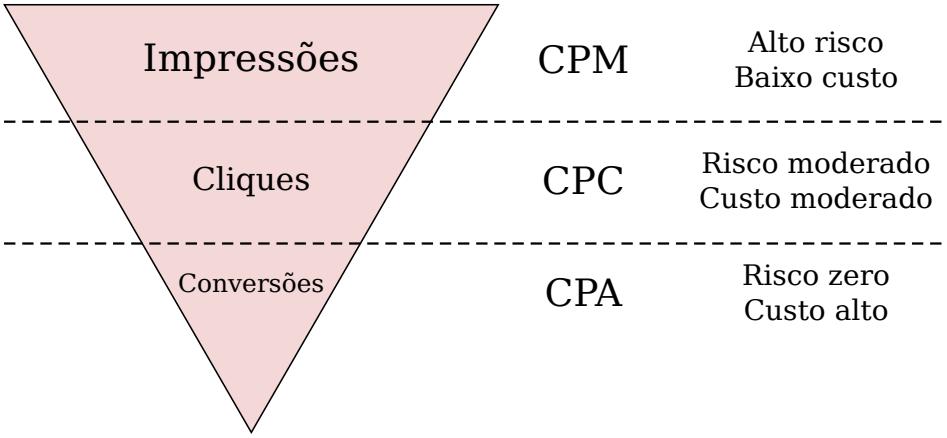


Figura 2.4: *Funil das etapas de conversão de usuários e a relação com cada modelo de rentabilização.*

2.1.4 Modelo de ordenação teórico

Em grandes redes de anúncios, anúncios com modelos de pagamento diferentes competem em um leilão pelos mesmos espaços de exibição. Cada anunciante aposta um valor (BID) que expressa o quanto está disposto a pagar por impressão (CPM), clique (CPC) ou ação (CPA), conforme o modelo escolhido. Há uma generalização teórica do problema de ordenação que permite que anúncios de modelos diferentes compitam em um único leilão. Podemos definir a fórmula de ordenação de anúncios apenas em função de dois valores, a aposta do anunciante (BID) e a probabilidade de se atingir o objetivo esperado com a veiculação do anúncio.

$$RANK = BID * \Pr(sucesso) \tag{2.1}$$

No modelo de CPM o objetivo é simplesmente a impressão do anúncio na tela. Como não há risco do anúncio não ser exibido, temos que $\Pr(impressão) = 1$, e note também que e logo:

$$RANK_{CPM} = BID * \Pr(impressão) = BID \tag{2.2}$$

No modelo CPC, em que o objetivo é o clique do usuário:

$$RANK_{CPC} = BID * \Pr(clique) \tag{2.3}$$

Para o modelo de CPA, além do clique inicial no anúncio, o usuário ainda precisa efetuar a ação desejada pelo anunciante:

$$RANK_{CPA} = BID * \Pr(clique) * \Pr(ação) \tag{2.4}$$

A taxa de usuários que clica em um anúncio exibido geralmente é menor do que 1% e a proporção desses que também realizam uma ação (e.g. compram um produto) é menor ainda. Entretanto, como a escala de sistemas de publicidade na Internet é de milhões de usuários e bilhões de acessos, estimar as probabilidades de clique e de ação é uma área de pesquisa que rende bons resultados. O trabalho de Bagherjeiran et al. explora uma abordagem de otimização para o modelo de CPA

[BHR10]. Há inúmeros aspectos que podem ser considerados para estimar Pr(*clique*) e Pr(*ação*) com relação ao usuário e o que o influencia em seu engajamento com conteúdo publicitário. Algumas das questões que buscaram ser analisadas são: o anúncio é relevante com o interesse do usuário naquele momento? [AGH+09] A exibição excessiva de anúncios irrita o usuário? [BCF+08a] O aspecto visual do anúncio está adequado? [AZZ+12]

A seleção dos anúncios é feita por um algoritmo que leva em conta inúmeros fatores, desde variáveis de qualidade, até o modelo de pagamento escolhido. Tal algoritmo é responsável por gerenciar o risco envolvido na exibição de anúncios que tem pouca chance de trazer retorno financeiro, vetar abusos (trapaças) no modelo de CPA, bem como maximizar os lucros. O bom desempenho do mesmo depende de sua capacidade de prever a *probabilidade de sucesso* na escolha de impressão de um anúncio. Assim, se um anúncio tem pouca chance de atingir o nível desejado no funil de conversão, também deve ter menos oportunidades para aparecer.

O modelo de ordenação pela probabilidade de se obter sucesso com relação ao objetivo da campanha não serve apenas para simplificar o leilão entre modelos de pagamento diferentes, ele, na verdade, descreve uma solução teórica do problema de ordenação de anúncios. Quando o sucesso é atingido na exibição de um anúncio, seja pela impressão, clique ou ação, temos o equilíbrio dos interesses dos participantes. Note que o anunciante atingiu o usuário no nível de envolvimento desejado, o usuário demonstrou interesse no que foi anunciado (clique ou ação), ou foi minimamente incomodado (impressão) e o publicador obteve receita. Este modelo de ordenação seria uma solução muito boa se as probabilidades de sucesso fossem todas conhecidas, pois ele otimiza o retorno financeiro do publicador sem comprometer o ROI do anunciante (todo o investimento é convertido no nível de envolvimento desejado), nem incomodar o usuário (são exibidos apenas os anúncios que mais lhe interessam).

2.1.5 Anatomia de um anúncio

Um anúncio é composto por diversas informações, e vai muito além do título, descrição e link que visualizamos. Por trás da exibição de um anúncio contextualizado existem informações que definem o público alvo, o modelo de pagamento, a página, etc. Nesta seção veremos essas diversas informações e como estão organizadas entre si.

Os dados de um anúncio são armazenados em uma estrutura hierárquica, que facilita o gerenciamento do ponto de vista publicitário. Esta seção baseia-se no trabalho de Bendersky et al. [BGJM10], que descreve esta estrutura para a publicidade contextualizada, chamando-na de *anatomia de um anúncio*. A figura 2.5 representa a hierarquia mencionada e a seguir explicamos a semântica de cada nível na tabela 2.2.

Naturalmente, dependendo da implementação de cada rede de anúncio, pode haver variações na estrutura de dados utilizada, bem como informações complementares utilizadas por funcionalidades específicas.

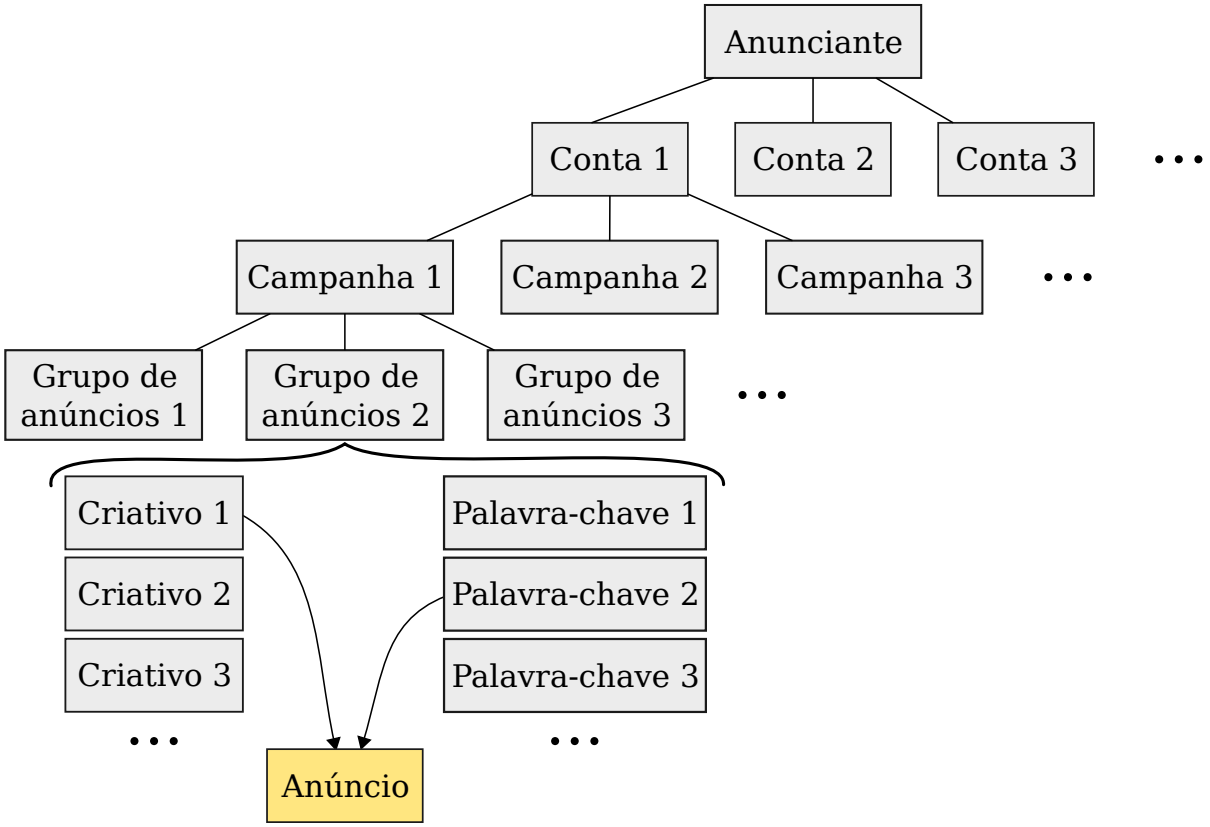


Figura 2.5: *Estrutura hierárquica das informações de um anúncio.*

Entidade	***Descrição***
Anunciante	Representa uma entidade do mundo real que será responsável pela elaboração de anúncios. Pode ser uma pessoa ou uma empresa.
Conta	Segundo nível de organização do anunciante, que pode ter várias contas. É bastante útil, por exemplo, no caso de agências de publicidade que podem organizar os anúncios de seus clientes em diversas contas.
Campanha	Representação de uma campanha publicitária de uma loja. Serve para agrupar anúncios sob um mesmo motivo, tais como: uma promoção (e.g. saldão de sapatos femininos), divulgação em uma data comemorativa (e.g. dia dos pais). Na verdade uma campanha encapsula grupos de anúncios e não anúncios propriamente ditos.
Segmentação	Definição demográfica do público alvo, que é uma informação importante para escolha de anúncios contextualizados. Considera características mais gerais do usuário, como região geográfica, sexo, idade, idioma, etc. Geralmente está associada a uma campanha, significando que todos os anúncios da campanha sejam dirigidos ao público-alvo determinado pela segmentação.
Grupo de Anúncios	Agrupa criativos e palavras-chave relacionados aos anúncios específicos de um único produto ou serviço.
Criativo	Parte visual de um anúncio. Pode ser texto, imagem, vídeo ou uma animação flash. Geralmente tem caráter apelativo e faz chamada a uma ação, usando frases como “compre já”, ou “vagas limitadas”.
Palavras-chave	São palavras ou frases que representam interesses específicos de um usuário. São utilizadas como uma segmentação de granularidade mais fina e utilizadas de maneira diferente dependendo do tipo de publicidade. Em busca patrocinada são casadas com a consulta realizada, enquanto que em publicidade contextualizada com o conteúdo da página, por exemplo.
Anúncio	Um anúncio é a composição de um criativo e uma palavra-chave de um mesmo grupo de anúncios. A rigor, cada combinação de criativo e palavra-chave é considerada como um anúncio diferente. Isso permite explorar diferentes formas de anúncios para um mesmo produto e descobrir as mais efetivas.
Aposta	Um valor em dinheiro associado ao anúncio exibido que deverá ser pago pelo anunciante ao publicador conforme o modelo de pagamento escolhido.

Tabela 2.2: *Entidades que compõe os dados de um anúncio.*

2.2 Sistema de exibição de anúncios

Neste trabalho estudamos e realizamos experimentos em um sistema de publicidade que segue o modelo de uma rede de anúncios. Nesta seção apresentamos uma visão geral do funcionamento de uma rede anúncios e detalhamos a tarefa do sistema que mais interessa nosso trabalho, a saber, a reordenação de anúncios.

É comum que sistemas de anúncios (SA) sirvam milhões de usuários e bilhões de requisições mensais. Seu tempo de resposta deve ser extremamente rápido (da ordem de 200 ms), de modo a evitar qualquer incômodo perceptível ao usuário. Para atender estes requisitos de performance eles são projetados de maneira semelhante a de máquinas de busca, lidando com a indexação e recuperação de páginas, bem como de *crawling* das páginas dos publicadores e anunciantes.

2.2.1 Visão geral

O funcionamento de um SA exige a realização de tarefas das mais diversas. Algumas com interação humana, outras completamente automatizadas. Elas ocorrem tanto para atender uma requisição de anúncios efetuada até aquelas que envolvem a preparação da infra estrutura. A figura 2.6 representa o funcionamento do SA e abaixo listamos uma descrição mais detalhada de cada etapa do fluxo de exibição de anúncios:

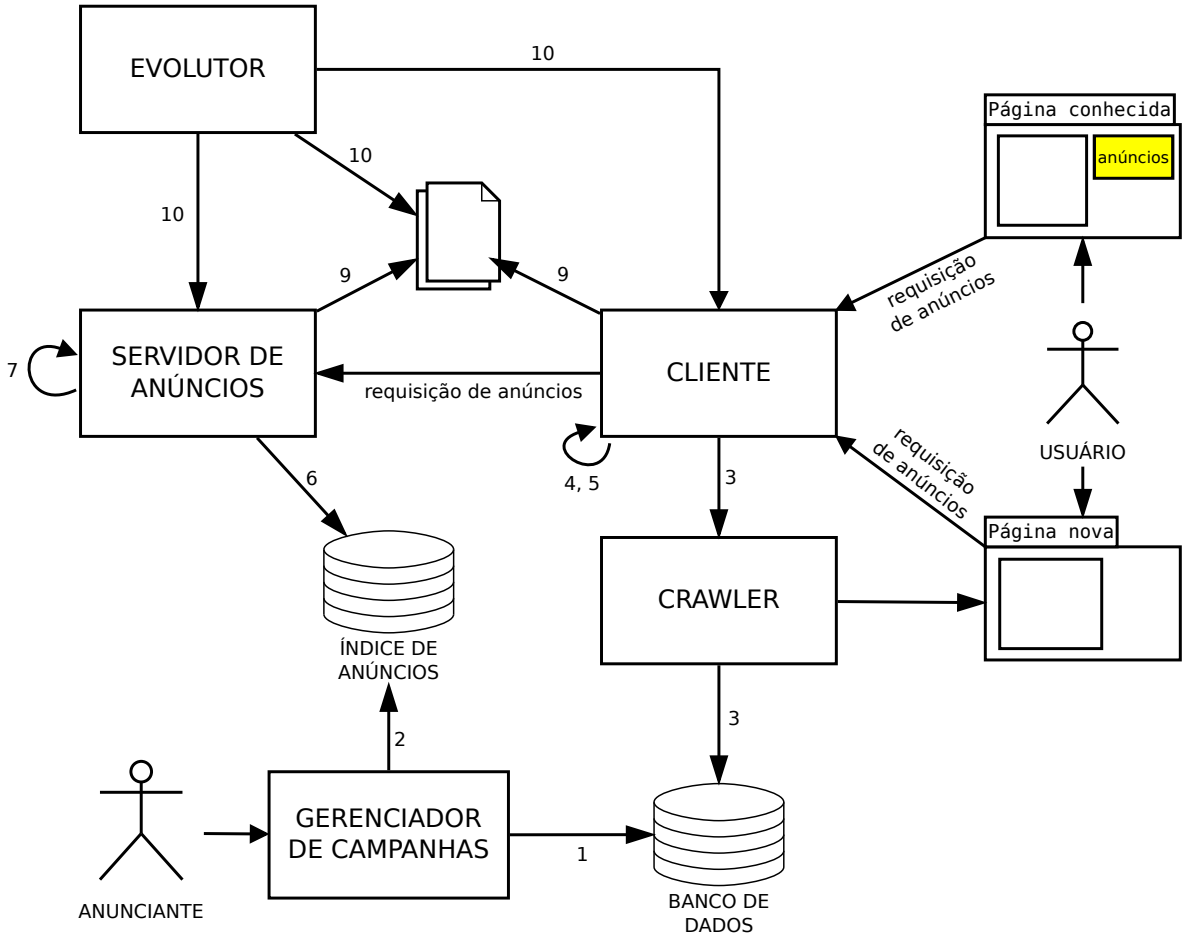


Figura 2.6: *O funcionamento de um sistema de anúncios.*

1	Inserção de dados de anúncio	6	Indexação dos anúncios
2	Coleta de palavras-chave	7	Geração de consulta para recuperação de anúncios
3	Agregação de outras fontes de dados	8	Recuperação de anúncios
4	Reordenação dos anúncios	9	Exibição dos anúncios
5	Coleta de dados	10	Aprendizado e evolução da escolha de anúncios

Tabela 2.3: *Lista de atividades representadas na figura 2.6*

1. Inserção de dados de anúncio

   O anunciante faz o cadastro dos dados da campanha de anúncios que deseja veicular. Geralmente há um sistema com interface visual amigável que permite a entrada de dados do anúncio, segmentação de usuários, tempo de veiculação, aposta máxima, criativo, etc. Todos os dados descritos na tabela 2.2 são inseridos nesta etapa. Em algumas implementações de SA ocorre ainda a captura de palavras da página de destino do do clique, informação que pode ser útil para a escolha do melhor anúncio, como no trabalho de Murdock et Al. [MCP07]

2. Indexação dos anúncios

   Periodicamente o SA executa a atualização de seu índice de anúncios, permitindo que os anúncios recém criados comecem a ser veiculados. A estrutura de dados mais convencional é a mesma utilizada em máquinas de busca, o índice invertido [MRS08]. Os termos do dicionário são retirados dos diversos níveis de definição de um anúncio: desde informações sobre o anunciante até as palavras-chave associadas aos grupos de anúncios. A indexação pode variar um pouco conforme cada implementação. O trabalho de Ribeiro-Neto, por exemplo, utiliza métodos de expansão de consulta para evitar falta de correspondência entre o vocabulário das páginas e dos anúncios [RNCGSdM05].

   Conforme vimos na seção 2.1.5, cada combinação entre um criativo e uma palavra-chave poderia ser considerada um anúncio diferente. Restrições comuns em uma rede anúncios permitem centenas de criativos e dezeas de milhares de palavras-chave relacionadas, o que leva a milhões de combinações de anúncios em um único grupo. Considerando centenas de milhares de contas de anunciantes com milhares de grupos de anúncios cada, teríamos uma quantidade de documentos (anúncios) da ordem de centenas de trilhões, número que chega aos limites até mesmo dos motores de busca mais poderosos. Bendersky et al. [BGJM10] comparam diferentes estratégias de indexação de anúncios; aproveitando-se da estrutura hierárquica dos mesmos, reportam que indexar grupos de anúncios e depois escolher combinações de anúncio entre os grupos recuperados apresenta ganho não só de velocidade, mas também de relevância. E isto é intuitivo se pensarmos que um grupo de anúncio é concebido de modo a divulgar um produto único.

3. Coleta de palavras-chave

   Os publicadores criam novas páginas diariamente que também devem servir anúncios, mas gerenciar o cadastro dessas novas páginas manualmente é inviável. A técnica comumente utilizada é identificar páginas que não são conhecidas e disparar o processo de reconhecimento das mesmas quando a primeira chamada ao SA é feita a partir delas. Dessa maneira, as primeiras exibições de anúncios nelas não tem contextualização adequada. O SA então faz o reconhecimento da página utilizando técnicas de recuperação de informação, como extração

de palavras-chave, identificação de língua, identificação de semântica, categorização por assunto, etc. Assim, tão logo o SA faça o reconhecimento da página, ela estará disponível para contextualização adequada.

4. Geração de consulta para recuperação de anúncios
   Uma vez que a URL de origem da requisição já tenha sido reconhecida e as informações de interesse extraídas e armazenadas, é possível gerar uma consulta *on-demand* ao índice de anúncios no momento em que uma requisição de anúncios é feita ao servidor.

5. Agregação de outras fontes de dados
   Além das informações associadas à página de exibição dos anúncios são agregados outros dados que permitem expandir a contextualização, como interesses do usuário, perfil demográfico, região geográfica, entre outros.

6. Recuperação de anúncios
   Gerada a consulta a ser disparada contra a base de anúncios, são feitas buscas em um ou mais índices invertidos, dependendo da estrutura e da complexidade da lógica utilizada na recuperação dos anúncios. Nesta etapa recupera-se um número de anúncios muito superior ao requisitados pois é essencial ter uma boa cobertura de todos os anúncios que possam ser relevantes para o contexto de exibição. Embora os sistemas de recuperação de informação já devolvam resultados ordenados por relevância, é a próxima fase é que será responsável por determinar a ordenação de relevância final, e consequentemente a escolha dos anúncios que serão exibidos.

7. Reordenação dos anúncios
   Os anúncios recuperados do índice de buscas são então submetidos a funções de reordenação que levam em consideração diversas variáveis de interesse (e.g. taxa de clique, taxa de conversão, valor da aposta). Estas funções podem ser desde heurísticas desenvolvidas por especialistas de domínio, até sofisticados modelos de aprendizado computacional. Muitos trabalhos acadêmicos, este inclusive, trabalham principalmente em como aprimorar esta etapa.

8. Exibição dos anúncios
   Determinados e ordenados os anúncios que devem ser exibidos, o SA deve realizar a inserção dos criativos em uma página da web que, geralmente, não é de sua propriedade. Para que isso possa ocorrer, o publicador teve de inserir em sua página um trecho de código fornecido pelo gerenciador do SA. Este código é responsável por requisitar anúncios e renderizar a resposta recebida nos espaços publicitários determinados pelo publicador. A figura 2.7 ilustra este processo.

9. Coleta de dados
   Durante o período de veiculação das campanhas são coletados dados de clique, comportamento do usuário e desempenho dos anúncios. Informações de impressão, clique e conversão são capturadas para cobrança dos anunciantes conforme o modelo de pagamento escolhido. O processamento desses dados também não pode levar muito tempo pois restrições de orçamento devem ser respeitadas, principalmente nos modelos pré-pagos.

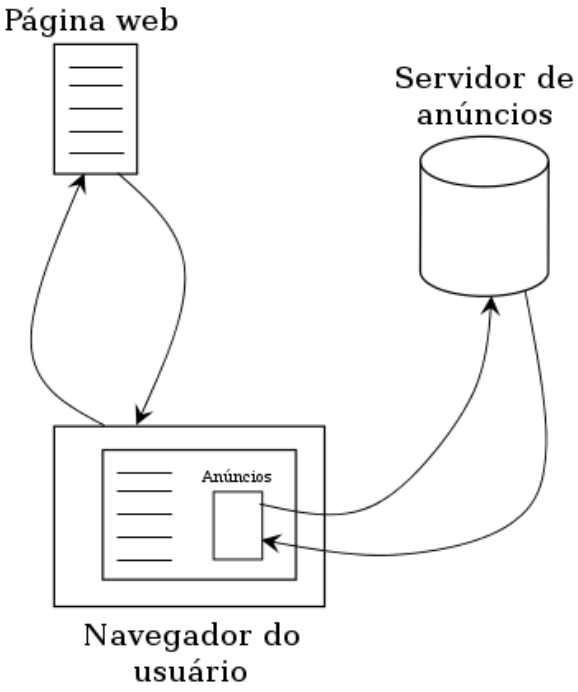


Figura 2.7: *Processo de uma requisição. O navegador requisita uma página web e os anúncios advindos do SA são inseridos via Javascript.*

10. Aprendizado e evolução da escolha de anúncios
    Os dados coletados são utilizados na atualização de parâmetros do sistema (e.g. CTR global, taxa de conversão de um site) e servem também como entrada de dados para algoritmos de aprendizado computacional. Indústria e academia atuam em conjunto para a evolução dos modelos de recomendação e escolha dos anúncios, de modo a otimizar os ganhos.

2.3 Reordenação de anúncios

A fase de reordenação de anúncios conforme descrita na seção anterior é o tema principal de estudo deste trabalho. A ordenação final dos anúncios recuperados tem um papel muito importante no desempenho do geral do sistema. Como os espaços reservados a publicidade são limitados (sendo apenas um único em muitos casos) a ordenação dos anúncios recuperados não determina apenas a ordem de exibição (destaque que o anúncio receberá), mas sim a própria exibição ou não do mesmo. O que claramente é essencial para o sucesso do sistema, uma vez que a exibição de publicidade relevante ao usuário foi o impulsionou o grande crescimento do mercado de publicidade computacional, e até mesmo a invenção e o sucesso dos modelos de pagamento CPC e CPA.

A reordenação é feita pelo SA durante o processo de resposta de uma requisição de anúncios, e portanto, tem fortes restrições de performance. Uma maneira eficiente para implementá-la é por meio de funções matemáticas de variáveis representativas associadas a cada anúncios. Quanto maior o valor que a função assumir, melhor colocado estará o anúncio. Chamamos essas funções de *estratégias de ordenação*. O SA estudado neste trabalho segue este modelo e a tabela 2.4 descreve as variáveis que de que ele dispõe.

Uma tendência recente da área de publicidade computacional é a de melhorar o conhecimento

Variável	*Descrição*
BID	A aposta do anunciante. É o valor máximo que o anunciante está disposto a pagar por um sucesso no modelo de conversão escolhido: uma impressão (CPM), um clique (CPC) ou uma ação (CPA).
CTR	O desempenho histórico do anúncio representado por sua taxa de clique. Calcula-se o CTR do anúncios dividindo o número de cliques recebidos pelo total de impressões do mesmo realizadas.
CTX	O *score* de relevância textual (*tf-idf*) devolvido pela ferramenta de recuperação de informação utilizada (e.g. *Apache Solr*). O CTX define uma ordenação que leva em conta apenas a semelhança textual entre o anúncio e a página acessada pelo usuário.
BTU	Outro *score* de relevância textual. Obtido da semelhança entre os anúncios e uma coleção de palavras que representa os interesses do usuário (e.g. "futebol", "carros velozes", "moda inverno" ou "celular galaxy s5").

Tabela 2.4: *Variáveis utilizadas pelo sistema estudado para reordenação de anúncios.*

sobre o usuário que está navegando, não de forma generalista, mas individualizada. Conhecer o perfil do usuário (sexo, faixa etária, região geográfica, interesses pessoais, etc...) permite que a recomendação de publicidade seja mais precisa, fornecendo ao usuário anúncios de seu interesse e melhorando o retorno obtido pelos anunciantes. A obtenção desse tipo de dado pode ser feita por meio de redes sociais, requisitando ao usuário acesso aos dados; ou por rastreamento de comportamento (*behavioral tracking* ou *behavioral targeting*), área que utiliza aprendizado de máquina para analisar padrões de navegação e classificar usuários em perfis. Essa é uma área promissora, que tem obtido melhorias significativas no engajamento do usuário [TLY+11].

2.3.1 Métricas de desempenho

A reordenação de anúncios é um problema de otimização com múltiplos objetivos, a saber, do usuário, do anunciante e do publicador. Tais objetivos são subjetivos e em muitos casos conflitantes, como exemplificado na introdução deste trabalho 1.1. Além disso, é de se esperar que os participantes mudem de objetivo ao longo do tempo, por exemplo um usuário que altere seu comportamento de compra nos finais de ano, ou o anunciante que decide fazer uma ação de marketing apenas para aumentar a exposição de sua marca, ao invés de simplesmente vender seus produtos.

É evidente que não existe uma solução ótima universal para a ordenação de anúncios. Logo, se quisermos encontrar funções de ordenação cada vez melhores, precisaremos de um método objetivo que nos permita julgar a qualidade das soluções. Fazemos esta avaliação por meio de métricas de desempenho representativas aos participantes do sistema, de modo que um aumento de seu valor numérico represente um aumento na satisfação do participante. Na tabela 2.5 definimos três métricas de interesse e descrevemos seu efeito sobre os participantes interessados. Elas servirão de base para análise do desempenho das funções de ordenação de anúncios.

Note que a métrica *ROI* é um dado que pertence ao anunciante e nem sempre está disponível para o SA. Em nosso SA estudado, embora haja viabilidade sistêmica para fornecimento deste dado, muitos anunciantes preferem não informá-lo pois o julgam como segredo de negócio.

Métrica	*Definição*	*Interessados*
CTR global	Número de cliques dividido pelo número de impressões totais do sistema.	Usuário e anunciante. O aumento do número de cliques representa maior engajamento do usuário e, consequentemente, maior tráfego para a página do anunciante.
ROI	Retorno recebido sobre valor investido em publicidade	Anunciante. O objetivo do anunciante é otimizar seus investimentos em publicidade.
ECPM	Valor financeiro arrecadado a cada mil impressões	Publicador. Geração de receita é o motivo principal da veiculação de anúncios.

Tabela 2.5: *Métricas representativas do interesse de cada participante do sistema.*

2.3.2 Função objetivo

As métricas de desempenho são indicadores associados ao interesse de cada participante no sistema de anúncios. Precisamos, entretanto, de uma métrica única para comparação do desempenho das diferentes estratégias de ordenação. Por esse motivo, com base no trabalho de Broinizi [BF15], introduzimos o conceito de *função objetivo* que nos permitirá comparar as diferentes fórmulas de ordenação e eleger a melhor.

Definimos a função objetivo como uma ponderação que representa a importância comercial dada pelo publicador a cada um dos participantes no equilíbrio do sistema:

$$Fobj = \alpha\,\overline{CTR} + \beta\,\overline{ROI} + \gamma\,\overline{ECPM} \tag{2.5}$$

onde

$$0 \leq \alpha, \beta, \gamma \leq 1$$

$$\alpha + \beta + \gamma = 1$$

e as métricas de desempenho de cada estratégia são normalizadas por uma estratégia de *baseline* que, como veremos em 4.3.1, é configurada como uma reordenação aleatória dos anúncios.

$$\overline{CTR} = \frac{CTR - CTR_{baseline}}{CTR_{baseline}}$$

$$\overline{ROI} = \frac{ROI - ROI_{baseline}}{ROI_{baseline}}$$

$$\overline{ECPM} = \frac{ECPM - ECPM_{baseline}}{ECPM_{baseline}}$$

Função objetivo regularizada

Como discutido anteriormente, o bom desempenho de um sistema de publicidade computacional depende do equilíbrio entre os interesses dos participantes do sistema. Introduziremos a variância dos interesses como um parâmetro de regularização a fim de favorecer estratégias que apresentem desempenho equilibrado:

$$\sigma^2 = \frac{(\overline{CTR} - \mu)^2 + (\overline{ROI} - \mu)^2 + (\overline{ECPM} - \mu)^2}{3}$$

$$\mu = \frac{\overline{CTR} + \overline{ROI} + \overline{ECPM}}{3}$$

Assim, a função objetivo passar a ser:

$$Fobj = \alpha\,\overline{CTR} + \beta\,\overline{ROI} + \gamma\,\overline{ECPM} - \delta\,\sigma^2 \tag{2.6}$$

σ^2 também é ponderado por um parâmetro $\delta \geq 0$, a fim de que o publicador seja capaz de controlar qual a importância do equilíbrio do sistema para ele. Este controle é fundamental para a estratégia de mercado do publicador. Podendo ele priorizar ora seu interesse comercial, ora o dos anunciantes ou do usuário.

Interpretação da função objetivo

É fácil ver que para a estratégia de *baseline* vale que $\overline{CTR} = \overline{ROI} = \overline{ECPM} = \sigma^2 = 0$ e portanto $Fobj_{baseline} = 0$. Dessa maneira, toda estratégia cujo valor da função objetivo seja positivo julgaremos melhor do que o *baseline* e pior caso contrário. Além disso, como todas as estratégias são normalizadas podemos comparar seus valores entre si e elegermos a melhor.

Função objetivo ajustada

Em um ambiente em que todas as variáveis necessárias estejam disponíveis a função objetivo a ser utilizada é exatamente a que apresentamos acima. No ambiente em que conduzimos nossos testes, e de maneira mais geral em todo o mercado brasileiro de publicidade online não temos a informação de ROI dos anunciantes (note que não precisamos do valor monetário real, apenas um número que indique a taxa de conversões). No mercado externo é muito mais comum a abertura dessa informação, para que as redes de anúncios possam otimizar as campanhas publicitárias, principalmente no modelo de CPA. As empresas brasileiras tem se acostumado mais com essa ideia, e acreditamos que em poucos anos a utilização do ROI já seja possível.

Por conta disto, ajustaremos nossa função objetivo para dar mais peso ao CTR, uma vez que ele serve como indicador indireto do interesse do anunciante.

$$Fobj = (\alpha + \beta)\,\overline{CTR} + \gamma\,\overline{ECPM} - \delta\,\sigma^2 \tag{2.7}$$

onde

$$\sigma^2 = \frac{(\overline{CTR} - \mu)^2 + (\overline{ECPM} - \mu)^2}{2}$$

$$\mu = \frac{\overline{CTR} + \overline{ECPM}}{2}$$

2.4 Teste A/B

Um "teste A/B" (*A/B test*) é um experimento controlado em que os usuários de um *website* são aleatoriamente expostos a duas variações do sistema: controle (A) e tratamento (B). [KLSH09] O grupo de controle é exposto a versão atual do sistema enquanto que o de tratamento a uma versão que se supõe ser melhor. Para tal julgamento, define-se uma métrica de avaliação representativa do objetivo que se deseja atingir. A avaliação é feita por meio de um teste de hipóteses. A hipótese nula é definida como as duas versões do sistema sendo equivalentes, sendo a mesma rejeitada ou não segundo um nível de significância.

O termo "teste A/B" foi bastante difundido nos últimos anos, em especial no meio de publicidade digital, área em que surgiram diversos livros sobre o assunto [SK13, Ash08, EQvDC08], bem como ferramentas e *software* para auxiliar sua execução. A terminologia é semelhante à utilizada em áreas como estatística e matemática. Ela esta descrita com mais detalhes na tabela 2.6.

Termo	***Definição***
Métrica de avaliação	Uma medida quantitativa do objetivo do experimento. Alguns termos equivalentes são: critério de avaliação global, métrica de desempenho ou função de ajuste (*fitness function*). Um experimento pode ter mais de uma métrica de avaliação, caso em que será necessária a avaliação de um analista, ou então adota-se uma ponderação matemática.
Fator	Uma variável controlável que supõe-se influenciar a métrica de avaliação. Aos fatores são atribuídos valores (também chamados de níveis ou versões). Em testes A/B simples há um único fator, que assume apenas dois valores: A e B.
Variante	Uma das possíveis versões de sistema a que os usuários estão sendo expostos. É definida ao se atribuir valores a cada um dos fatores do teste. Em testes A/B simples há apenas duas variantes, a de controle e a de tratamento, determinadas pelos fatores A e B, respectivamente.
Unidade experimental	A entidade sobre a qual as métricas de avaliação são calculadas antes da comparação com cada variante. Assume-se que as unidades são independentes. Na Internet, é comum que o usuário seja a unidade experimental, embora se use também usuários do dia, sessões de usuário e visualizações de página (*pageviews*).

Tabela 2.6: *Terminologia usada em testes A/B*

2.5 Análise de Componentes Principais

A Análise de Componentes Principais (*Principal Component Analysis* ou PCA) é um procedimento estatístico utilizado em análise multivariada e análise exploratória de dados. A ideia central é reduzir a dimensionalidade de um conjunto de dados contendo um grande número de variáveis relacionadas, e concomitantemente reter a maior parte possível da variância original. Isto é obtido através de uma transformação dos dados originais em um novo conjunto de variáveis, as chamadas componentes principais (PCs), que são não correlacionadas, e estão ordenadas de modo que a primeira retém a maior parte da variância original. Os valores das novas variáveis, ou as componentes principais, são calculados por meio de uma mudança de base realizada no espaço das variáveis originais, de tal modo que a nova base esteja alinhada com os eixos de maior dispersão dos dados originais.

2.5.1 Definição

Seja x um vetor de p variáveis aleatórias, o primeiro passo é encontrar uma função linear $a_1'x$ dos elementos de x contendo variância máxima, onde a_1 é um vetor de p constantes $a_{11}, a_{12}, ..., a_{1p}$ tal que:

$$a_1'x = a_{11}x_1 + a_{12}x_2 + ... + a_{1p}x_p = \sum_{j=1}^{p} a_{1j}x_j$$

Em seguida, encontramos uma função linear $a_2'x$, não correlacionada com $a_1'x$ tendo máxima variância, e assim por diante, de modo que no k-ésimo passo uma função linear $a_k'x$ seja achada tendo máxima variância sujeita a ser não correlacionada com $a_1'x, a_2'x, ..., a_{k-1}'x$. A k-ésima variável derivada, $a_k'x$ é a k-ésima componante principal (PCk).

Uma vez definidas as componentes principais, precisamos saber como encontrá-las. Considere o caso em que o vetor de variáveis aleatórias tenha uma matriz de covariâncias Σ conhecida (no caso mais realístico em que Σ é desconhecido, seguimos substituindo por uma matriz estimada S) definida por:

$$\Sigma_{i,j} = \mathrm{E}[(X_i - \mu_i)(X_j - \mu_j)]$$

que são as covariâncias entre as variáveis X_i e x_j, para $i! = j$ e a própria variância de X_i quando $i = j$. A i-ésima componente principal pode ser obtida por:

$$z_k = a_k'x$$

onde a_k é um autovetor de Σ correspondente a seu k-ésimo maior autovalor (λ_k). Além disso, se a_k é escolhido para ter comprimento unitário ($a_z'a_k = 1$), então a variância $var(z_k) = \lambda_k$.

Não é escopo deste trabalho apresentar uma demonstração detalhada de como o método de PCA é construído. Para obter mais informações neste sentido o leitor pode ler o livro de Jolliffe [Jol02], citado na bibliografia.

2.5.2 Procedimento

Como visto acima, as componentes principais podem ser obtidas por meio da matriz de covariância do conjunto de dados. Eis o procedimento para obtê-las:

1. Normalizar cada uma variáveis do conjunto de dados (pois o PCA é sensível aos valores absolutos das dimensões)
2. Calcular a matriz de covariância dos dados normalizados [CDC87];
3. Calcular os autovalores e autovetores da matriz de covariância;
4. O autovetor (z_1) com maior autovalor (λ_1) associado é a primeira componente principal;
5. O autovetor (z_2) com o segundo maior autovalor (λ_2) associado é a segunda componente principal, e assim por diante;
6. Escolher a quantidade de componentes principais ($z_1, ..., z_k$) baseada na soma acumulada das variâncias. (e.g. para reter 80% da variância original dos dados escolha o primeiro k tal que $\lambda_1 + \lambda_2 + ... + \lambda_k > 0.8$)

2.5.3 Discussão

A primeira componente principal está na direção de maior dispersão dos dados, ou seja é a componente com a maior variância dos dados originais. A segunda componente é a direção de maior dispersão sob a restrição de ser ortogonal a primeira. E assim por diante, a i-ésima componente principal está na direção de maior variância sob a restrição de ser ortogonal a todas as componentes principais anteriores. Podemos imaginar as componentes principais como os eixos de um elipsóide que circunscreve o conjunto de dados, o eixo de maior amplitude é a primeira componente e assim por diante. A figura 2.8 [1] mostra esta ilustração geométrica para um conjunto de dados em 3 dimensões.

PCA é comumente utilizado para redução de dimensionalidade em conjuntos de dados com muitas variáveis, pois permite encontrar uma projeção dos dados que preserve grande parte da variância original. Dessa forma diminuímos a complexidade do problema perdendo pouco ou quase nada da expressividade dos dados. Um outro uso comum para o PCA é na análise exploratória de dados. Utilizando-se de visualizações de projeções das variáveis originais na nova base podemos entender melhor as correlações entre as variáveis originais, e este é o principal uso de PCA neste trabalho.

Embora o PCA já seja um método estabelecido na literatura, há trabalhos recentes sendo feitos para aprimorá-lo [SXY13, SLCC14], principalmente no que diz respeito a sua aplicabilidade sobre grande volumes de dados. Há uma técnica mais recente, denominada *t-SNE* [VdMH08], que é particularmente eficiente para visualização de conjuntos de dados com alta dimensionalidade (centenas ou milhares de dimensões). Em nosso caso, como estamos lidando com um espaço de baixa dimensionalidade (apenas 4 variáveis) não foi necessária a utilização do t-SNE. O PCA é uma técnica amplamente utilizada, mas é conhecido por diferentes nomes em cada área de atuação, como KLT e SVD entre outos.

[1] Retirado de http://www.joyofdata.de/public/pca-3d, acessado em abril de 2015.

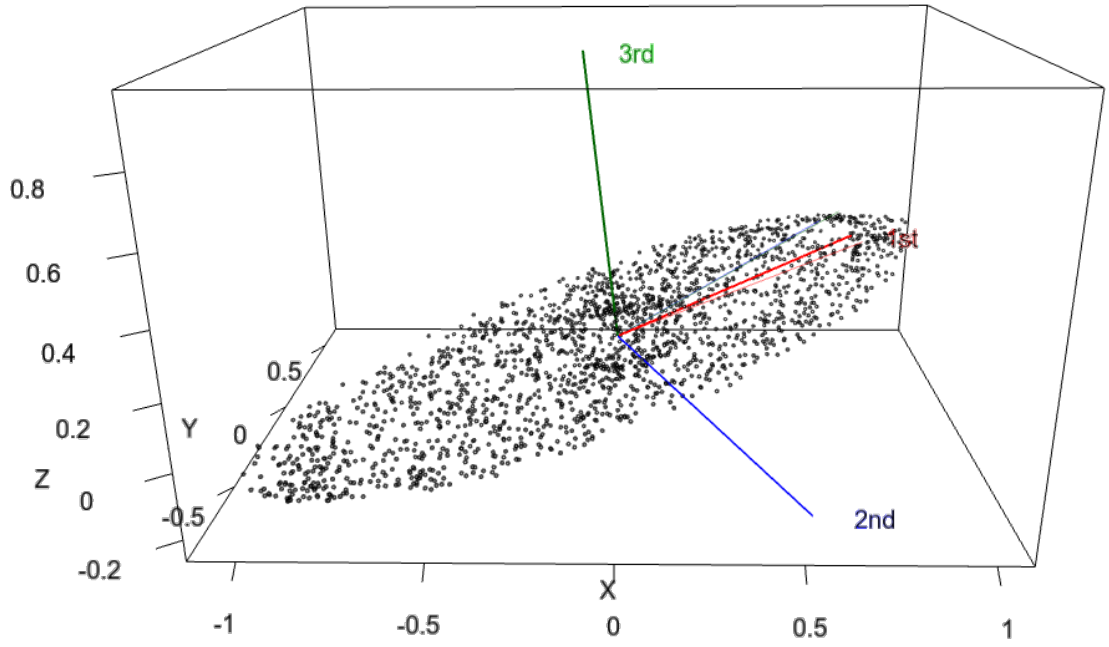


Figura 2.8: *Componentes principais em um conjunto de pontos*

Capítulo 3

Trabalhos Relacionados

No desenvolvimento de nosso trabalho, nos posicionamos no panorama de pesquisa mais atual analisando trabalhos publicados nas últimas edições de seis importantes conferências e duas revistas. Todos contendo conteúdo voltado para a área de publicidade computacional. A tabela 3.1 mostra a distribuição dos artigos nas diferentes fontes e edições.

Fonte	***Total***	***Edições***
Conference on Information and Knowledge Management (CIKM)	20	2007 (1), 2008 (2), 2009 (1), 2010 (1), 2011 (5), 2012 (8), 2014 (2)
Conference on Knowledge Discovery and Data Mining (KDD)	12	2007 (1), 2008 (1), 2012 (5), 2013 (1), 2014 (4)
Communications of the ACM, novembro 2011	1	2011 (1)
Symposium On Applied Computing (SAC)	1	2012 (1)
Special Interest Group on Information Retrieval (SIGIR)	5	2005 (1), 2006 (1), 2007 (1), 2009 (1), 2010 (1)
Transactions on Intelligent Systems and Technology (TIST)	5	2011 (1), 2014 (3), 2015 (1)
International Conference on Web Search and Data Mining (WSDM)	12	2010 (1), 2012 (4), 2013 (1), 2014 (5), 2015 (1)
International World Wide Web Conference (WWW)	15	2006 (1), 2008 (1), 2009 (1), 2010 (1), 2012 (6), 2013 (2), 2014 (3)

Tabela 3.1: *Tabela com a distribuição dos artigos utilizados em nossa pesquisa.*

Grande parte dos artigos avaliados são diretamente relacionados com nosso trabalho, e outros são marginais. Ao todo avaliamos 71 artigos e destacamos a seguir os principais assuntos abordados. No quesito contexto de publicidade 2.1.1, 24 abordam publicidade contextualizada e 12 busca patrocinada. Com relação ao formato visual 2.1.2, 16 lidaram com anúncios de *display* (flash ou imagem) e 13 com anúncios de texto e imagem. A seleção de anúncios de modo geral foi o tema de 28 artigos, sendo que 3 deles tratam da fase de recuperação, 14 da de reordenação e outros 11 no modelo de recomendação. Inúmeros trabalhos abordam o problema de ordenação utilizando aprendizagem computacional, sendo a regressão logística o método mais utilizado (8 artigos). A

formulação do problema como a probabilidade de sucesso 2.1.4 é utilizada por 8 artigos. Por fim, as métricas de desempenho mais utilizadas foram o *score* de relevância textual e a taxa de clique (CTR), com 12 e 10 artigos respectivamente.

A área de publicidade computacional é fortemente influenciada pelas práticas do mercado, como os formatos de publicidade mais vendidos, as necessidades dos clientes e a tecnologia dos concorrentes. É uma área de pesquisa da computação diretamente aplicada ao mundo real, e que possui forte incentivo financeiro para expandir seus limites, como visto no capítulo de introdução. Em nossa pesquisa identificamos três principais frentes de trabalho que emergiram naturalmente ao longo do tempo e a evolução dos sistemas de publicidade:

- primeiramente a ***recuperação*** dos anúncios em grandes volumes de dados;
- posteriormente a ***reordenação*** dos anúncios recuperados para melhorar a relevância ao usuário;
- e mais recentemente a ***evolução*** dos sistemas de anúncios com base em dados.

1ª geração: Recuperação

A recuperação de anúncios é uma área de pesquisa mais consolidada e, portanto, com menos trabalhos recentes publicados. Os primeiros trabalhos [RNCGSdM05, BFJR07, YGC06] abordam o problema somente do ponto de vista de recuperação de informação, e tem como objetivo melhorias na relevância textual. Outros mais recentes, entretanto, já acoplaram capacidades de aprendizagem computacional, como por exemplo [AG12].

O artigo [RNCGSdM05] mostra que a discrepância de vocabulário existente entre anúncios e páginas atrapalha a recuperação textual. Os termos utilizados nos anúncios são geralmente genéricos e se referem a uma ampla área de atuação (e.g. carro esportivo), enquanto que um artigo que fale sobre amortecedores pode não usar tal expressão. Para mitigar o problema de perda de cobertura ocasionado pela diferença de vocabulário, Ribeiro-Neto et al. [RNCGSdM05] propõem dez novas estratégias de associação entre anúncios e páginas. As cinco mais simples utilizam palavras-chave para melhor contextualização, e as cinco mais elaboradas utilizam uma rede baeysiana para expandir o vocabulário das páginas com termos relevantes que são utilizados pelos anunciantes. O trabalho de Broder et al. [BCF+08b] abordou a reordenação de anúncios em busca patrocinada e obteve ganhos significativos por meio de técnicas de expansão de consulta. A seleção de anúncios é feita logo após a recuperação dos resultados de busca, e se utiliza do conteúdo dos primeiros resultados para aumentar a consulta que é feita ao servidor de publicidade. Os termos mais significativos são incluídos na consulta e também submetidos a uma classificação taxonômica que relaciona o anúncio a uma classe comercial.

2ª geração: Reordenação

Nos últimos anos, entretanto, vemos que o foco de pesquisa na área tem sido o problema de reordenação dos anúncios pós recuperação. São muitas as propostas para enriquecimento do modelo de ordenação (recomendação) de anúncios, dentre elas: características visuais [CZA+12, AZZ+12], localização geográfica [AC14], expansão do vocabulário dos anúncios [WLF+09], interesses pessoais do usuário [TPGJ11, FKLT12], indisposição do usuário com publicidade [BCF+08a] e até se o

mesmo está entediado [KSSS15]. Cada uma dessas propostas captura novos aspectos do contexto do usuário, e são diferenciais competitivos dos publicadores de anúncios. Tanto no fato de terem mais insumos para seu algoritmo de reordenação (melhor escolha dos anúncios recomendados), quanto em oferecer aos anunciantes mais funcionalidades para escolha de seu público-alvo.

O artigo [MCP07], assim como [RNCGSdM05], trabalha com a diferença de vocabulário existente entre anúncios e páginas. Olhando do ponto de vista de tradução estatística, ele considera que a mensagem do anúncio é a mesma mensagem da página de destino, mas em uma outra linguagem: a linguagem apropriada para publicidade. Esses autores propõem duas novas classes de variáveis que melhoram a contextualização textual do anúncio com a página de destino: uma baseada em tradução automática e a outra baseada em avaliação de tradução. Elas poderiam ser incorporadas à variável de CTX que descrevemos, ou então a uma variável específica para esse fim. O artigo [CAJ08], além de descrever como obtém suas variáveis de contextualização textual (que são diferenciadas por regiões da página), também utiliza resposta de clique dos usuários como entrada da ordenação dos anúncios recuperados (variável CTR). Exibe um modelo híbrido entre CTX e CTR construído com regressão logística que apresenta resultados muito superiores ao uso exclusivo de CTX.

Um trabalho mais recente da área é o artigo de Agarwal e Gurevich [AG12], que considera o método de aprendizado utilizado para reordenação como uma caixa preta. Ele encontra uma função linear que seja uma boa aproximação para o mesmo. Essa aproximação pode ser calculada e indexada para todos os itens da coleção, de modo que a fase de recuperação tenha mais acurácia e, consequentemente, a resposta final seja muito similar à aplicação global do algoritmo de aprendizagem. Nesse aspecto, nosso trabalho também não está atrelado a um algoritmo de ordenação específico. Além disso, nossa análise também pode fornecer insumos para ajuste do algoritmo utilizado.

Alguns trabalhos abordam o problema de ordenção de anúncios do ponto de vista do modelo teórico que apresentamos na seção 2.1.4, que é a ordenação pela probabilidade de sucesso (i.e. impressão, clique e conversão). O trabalho de Bagherjeiran et al. [BHR10] otimiza para melhorar as conversões, embora trabalhe no modelo CPC. Utilizando regressão linear, ele induz uma ordenação parcial fraca que prioriza os anúncios com mais chance de conversão, seguidos pelos mais propensos ao clique e, por último, os que seriam apenas impressos. Como na maioria dos trabalhos mais recentes a ordenação não se limita ao uso de CTX, são incluídas duas outras *features*: a reputação do anunciante e a reputação do publicador.

3ª geração: Evolução

Um dos maiores acontecimentos recentes da Internet foi o que é informalmente chamada por muitos da era do "*Big Data*", marcada pela constatação de que podemos utilizar a enorme quantidade de dados advindos da interação dos usuários com sistemas, *sites*, aplicativos e sensores digitais para melhorar a experiência do usuário. Muitos sistemas de recomendação foram concebidos pela análise desse tipo de dado e por meio de experimentação vêm obtendo resultados excelentes. A ciência de dados é uma área recente de pesquisa que tem fornecido diretrizes claras para experimentação, análise e interpretação dos dados em diversas áreas do conhecimento. Ela tem aproximado a computação das diversas áreas de pesquisa (e.g. medicina, biologia, física, etc.) bem como a colaboração entre mercado e academia, principalmente quanto à inovação tecnológica.

Nosso trabalho, no que diz respeito a proposta de análise descrita no capítulo 4, é uma aplicação de ciência de dados ao problema de ordenação de anúncios em publicidade computacional. A criação

de sistemas de recomendação tem início na análise exploratória dos dados, seguido por rodadas de modelagem e experimentação. Nosso trabalho está inserido exatamente neste contexto: descrevemos e validamos uma análise de dados que facilita a evolução de sistemas de recomendação de anúncios. Nossa análise impõe uma única restrição sobre os dados de entrada: serem valores numéricos reais. Isso não deve ser um problema na maioria dos casos, pois facilmente podemos transformar dados categóricos em valores ou escalas numéricas. Muitos dos trabalhos relacionados citados transformam algum aspecto do contexto de publicidde em um valor numérico para inferir uma ordenação sobre os anúncios, esses são os valores que esperamos como entrada de nossa análise de dados. Nosso proposta não é a combinação propriamente dita desses valores, mas sim fornecer insumos para alterações no algoritmo de ordenação utilizado, seja ele um classificador ou uma fórmula matemática.

Alguns trabalhos já foram conduzidos para levantar o panorama atualizado da área de publicidade computacional e trazem esclarecimento sobre os caminhos que temos tomado e por onde podemos prosseguir [SM12a, SM12b]. Uma outra aplicação de engenharia e ciência de dados é o trabalho de Barford et al. [BCK+14] conduzido recentemente, que realizou uma pesquisa de campo na Internet do ponto de vista do usuário. Um *crawler* [1] visitou centenas de *websites* utilizando centenas perfis artificiais de usuários, coletou centenas de milhares de anúncios que os sistemas de publicidade exibiram. Eles identificaram que publicidade direcionada não é mais novidade, pois mais de 80% dos anúncios únicos exibidos eram direcionados ao perfil utilizado na navegação, seja por perfil demográfico (sexo e idade), geolocalização ou interesses pessoais. Por outro lado, ainda existe a prática de veiculação de ampla audiência, na qual o anúncio é exibido independente do perfil.

Dada a natureza mutante da Internet e o comportamento de seus usuários, a experimentação científica é o caminho mais seguro para a evolução consistente dos sistemas de recomenação. O trabalho [ALT+14] descreve o sistema de anúncios do *LinkedIn* e diversas das características de sua implementação. Dentre elas, destacam-se a infraestrutura distribuída para processamento em larga escala e seu sistema de experimentação para novos algoritmos de reordenação de anúncios. Os testes são feitos em tempo-real, sem a necessidade de alterações no código fonte do sistema. Essa tendência atual de experimentação também é seguida em nosso trabalho e a plataforma de experimentação que nosso SA de teste possui é descrita em [BMF14].

Em última instância, a meta que temos para sistemas de recomendação é que o mesmo tenha a capacidade de auto-adaptação automática baseada nos dados a que tem acesso. Trabalhos já foram realizados em outros problemas quanto a adaptação automática de sistemas de ordenação, e uma das técnicas mais utilizadas é programação genética [LCG+06]. O artigo [DCR+14] descreve um sistema de publicidade computacional que evolui constante e automaticamente a partir dos dados a que tem acesso. Ele faz transferência do conhecimento obtido em campanhas publicitárias anteriores para aquelas que acabaram de começar a veicular.

A evolução automática de sistemas baseada em dados vai de encontro à última mudança de paradigma ocorrida na área de publicidade: o leilão em tempo real (*real-time bidding*). Neste modelo o sistema de publicidade funciona como uma bolsa de valores, de um lado os publicadores tem espaço publicitário para vender e do outro os anunciantes desejam adquirir este inventário para veiculação de suas campanhas. Esse processo ocorre em tempo real a cada acesso de um usuário, e é intermediado por até 3 partes: a SSP (*Supply-Side Platform*), a DSP (*Demand-Side Platform*)

[1] Componente de software que navega pela Internet de forma metódica e automatizada.

e a *Exchange*. Embora publicadores e anunciantes possam construir sistemas para participar dos leilões na *Exchange*, em geral eles tercerizam esse trabalho para as plataformas de uma SSP ou DSP, respectivamente. As DSPs utilizam de diversos tipos de dados e apostam um valor de lance que julguem apropriado para cada impressão de anúncio. Nesse ambiente, as mudanças ocorrem muito rapidamente, logo, os sistemas de publicidade devem acima de tudo ser adaptativos. O trabalho de Zhang et al. [ZYW14] deriva funções de aposta adaptativas que dependem do valor de mercado de cada impressão, de modo a otimizar o valor pago por uma SSP na veiculação de publicidade. Já o artigo de Lee et al. [LODL12] utiliza regressão logística para combinar dados de conversão de campanhas de SSP, em diferentes contextos de exibição, de modo a minimizar o problema da esparsidade de dados.

Trajetória

Este trabalho de mestrado foi desenvolvido juntamente com a tese de doutorado de Marcos Broinizi, intitulada "Ordenação evolutiva de anúncios em publicidade computacional" [BF15]. Em seu trabalho, Broinizi utiliza programação genética para evoluir automaticamente funções de ordenação baseadas nas mesmas variáveis que apresentamos na tabela 2.4. As funções de ordenação são indivíduos que competem entre si por meio de uma função de *fitness*, no caso a função utilizada é a mesma função objetivo que apresentamos em 2.7. A cada dia os indivíduos melhor qualificados são reproduzidos para a geração seguinte e novos são gerados aleatoriamente por meio de mutação e reprodução. Dessa forma, são geradas centenas de fórmulas a cada semana em busca de uma solução cada vez mais adaptada ao atual ambiente de publicidade.

Nós utilizamos parte da estrutura de software desenvolvida em seu trabalho e corroboramos alguns de seus resultados. Nosso intuito inicial era construir um ordenador automático baseado em métodos de regressão e/ou agrupamento para servir de *baseline* para o trabalho de doutorado, entretanto os resultados obtidos não foram satisfatórios. A figura 3.1 mostra a dispersão 3D das variáveis CTR, CTX e BID normalizadas para um conjunto de impressões de anúncios. Os pontos vermelhos são as exibições de anúncios que receberam clique. Como se pode ver pela figura, e como constatamos em sucessivas tentativas, não há regiões com evidente concentração de cliques. Isso se deve ao fato de que para duas exibições de anúncios iguais (mesmo usuário, página e anúncios exibidos) em uma o usuário pode ter clicado e na outra não, o que caracteriza uma base cheia de informações contraditórias para a máquina de aprendizado. Além disso, o evento de clique em um anúncio é extremamente raro (e.g. 0,01% de chance) e a maioria dos modelos de classificadores não lida muito bem com isso.

Dessa maneira, como não foram obtidos resultados melhores do que as funções de ordenação feitas manualmente por analistas de negócio, alteramos o foco de nosso trabalho para entender melhor qual o papel de cada variável na função de ordenação e sua influência no desempenho global do sistema de anúncios. Nosso trabalho é complementar ao de Broinizi, visto que apresentamos um roteiro de análise que pode ser utilizado para entender as variáveis utilizadas e as funções evoluídas.

Conclusão

Em suma, a área de publicidade computacional tem evoluído muito nos últimos anos e tamanho avanço, inclusive científico, tem sido dirigido pelo interesse financeiro do mercado. A área emergiu naturalmente com o avanço da audiência na Internet e se tornou a principal fonte de receita de

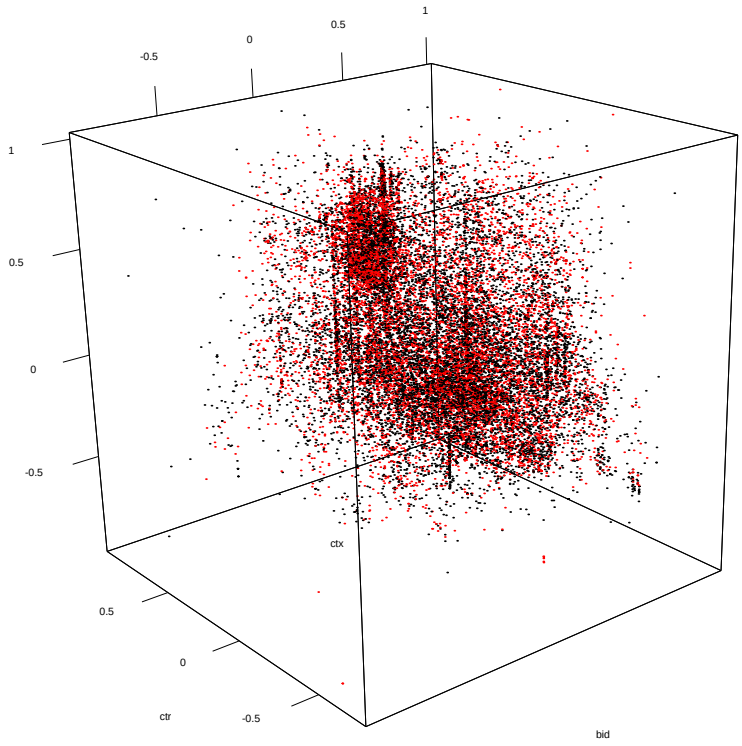


Figura 3.1: *Visualização 3D das variáveis CTR, CTX e BID normalizadas*

inúmeros produtores de conteúdo e serviços gratuitos na rede. Sua expressão como uma área de pesquisa científica surgiu principalmente quando os anúncios passaram a ser exibidos de maneira relacionada com o conteúdo da página. Mais notadamente nos sistemas de busca patrocinada, o que leva aos primeiros desafios em indexação e recuperação rápida em grandes bases de dados de anúncios. Após essa fase o foco passou a ser o enriquecimento dos modelos de reordenação de anúncios, levando a inúmeros trabalhos científicos. Por fim, em um momento mais recente averiguamos a evolução de sistemas baseada em dados, linha em que nosso trabalho está inserido. Dentre as nossas principais contribuições estão a descrição de um método de análise para evolução de sistemas de recomendação de anúncios.

A discussão de trabalhos relacionados aqui apresentada não contemplou o formato convencional de uma revisão sistemática por se tratar de uma área recente de pesquisa em rápida tranformação. Principalmente pela forte motivação de mercado, a área de publicidade computacional muda constantemente tanto nos formatos e modelos de publicidade quanto na variedade de técnicas computacionais utilizadas para a otimização de resultados. Ainda assim nossa cobertura foi extensa, abrangendo as principais conferências e periódicos relacionados, nossa discussão encontrou os principais artigos que influenciaram e desenharam a evolução da publicidade digital.

Capítulo 4

Proposta e Análises

Atualmente, uma forte tendência no mercado *online* é a utilização de dados provenientes de ambientes reais para aprimorar a modelagem de sistemas. Uma prática relativamente nova na área de sistemas mas já consagrada em outras ciências como a física e a química. Os avanços das tecnologias de transmissão de dados via rede viabilizaram a coleta e envio de dados estatísticos sobre o usuário e a utilização do sistema sem afetar negativamente a experiência de navegação. É na mineração e utilização desses dados que o mercado de publicidade digital tem vislumbrado as maiores possibilidades, e é sobre a análise deles que trata nossa proposta neste capítulo. Antes de prosseguir, relembremos a definição dada em 1.1 do problema que queremos resolver:

Como fazer alterações na função de ordenação de anúncios baseado em dados reais a fim de melhorar o desempenho do sistema de publicidade computacional?

Com esta pergunta em mente, desenvolvemos uma abordagem de análise dos fatores de ordenação que provê ao especialista de domínio informações e *insights* essenciais para compreensão dos pontos fortes e fracos de seu sistema de publicidade, bem como indica caminhos de melhoria para o mesmo.

Este capítulo está organizado da seguinte maneira: na seção 4.1 delineamos o escopo em que nossas análises foram desenvolvidas, em 4.2 definimos a análise das variáveis de ordenação proposta e a seguir em 4.3 exibimos a validação da mesma através de um primeiro experimento no ambiente de produção do UOL Cliques. Uma vez apresentada e validada nossa proposta de análise, em decorrência dos *insights* obtidos pelo primeiro experimento, enriquecemos nosso trabalho apresentando análises de outros dois experimentos em 4.5 e 4.6.

4.1 Escopo do trabalho

Nos capítulos anteriores discutimos de maneira abrangente a área e o mercado de publicidade computacional. Entretanto, as análises propostas em nosso trabalho foram concebidas e validadas em um único sistema de publicidade: o UOL Cliques, um sistema de anúncios utilizado pelo UOL, que é um dos maiores portais da internet brasileira. Portanto, embora seja possível aplicar os conceitos e ideias aqui apresentados em outros contextos, salientamos o escopo com o qual trabalhamos:

- *Contexto de publicidade*: publicidade contextualizada (cf. seção 2.1.1)

- *Formato visual*: texto e imagem (cf. seção 2.1.2)
- *Modelo de pagamento*: CPC (cf. seção 2.1.3)
- *Modelo de ordenação*: combinação de BID, CTR, CTX e BTU (cf. tabela 2.4)
- *Dados de campanha (anatomia)*: anunciante, campanha, criativo, palavras-chave, aposta e segmentação por região geográfica, sexo e idade. (cf. tabela 2.2)
- *Garantia de entrega*: não há garantia mínima de cliques, o pagamento é pré-pago e a veiculação é feita enquanto houver orçamento para a campanha

4.2 Análise de variáveis em ordenação de anúncios

Dividimos nossa análise em duas etapas, a primeira fornece uma análise isolada de cada uma das variáveis de ordenação, e a segunda uma análise combinada. Juntas elas servem como base para os especialistas de domínio fazerem alterações nas fórmulas de ordenação.

4.2.1 Etapa 1 - Teste A/B

A primeira parte é a realização de um teste A/B para avaliação isolada de cada uma das variáveis de ordenação de anúncios. Dessa forma, o primeiro passo é definir o teste A/B que será realizado, especificando cada um dos itens presentes na tabela 2.6. Em seguida executamos a implementação do mesmo em ambiente de produção, coletamos os dados e por fim analisamos o desempenho de cada variável em relação a cada métrica.

1. Especificação do teste A/B
    - ***Métrica de avaliação***: utilizamos as métricas de avaliação descritas na secão 2.3.1 como representação do interesse de cada um dos participantes do sistema. A saber o CTR global para o interesse do usuário, o ECPM para o publicador e o ROI para o anunciante. Quando não disponível pode-se utilizar o CTR global como métrica de interesse do anunciante.
    - ***Fator***: cada uma das variáveis de ordenação disponíveis no sistema (e.g. CTR, CTX, BID, etc.). Isolamos os fatores definindo uma função de ordenação exclusiva para cada um, e acrescentamos ainda uma função de reordenação aleatória para utilizar como *baseline*.
    - ***Variante***: a implementação de cada uma das estratégias de ordenação pronta para utilização em produção.
    - ***Unidade experimental***: definição aberta. Geralmente uma das seguintes opções: requisição de página, usuário ou sessão de usuário. Note que a unidade experimental de usuário permite que sejam feitas análises que considerem a evolução da métrica de desempenho ao longo do tempo, pois os usuários permanecem atrelados à mesma variante, contabilizando desempenho sempre ao mesmo fator.

2. Execução em produção e coleta de dados
   Uma vez definido o teste A/B e sua implementação disponível, coloca-se em execução durante um período arbitrário (e.g. alguns dias) enquanto os dados são coletados. Geralmente apenas uma parte das unidades experimentais é reservada ao experimento (e.g. 20%), sendo dividida entre o grupo de tratamento e o de controle, enquanto que a maior parte permanece fora do experimento (e.g. 80%). A implementação de coleta de dados deve contemplar:

   - as métricas de avaliação (i.e. CTR global, ECPM e ROI) segmentadas por estratégia;
   - e os valores absolutos das variáveis de ordenação, pois serão necessários na etapa 2 pela análise de PCA.

3. Análise de desempenho
   Em posse dos dados coletados avaliamos o desempenho relativo de cada variante do sistema com relação às métricas de interesse. Calculamos a função objetivo de cada estratégia e determinamos qual a estratégia de ordenação mais apropriada para os objetivos do publicador. O apêndice A.1 contém o *software* utilizado para execução deste passo e pode ser utilizado livremente na reprodução deste método de análise.

4.2.2 Etapa 2 - Análise de componente principais

A segunda etapa é a análise de componentes principais (PCA, *Principal Component Analysis*) sobre as variáveis de ordenação. O conjunto de dados utilizados nesta análise são os valores de cada variável na ordenação de anúncios. As que estavam disponíveis no SA que utilizamos, por exemplo, eram BID, CTR, CTX e BTU. Esses dados devem ser coletados durante a primeira etapa.

Para um dado conjunto F de variáveis de ordenação $f_1, f_2, ...f_n \in \mathbb{R}$ disponíveis em um SA, temos que cada anúncio recuperado para exibição em uma requisição de anúncios é um vetor $a \in \mathbb{R}^n$ tal que cada $a_i, i = 1..n$ corresponde ao valor de f_i naquela requisição. Os valores $a_1..a_n$ são utilizados pelas estratégias para definir a ordenação final dos anúncios. Nesse espaço vetorial, em que cada dimensão corresponde a uma variável de ordenação, aplicamos PCA para analisar quais delas influenciam a ordenação, bem como os relacionamentos entre elas.

1. Seleção da amostra de dados (opcional)
   Os dados coletados no passo anterior geralmente tem um volume muito alto, pois são advindos de milhões de requisições por dia. Isso dificulta não só sua manipulação e processamento, mas também interfere nos resultados da análise, pois as combinações possíveis dos valores de $f_1, f_2, ...f_n$ aumentam muito, podendo influenciar a análise de PCA introduzindo relacionamentos espúrios entre variáveis. Para nós (e é provável que sempre) faz-se necessário utilizar alguma heurística para selecionar uma amostra menor dos dados.

2. Análise de componentes principais
   Nesse momento, executamos PCA no conjunto de dados obtido na etapa 1 e obtemos a projeção dos vetores da base original B na base nova B'. Por exemplo:

$$\overrightarrow{CTR} = (0,1,0,0)_B = (p_1, p_2, p_3, p_4)_{B'} = p_1\overrightarrow{PC_1} + p_2\overrightarrow{PC_2} + p_3\overrightarrow{PC_3} + p_4\overrightarrow{PC_4}$$

na qual cada $p_1, ...p_n \in \mathbb{R}$. E, algebricamente, também podemos encontrar para cada componente principal seu valor na base original, por exemplo:

$$\overrightarrow{PC_1} = (1,0,0,0)_{B'} = (v_1, v_2, v_3, v_4)_B = v_1\overrightarrow{BID} + v_2\overrightarrow{CTR} + v_3\overrightarrow{CTX} + v_4\overrightarrow{BTU}$$

As variáveis principais na ordenação são as que apresentam os maiores valores de v_i nas primeiras componentes principais, pois essas são as variáveis originais com maior dispersão dos dados.

Para facilitar a interpretação das componentes principais utilizamos um *loading plot* [1] da projeção dos vetores da base original no espaço das componentes principais. Variáveis que sejam expressivas e bem alinhadas no gráfico têm alta correlação, enquanto que variáveis que apontem em direções diferentes são mais independentes. Além disso, variáveis de ordenação que sejam expressivas nas primeiras componentes principais são os que apresentam maior dispersão e que, portanto, mais influenciam na discriminação (ordenação) dos anúncios. O apêndice A.2 contém o *software* utilizado para execução deste passo e pode ser utilizado livremente na reprodução deste método de análise.

3. Conclusão e tomada de decisão
   Por fim, em posse dos resultados do teste A/B e do PCA, somos capazes de resolver nosso problema original. Podemos fazer alterações na função de ordenação de anúncios de modo a otimizar o desempenho do sistema de acordo com nossos objetivos. Identificamos as variáveis com melhor impacto na métrica que desejamos melhorar, analisamos sua correlação com outras, e então aumentamos sua influência no resultado final de ordenação.

4.3 Experimento 1: validação da análise

Para confirmar a validade da análise proposta na seção anterior 4.2 conduzimos um experimento no ambiente de publicidade computacional do UOL Cliques. Definimos e executamos o teste A/B, coletamos os dados necessários, e aplicamos PCA. Com base nesses insumos extraímos conclusões e sugerimos uma alteração na função de ordenação que foi adotada.

Características do SA utilizado no teste

A validação de nossa proposta foi realizada na rede de anúncios do maior portal de conteúdo da Internet brasileira. Nele há três tipos principais de publicadores de anúncios: os canais, os parceiros e os afiliados. Os canais são seções do portal UOL com conteúdo específico, como notícias, esportes, entretenimento, entre outros; os parceiros são sites de nicho que possuem parceria de audiência; e os afiliados são quaisquer sites que integraram às suas páginas uma seção para exibição de conteúdo publicitário por meio da rede. 2.1.1 Essa ampla gama de sites para exibição de anúncios atrai anunciantes dos mais variados, empresas de pequeno, médio e grande porte de todos os ramos de atuação.

O SA está submetido a fortes restrições de desempenho, como volume de requisições e tempo de resposta. Ele é responsável por atender mensalmente centenas de milhões de usuários únicos e mais

[1] Tipo de gráfico comum para interpretação de componentes principais. Veja na figura 4.4

de 5 bilhões de requisições de anúncios. A configuração e alteração das diferentes estratégias de ordenação é feita por meio de uma plataforma de configurações remotas, descrita em mais detalhes no trabalho de Broinizi e Ferreira [BMF14].

4.3.1 Validação da Etapa 1

Especificação do teste A/B

- ***Métricas de avaliação***: *CTR global* e *ECPM*. ROI não está disponível;
- ***Fatores***: *BID*, *CTR*, *CTX* e *BTU*, que são todos os disponíveis;
- ***Variantes***: *stBID*, *stCTR*, *stCTX*, *stBTU* e *stRND*. As quatro primeiras são as variantes de cada fator e a última utiliza uma função de reordenação aleatória para ser utilizada como *baseline*;
- ***Unidade experimental***: requisições.

Execução em produção e coleta de dados

O teste A/B foi executado ao longo de sete dias. Cada variante recebeu uma amostra aleatória de mesmo tamanho de todas as requisições feitas ao SA, totalizando mais de 300 milhões de anúncios exibidos. Os dados de impressão e clique dos anúncios exibidos foram armazenados em arquivos texto, processados e consolidados por uma aplicação *Java* e então armazenados em um banco de dados acessível para consulta.

Análise de desempenho

A figura 4.1 apresenta o gráfico do desempenho diário normalizado de cada estratégia em relação às métricas de CTR e ECPM. A estratégia com a melhor taxa de clique foi stCTR e a que gerou mais retorno financeiro foi stBID. A alta taxa de clique em stCTR a fez superar o patamar de retorno financeiro do baseline, em contrapartida, mesmo stBID tendo a pior taxa de clique, excedeu em muito as outras estratégias em termos de ECPM. Enquanto a estratégia stCTX obteve resultado levemente superior a stRND tanto em CTR quanto em ECPM, stBTU se mostrou bem semelhante: em termos de CTR sem um padrão definido; e em termos de ECPM equiparados em um patamar baixo.

A figura 4.2 mostra o gráfico da função objetivo para cada uma das estratégias avaliadas. Utilizamos a fórmula descrita na equação 2.7, e os parâmetros foram $\alpha = \beta = \gamma = \frac{1}{3}$ e $\delta = 1$. Nesta configuração, a estratégia stCTR alcançou desempenho muito superior do que todas as outras e seria nossa melhor escolha para utilizar como função de ordenação oficial. Embora stBID tenha se destacado na métrica de ECPM, não alcançou bons resultados de Fobj devido ao parâmetro de regularização $\delta\sigma^2$ que penaliza a discrepância nos resultados das métricas. A figura 4.3 ilustra a influência desse parâmetro utilizando $\delta = 0$, nela stBID tem desempenho equiparado a stCTR. Para fins acadêmicos esta configuração de parâmetros está bem apropriada mas, como dissemos, a mesma é de total controle do publicador segundo seus interesses.

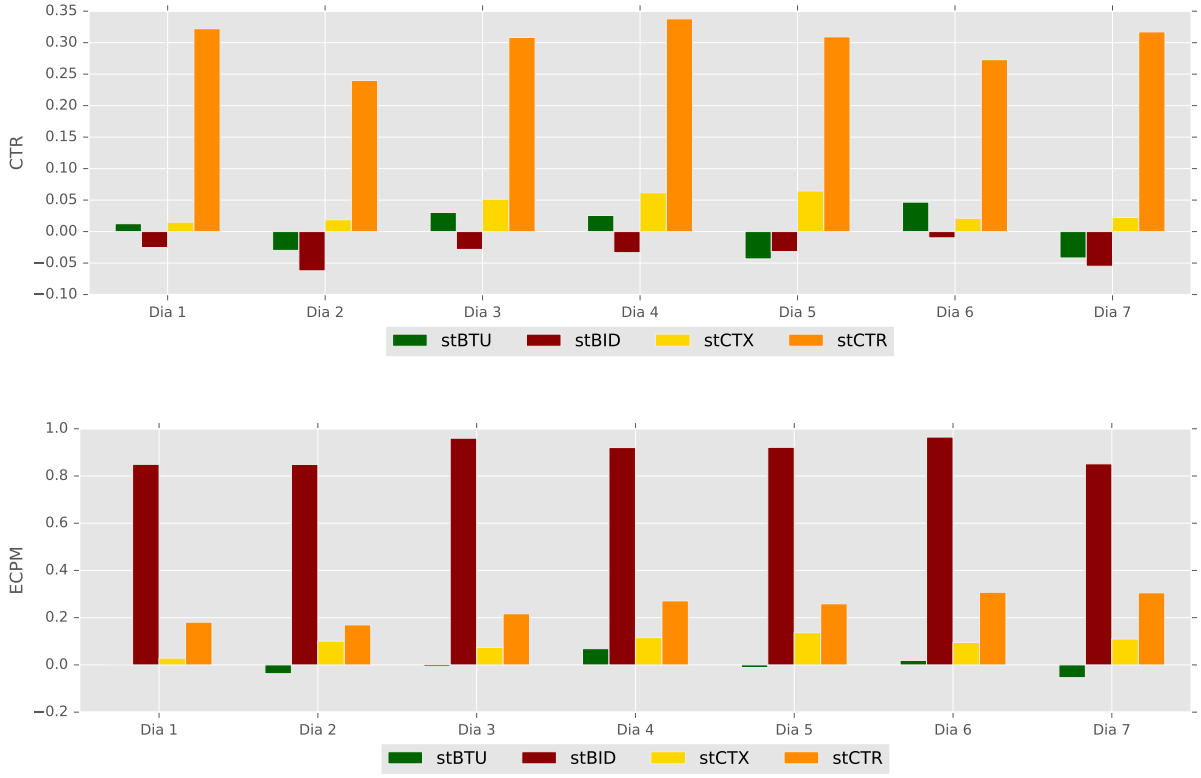


Figura 4.1: *Desempenho diário das estratégias no teste A/B.*

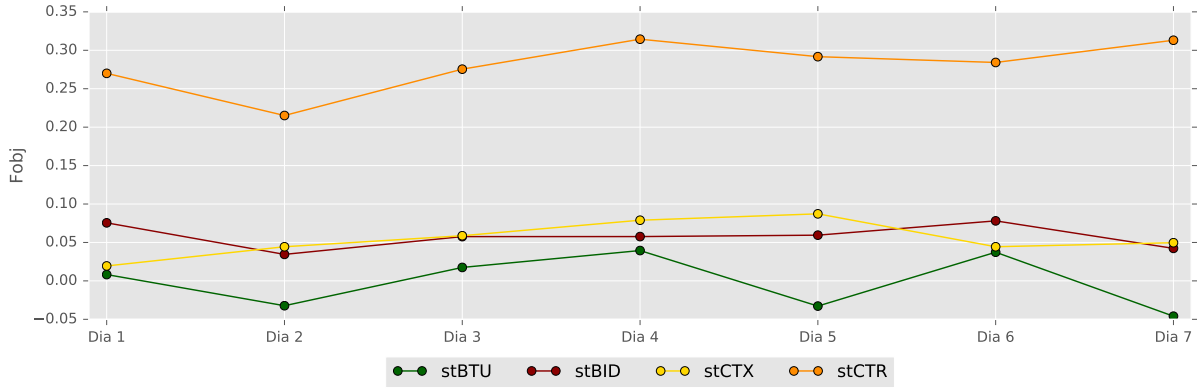


Figura 4.2: *Valor diário da função objetivo para os parâmetros* $\alpha = \beta = \gamma = \frac{1}{3}$ *e* $\delta = 1$

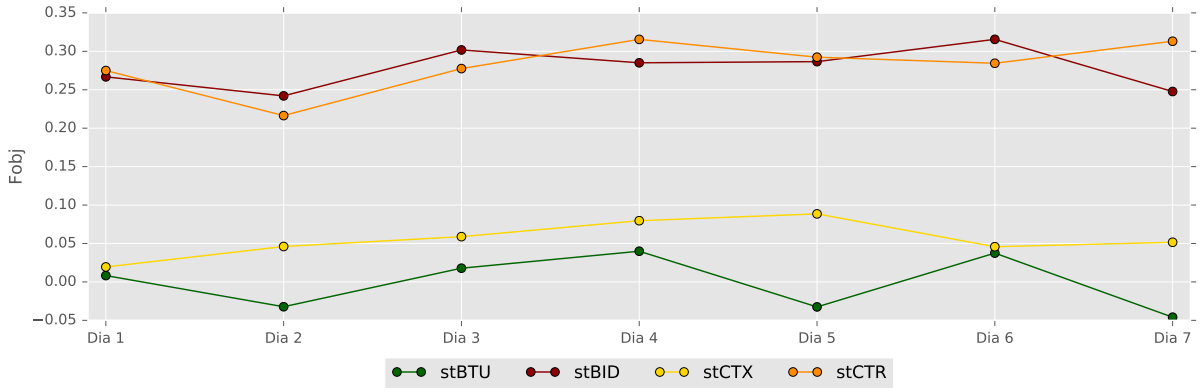


Figura 4.3: *Valor diário da função objetivo para os parâmetros* $\alpha = \beta = \gamma = \frac{1}{3}$ *e* $\delta = 0$

4.3.2 Validação da Etapa 2

Seleção da amostra de dados

O conjunto de dados de que dispunhamos para a análise PCA compreendia 33% de todas as requisições feitas ao SA em um período de sete dias. Entretanto esse dado apresenta muita dispersão e os resultados da análise de PCA foram inconclusivos. Adotamos então uma heurística para seleção da amostra de dados a ser utilizada: consideramos apenas os dados provenientes de anúncios que receberam cliques e os exibidos juntamente com eles.

Nossa intuição é de que esses dados expressam melhor as características de nosso problema, pois selecionam apenas anúncios que tiveram interação com o usuário. O que é uma boa forma de descartar anúncios que foram exibidos, mas nunca sequer vistos pelo usuário. Com isto reduzimos nosso conjunto de dados para aproximadamente 22500 pontos, cerca de 150 vezes menos.

Análise de componentes principais

O cálculo das componentes principais, bem como sua visualização foram realizadas por uma aplicação construída na linguagem R [2]. A tabela 4.1 mostra a projeção de cada uma das variáveis de ordenação na nova base de componentes principais. A preservação da variância original é de: 30.5% em PC1, 25.8% em PC2, 24.8% em PC3 e 18.6% em PC4.

	PC1	***PC2***	***PC3***	***PC4***
BID	-0.72014168	0.1376585	-0.02744383	-0.6794799
CTR	0.05708159	-0.4651475	-0.87528717	-0.1193811
CTX	0.68025155	0.3323004	-0.04331807	-0.6518879
BTU	0.12407854	-0.8088661	0.48087707	-0.3147974

Tabela 4.1: *Matriz de transofrmação dos vetores da base original na nova base.*

Como apresentado no roteiro de análise, para facilitar a interpretação da análise de PCA, utilizamos a exibição de um gráfico de *loading plots*. A figura 4.4 apresenta a projeção dos vetores da base original (BID, CTR, CTX e BTU) em dois planos definidos pela nova base (PC1, PC2, PC3 e PC4).

[2] Mais informações em http://www.r-project.org/, acessado em maio de 2015.

No plano de PC1×PC2, vemos que as duas primeiras componentes principais são compostas por uma combinação de BID, CTX e BTU, e uma expressão menor de CTR. A diferença de direção entre BID, CTX e BTU revela que essas variáveis são pouco correlacionadas, já o alinhamento de CTR com BTU revela a utilidade da captura de interesses do usuário para atrair sua atenção. As duas primeiras componentes principais cobrem apenas 56,3% de toda a variância, portanto necessitamos analisar também as duas últimas. No plano de PC3×PC4, vemos que a terceira componente (PC3) é composta principalmente por CTR, o que, em conjunto com a diferença de direção de BID, CTX e BTU no plano de PC1×PC2, revela que as quatro variáveis são pouco correlacionadas e capturam aspectos diferentes do contexto. A quarta componente (PC4) está em uma direção de dispersão residual das variáveis BID, CTX e BTU, e revela a existência de uma relação direta entre o valor pago pelo anunciante (BID) e a contextualização entre o anúncio e a página (CTX).

Conclusão e tomada de decisão

Neste momento somos capazes de extrair conclusões dos resultados obtidos e sugerir melhorias aos especialistas de domínio. Podemos destacar as seguintes:

Conclusão	***Ação sugerida***
CTR é a variável de ordenação que tem maior impacto para o usuário.	Para favorecer o interesse dos usuários e anunciantes, aumente o peso da variável CTR na fórmula de ordenação.
Pouca correlação entre as variáveis.	Por capturarem diferentes contextos de exibição, é uma boa prática sempre utilizar todas as variáveis em conjunto.
Correlação pequena entre BTU e CTR, e pouca eficiência de BTU.	Outros trabalhos [FKLT12, TLY+11] reportam aumento significativo na taxa de clique com a utilização de técnicas de captura do comportamento do usuário (relativo a BTU). É preciso melhorar a implementação da variável BTU.

Tabela 4.2: *Conclusões obtidas na validação de nossa proposta*

As três conclusões listadas acima foram confirmadas pelos especialistas. Fez-se então uma comparação de desempenho da fórmula oficialmente utilizada com uma variante em que o peso da variável CTR fosse mais alto. Observou-se ganho de desempenho na métrica de CTR global sem comprometer em nada a métrica de ECPM, culminando, assim na modificação da fórmula oficial em produção para o SA. O alto desempenho da variável CTR na métrica de CTR global indica que podemos olhar para o problema de exibição de publicidade do ponto de vista de recomendação, no qual usuários recomendam anúncios para outros que tenham comportamento semelhante. Tendência já apontada no artigo de Garcia-Molina [GMKP11].

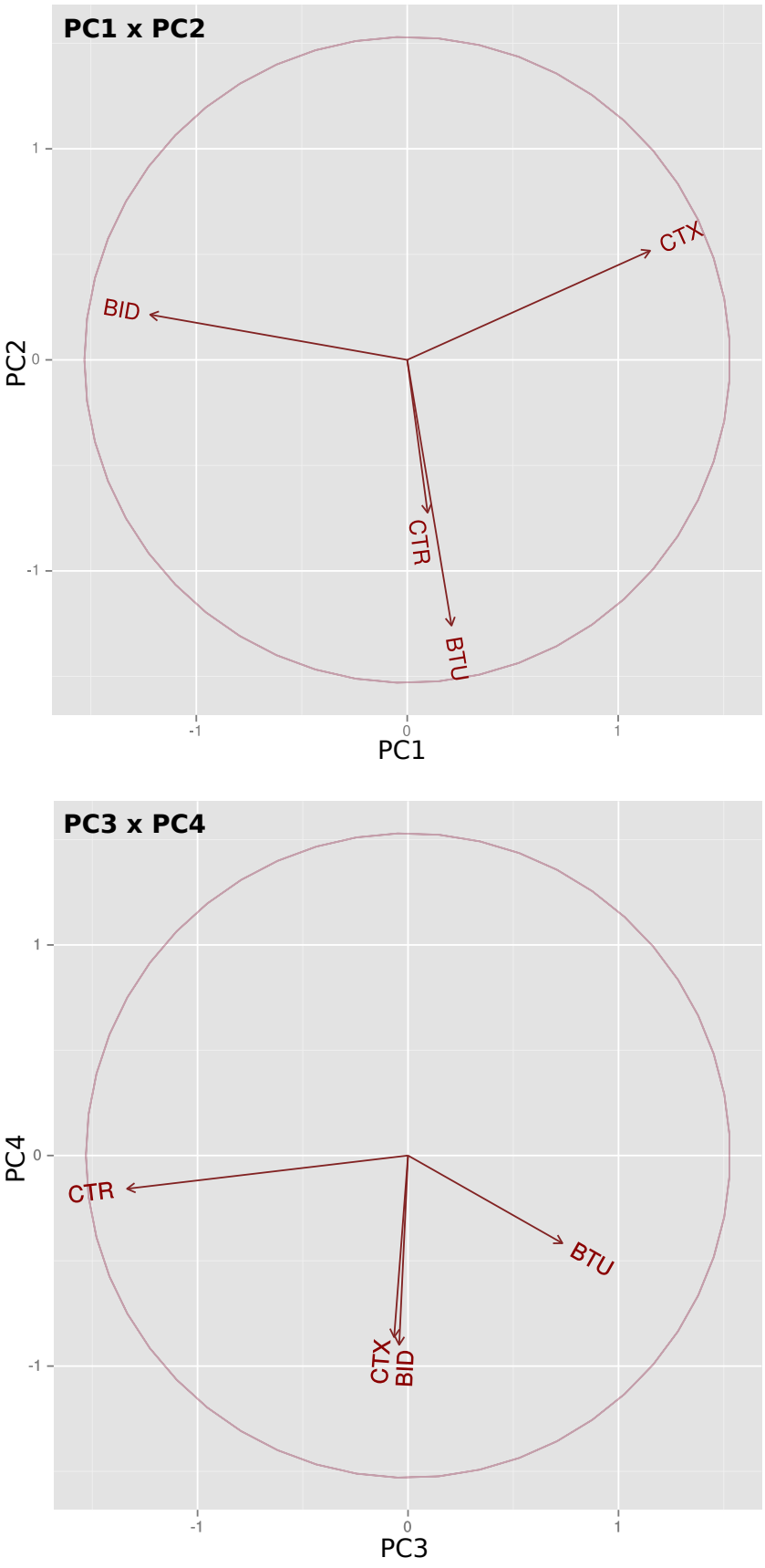


Figura 4.4: *Loading plots das componentes principais*

4.4 Relevância da proposta

Nossa proposta de análise é diretamente aplicada ao mercado e fornece aos administradores dos sistemas de anúncios uma base para compreensão de suas forças e fraquezas. Com ela eles tornam-se muito mais capazes de entender as características de sua audiência e de traçar um plano para otimizar os resultados das campanhas de seus anunciantes. Ela é principalmente relevante no cenário de transição entre sistemas de publicidade da segunda para terceira geração ('Reordenação' e 'Evolução' repectivamente, ambas discutidas no capítulo 3), pois fornece insumos para evolução de fácil interpretação e resultado aplicado comprovado. Ainda que a evolução não ocorra de maneira automatizada, é essencial que ela ocorra pois, sem evolução uma empresa de publicidade online perde mercado muito rapidamente devido a extrema competitividade da concorrência em um mercado que avança rapidamente.

4.5 Experimento 2: combinação de variáveis

O primeiro experimento (seção 4.3) mostrou a influência de cada variável no desempenho global do sistema, uma pergunta natural que se pode fazer é o que acontece quando combinamos duas ou mais variáveis em uma única fórmula de ordenação. Para obter essa resposta realizamos um segundo experimento, trouxemos as duas melhores estratégias do primeiro experimento para competir com estratégias tradicionais de ordenação, e que combinam o uso de diversas variáveis. A tabela 4.3 mostra cada estratégia, sua fórmula e respectiva descrição. Este experimento durou sete dias e coletou dados de mais de 300 milhões de anúncios exibidos.

Estratégia	***Fórmula***	***Descrição***
stCTR	CTR	Estratégia vencedora em CTR global no primeiro experimento.
stBID	BID	Estratégia vencedora em ECPM no primeiro experimento.
stTRAD	$CTR * BID$	Estratégia tradicional de reordenação criada pelo Google em 2002.
stUSER	$\frac{CTR-\overline{CTR}}{(\Delta CTR)} + \frac{CTX-\overline{CTX}}{(\Delta CTX)} + \frac{BTU-\overline{BTU}}{(\Delta BTU)}$	Estratégia que representa apenas o interesse do usuário, combinando as variáveis CTR, CTX e BTU.
stMKT	*confidencial*	Uma estratégia de mercado real, que combina todas as variáveis disponíveis. Estratégia oficial do UOL Cliques.

Tabela 4.3: *Estratégias de ordenação do segundo experimento.*

A figura 4.5 apresenta o gráfico normalizado do desempenho diário de cada estratégia em relação às métricas de CTR e ECPM. A estratégia stUSER obteve resultado levemente inferior ao de stCTR com relação à taxa de clique e equivalente em ECPM. Esperávamos um desempenho superior em clique uma vez que a stUSER agrega todas as variáveis de interesse do usuário, mas não foi o que aconteceu. Uma hipótese que não pudemos validar é se os cliques gerados por stUSER

são mais qualificados do que os de stCTR, pois para isso precisávamos analisar a métrica de ROI do anunciante. A estratégia stTRAD foi a mais equilibrada com relação às duas métricas de desempenho, mantendo patamares relativamente altos tanto em clique quanto em retorno financeiro. A estratégia stMKT por sua vez, obteve o melhor resultado em ECPM, superando stBID não somente em desempenho financeiro, mas também em taxa de clique (quase duas vezes superior). Com isso, validamos que a reordenação de anúncios com fórmulas baseadas nessas variáveis pode produzir ganhos para cada um dos participantes simultaneamente.

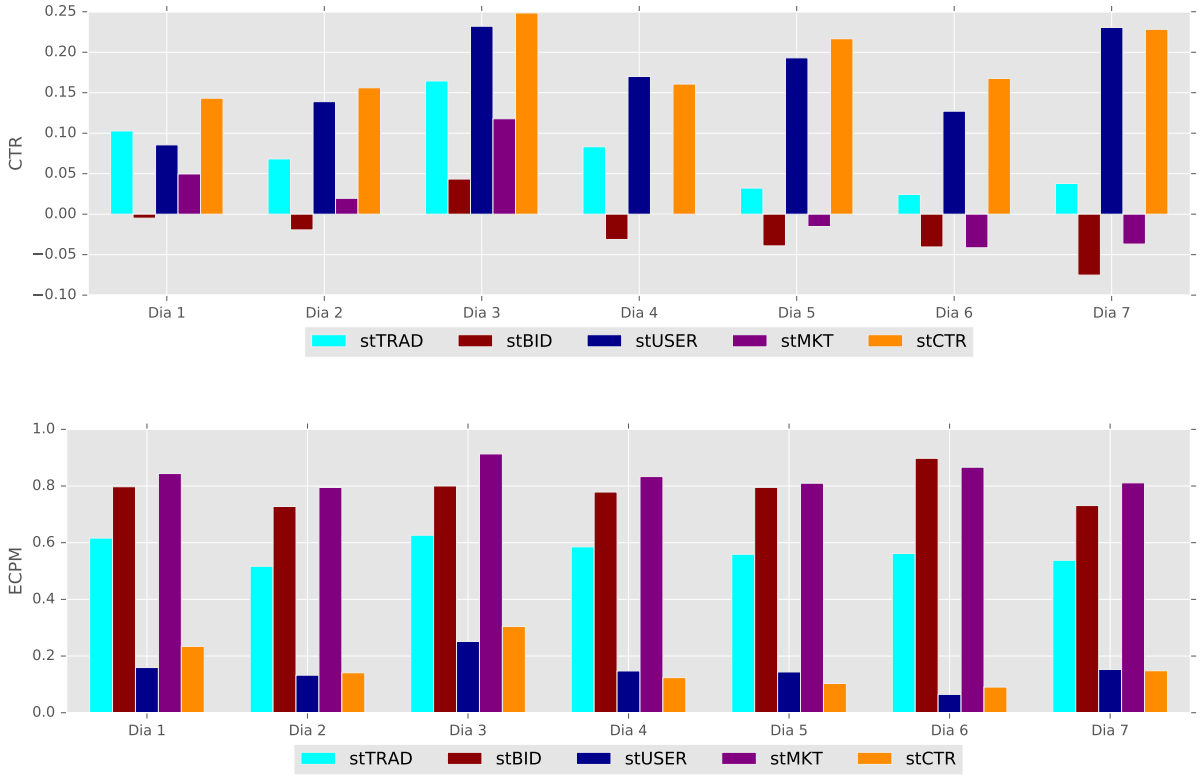


Figura 4.5: *Desempenho diário das estratégias no experimento 2.*

A figura 4.6 mostra o gráfico da função objetivo para cada uma das estratégias utilizando os parâmetros $\alpha = \beta = \gamma = \frac{1}{3}$ e $\delta = 1$. Todas as estratégias tiveram desempenho positivo moderado, oscilando no patamar de 10% a 20% de ganho em relação ao *baseline*. Comparando com o desempenho mais baixo de várias estratégias do experimento 1 (que utilizam uma única variável, c.f. 4.2), concluímos que a combinação das variáveis de ordenação eleva os resultados de função objetivo, ou seja, traz ganhos ao sistema de publicidade como um todo, e a cada um de seus participantes. As estratégias com melhores desempenhos foram stTRAD e stCTR, seguidas por stUSER. As estratégias stMKT e stBID, que têm um foco mais acentuado no retorno financeiro, obtiveram os resultados mais baixos de função objetivo. À longo prazo isto pode indicar uma degradação do sistema, levando a um maior retorno financeiro a curto prazo, mas a uma diminuição do interesse do usuário pelo conteúdo publicitário (o terceiro experimento 4.6 traz mais detalhes sobre isto). A figura 4.7 mostra o mesmo gráfico da função objetivo para cada uma das estratégias mas utilizando $\delta = 0$. Com isso, as estratégias de ordenação stBID e stMKT ganham destaque, sendo que stMKT obtém o melhor desempenho, e stBID a segunda posição. A definição dos parâmetros é responsabilidade

do publicador (dono do sistema de anúncios), mas vale notar que a estratégia stTRAD mostrou desempenho satisfatório independente do valor de δ. Isso se explica por que a sua fórmula é uma simples multiplicação entre CTR e BID, as variáveis que, como vimos no primeiro experimento 4.3.1, tem a maior influência nas métricas de CTR global e ECPM respectivamente. Assim sendo, utilizar stTRAD é um ótimo ponto de partida para a implementação de sistemas de publicidade computacional.

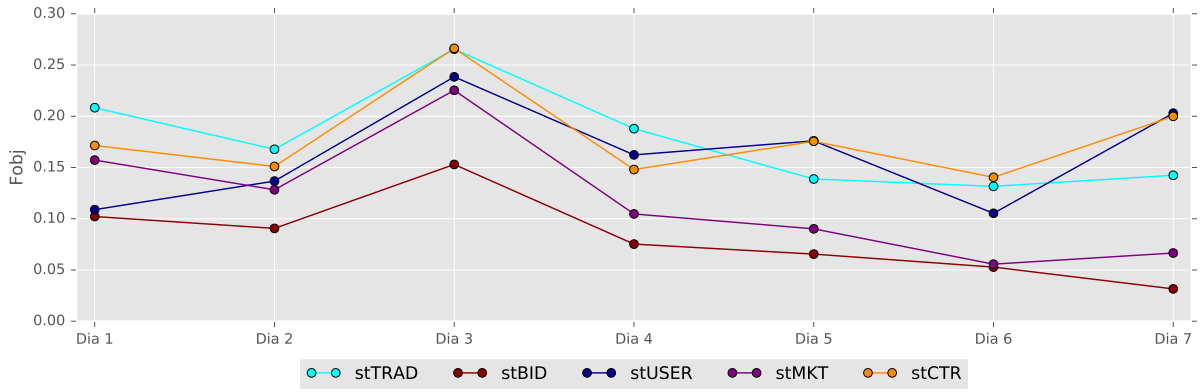


Figura 4.6: *Valor diário da função objetivo para os parâmetros* $\alpha = \beta = \gamma = \frac{1}{3}$ *e* $\delta = 1$ *no experimento 2.*

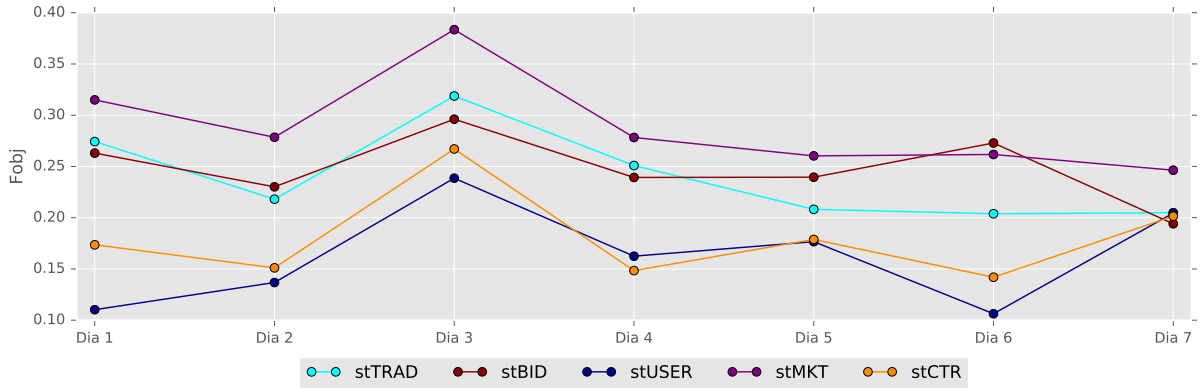


Figura 4.7: *Valor diário da função objetivo para os parâmetros* $\alpha = \beta = \gamma = \frac{1}{3}$ *e* $\delta = 0$ *no experimento 2.*

4.6 Experimento 3: desempenho ao longo do tempo

Os dois primeiros experimentos utilizaram como unidade experimental as requisições de anúncios. Dessa maneira, o mesmo usuário fica sujeito a diferentes variantes do teste A/B, sendo exposto a diferentes estratégias de ordenação. Tais testes trazem resultados válidos, mas limitados em análises de longa duração, pois os efeitos que uma estratégia tem sobre o usuário são diluídos nos efeitos de outra. Então, para realizarmos uma avaliação dos efeitos que a exposição contínua de uma estratégia tem sobre o usuário, fizemos uma modificação no sistema de publicidade do UOL Cliques para que fosse possível a veiculação de anúncios segmentados por grupos de usuários.

A implementação realizada divide todos os navegadores únicos que trafegam pelas páginas dos publicadores em grupos de usuários representativos de pequena parte do tráfego. Dividimos todos

os usuários em 256 grupos diferentes (representados por dois dígitos hexadecimais) e possibilitamos a escolha da estratégia de ordenação para cada grupo. Com isso, tornou-se possível a realização de experimentos tendo como unidade experimental o usuário, possibilitando, portanto, a análise de cada estratégia ao longo do tempo.

A configuração deste experimento, contemplou nove estratégias de ordenação: stRND (*baseline*), stBID, stCTR, stCTX, stBTU, stTRAD, stUSER, stMKT e stLCTR. As oito primeiras são as mesmas que já foram apresentadas nos experimentos anteriores, e stLCTR é uma ordenação única pela variável LCTR, que foi adicionada ao sistema durante o desenvolvimento deste trabalho. A variável LCTR armazena o CTR de cada anúncio segmentado por domínio, ou o CTR local em cada site (*Local CTR*). Seu uso e desenvolvimento pressupõe que há sites em que tanto a audiência quanto o conteúdo estão mais relacionados a determinados grupos de anúncios, e, por isso, a priorização dos mesmos nesses locais trará melhor desempenho na campanha de publicidade. Em requisições em que ainda não há dados suficientes para que o cálculo de LCTR seja considerado confiável, o valor LCTR é simplesmente o valor do CTR global daquele anúncio (i.e. variável CTR).

Mapeamos 5 grupos de usuários para cada estratégia, de modo que cada uma receberá as requisições feitas por aproximadamente 2% dos usuários ($\frac{5}{256}$). Como a distribuição dos usuários nos grupos é uniforme e aleatória, isto deve representar um número equivalente das requisições totais feitas ao sistema para cada estratégia. E além disto, pelo mesmo motivo, esta escolha não introduz nenhum viés estatístico nos resultados.

A figura 4.8 mostra os desempenhos diários alcançados pelas estratégias do experimento 3 nas métricas de CTR global e ECPM. Os dados do 13º dia do experimento não estão disponíveis devido a um problema ocorrido no sistema de contabilização dos cliques; entretanto isso não deve interferir em nossas conclusões. A estratégia stMKT obteve os melhores resultados tanto na métrica de CTR quanto de ECPM, o que representa melhores resultados para todos os participantes do sistema de publicidade. Algo que devemos ressaltar é que, se compararmos o desempenho de stMKT nos experimentos anteriores, veremos que ela sofreu uma elevação significativa na métrica de CTR global. Isto se explica pela ação tomada como resultado da análise do experimento 1 (c.f. 4.3.2). [3]

A segunda fórmula com melhor desempenho em cliques foi stLCTR, o que comprova que a segmentação da variável CTR por site trouxe ganhos em cliques, entretanto há uma leve queda no desempenho financeiro (veja a figura 4.9, que mostra o mesmo gráfico apenas para as estratégias stCTR e stLCTR). Isto pode acontecer por que um anúncio bem segmentado em relação a um site não necessariamente é o que paga mais por um clique (BID). Na verdade, em uma estratégia que faça uso da variável de BID (realidade do mercado), os anúncios mais bem segmentados por site precisarão apostar menos por clique (BID) para conseguir ganhar o leilão e conseguir uma impressão.

A figura 4.10 mostra o gráfico da função objetivo para cada uma das estratégias utilizando os parâmetros $\alpha = \beta = \gamma = \frac{1}{3}$ e $\delta = 1$. As estratégias stCTR, stLCTR e stUSER foram as únicas que obtiveram desempenho predominantemente positivo, apresentando ganhos entre 5% e 45%. Isso já era esperado devido a influência da variável CTR já explanado nas seções anteriores, e a supresa fica por conta do pequeno destaque de stLCTR, que obteve o melhor resultado. As estratégias stBTU e stCTX obtiveram desempenho equivalente ao baseline, flutuando ora acima ora abaixo do

[3] Esta elevação não foi observada no segundo experimento pois os resultados ainda não tinham sido apresentados para a área de negócio e a fórmula oficial (stMKT) ainda não havia sido modificada.

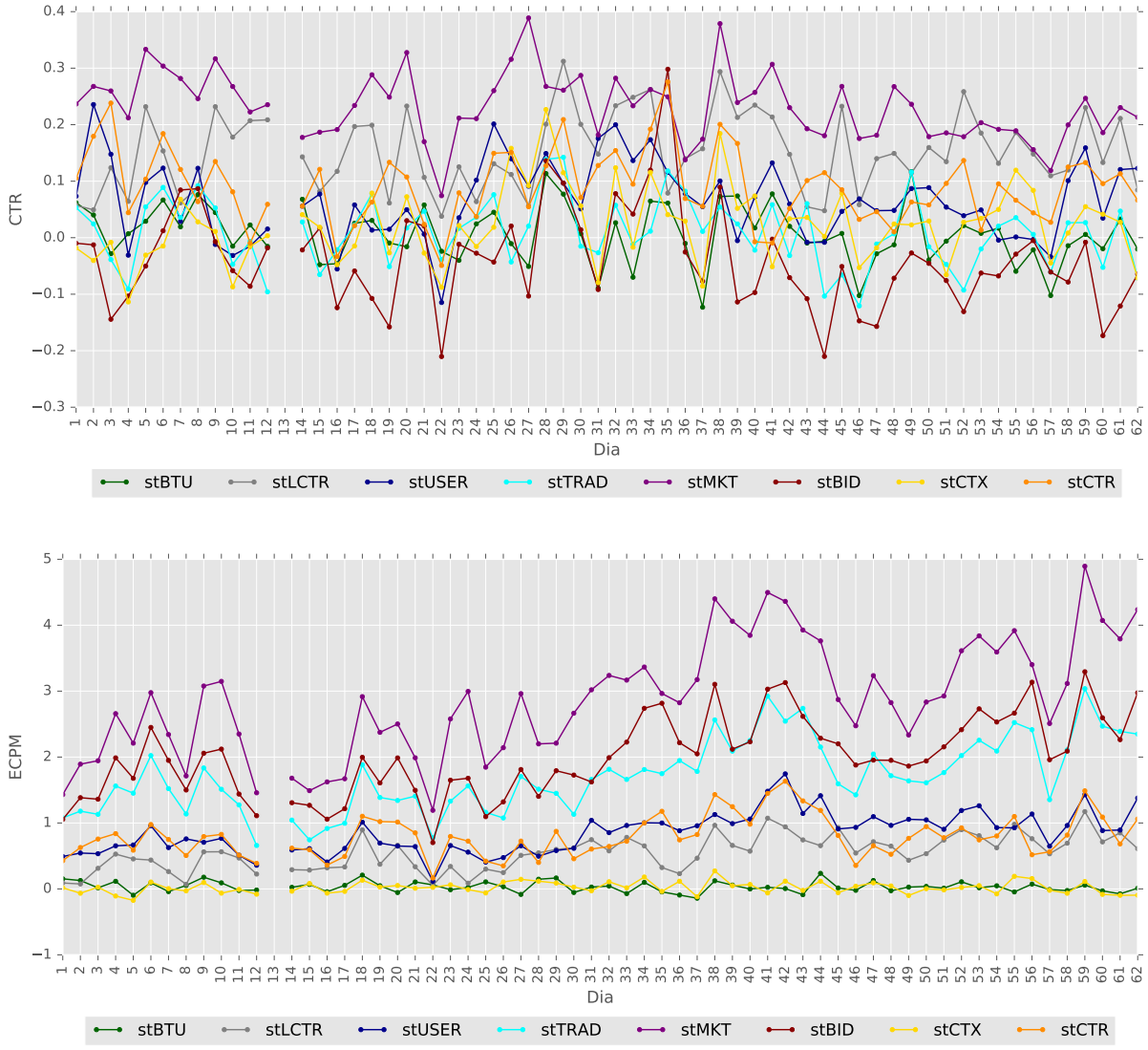


Figura 4.8: *Desempenho diário das estratégias no experimento 3.*

Figura 4.9: *Desempenho diário das estratégias stCTR e stLCTR no experimento 3.*

mesmo, mas sem comportamento consistente. Todas as estratégias que utilizam a variável BID em sua fórmula (stBID, stTRAD e stMKT) obtiveram desempenhos de função objetivo cada vez piores com o passar do tempo. Entretando, se removermos a necessidade de equilíbrio entre os interesses dos participantes ($\delta = 0$), veremos estas mesmas estratégias obtendo os melhores desempenhos com o passar do tempo (veja figura 4.7), sendo o grande destaque para a estratégia stMKT.

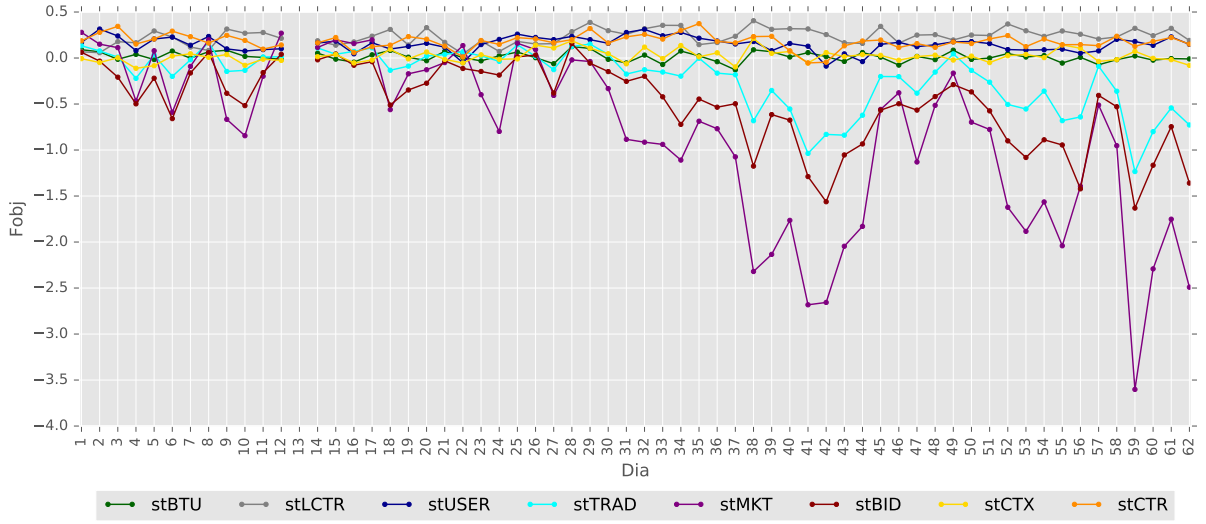


Figura 4.10: *Valor diário da função objetivo para os parâmetros* $\alpha = \beta = \gamma = \frac{1}{3}$ *e* $\delta = 1$ *no experimento 3.*

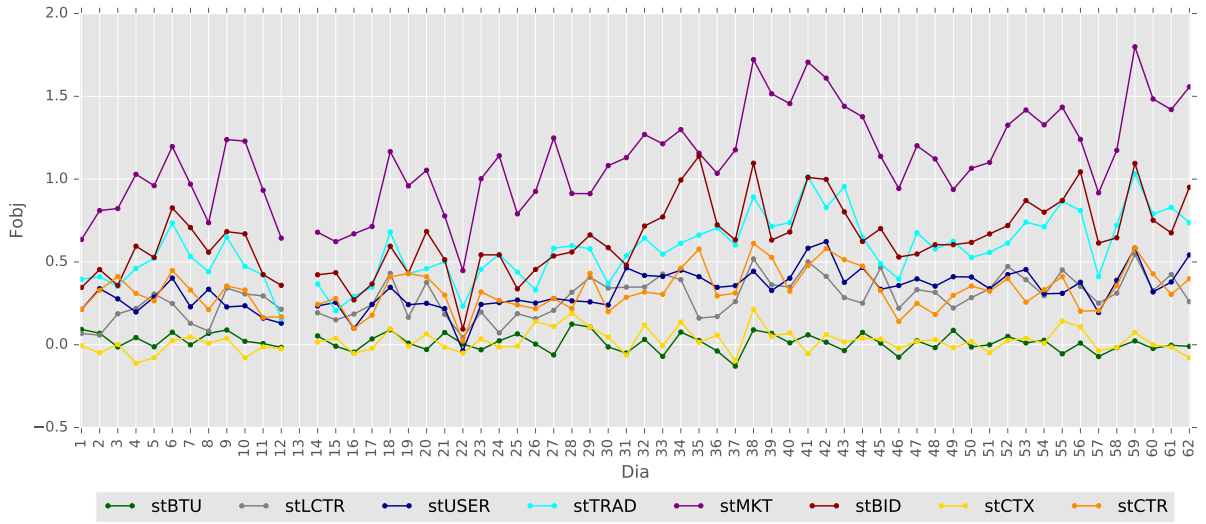


Figura 4.11: *Valor diário da função objetivo para os parâmetros* $\alpha = \beta = \gamma = \frac{1}{3}$ *e* $\delta = 0$ *no experimento 3.*

Poderia-se argumentar que este efeito ocorre devido a uma degradação da estratégia de baseline, ao supor que com o passar do tempo um usuário exposto a ordenação aleatória perde o interesse nos anúncios exibidos por não estarem bem relacionados com seu interesse. Entretanto não é isto que acontece, os ganhos em ECPM ocorrem devido a um fato conhecido dos administradores do UOL Cliques, que é a sazonalidade de campanha ocorrida no final de cada mês. Para um melhor entendimento desse fato apresentamos desempenho em impressões, cliques e receita das estratégias

stRND (baseline), stLCTR e stMKT (vencedoras em Fobj com $\delta = 1$ e $\delta = 0$ respectivamente). Como os valores absolutos são sigilosos, apresentamos na figura 4.12 os dados absolutos obfuscados por uma transformação linear, ou seja, multiplicamos todos os valores por um escalar fixo não informado.

O gráfico de impressões mostra que cada estratégia recebeu um número equivalente de requisições de anúncios, sendo que os vales ocorrem devido à diminuição do tráfego durante os finais de semana. Os dias de número 25 e 56 do experimento 3 são os primeiros dias de julho e agosto respectivamente. Dessa forma, no gráfico de receita, podemos notar nitidamente o efeito de sazonalidade que ocorre nos finais de mês. O comportamento comum dos anunciantes é que ao longo do mês as campanhas vão sendo configuradas, sendo que frequentemente o dia escolhido para término da veiculação é o último dia do mês. Isso explica a redução brusca de receita obtida em todas as estratégias do dia 24 para 25 e do dia 55 para 56. Um outro ponto a elucidar é o fato de que há uma diferença crescente nos valores ECPM (e, portanto de Fobj) entre as estratégias que fazem uso da variável BID e as que não fazem. Isto ocorre porque os administradores das campanhas publicitárias tendem a aumentar o valor de BID nos finais de mês para utilizar todo o orçamento alocado e aumentar a entrega de seus anúncios. Com essa análise confirmamos o efeito de sazonalidade de final de mês e refutamos a suspeita de degradação da estratégia de baseline.

Analisando o gráfico de receita e de cliques podemos concluir que, de fato a melhor escolha para função de ordenação oficial seria stMKT, pois a mesma obtém resultados consistentemente superiores a stLCTR tanto em CTR global quanto em ECPM. Isso nos mostra que a escolha dos parâmetros de Fobj deve ser feita com cuidado pois pode esconder determinadas nuances das estratégias analisadas.

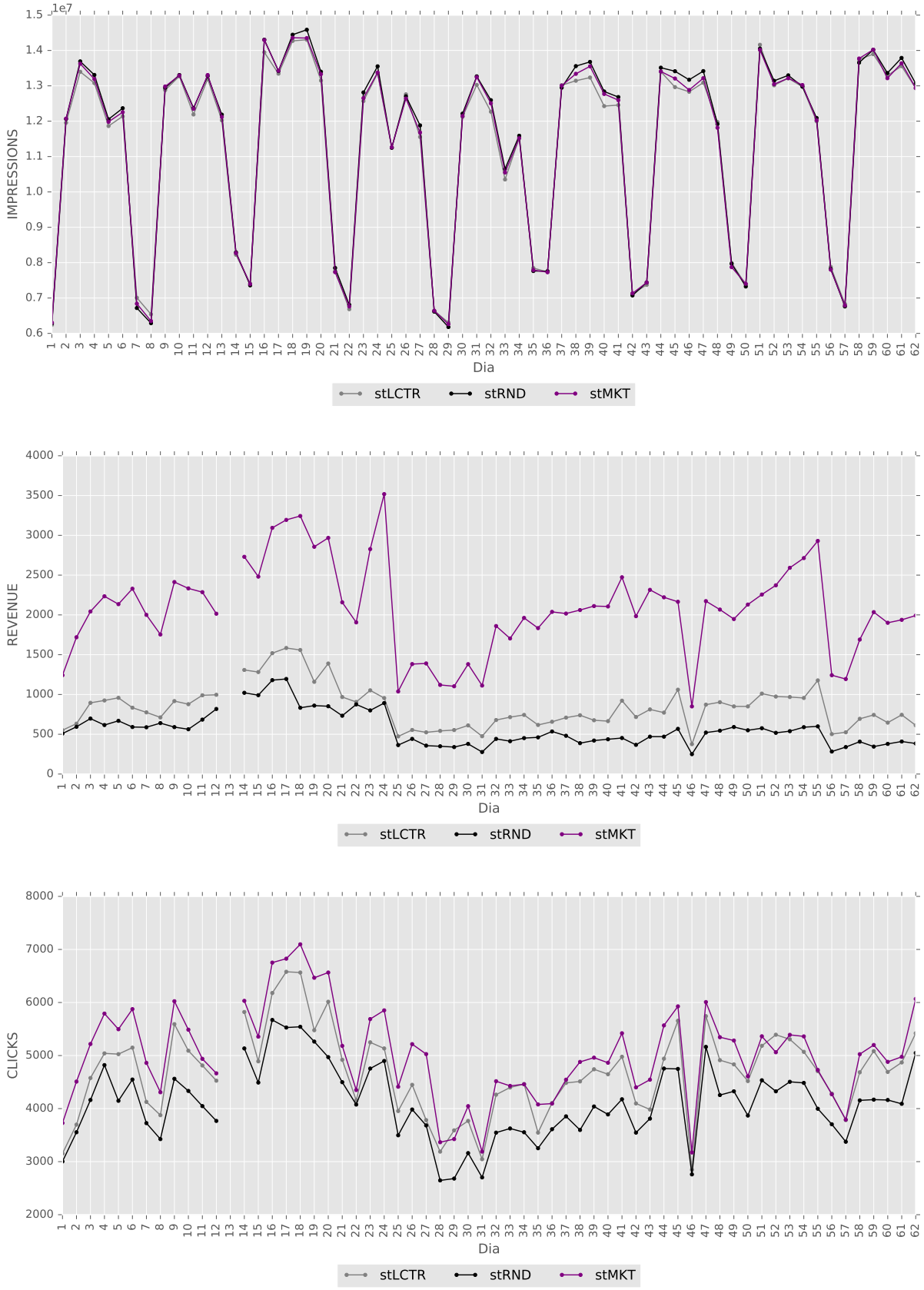


Figura 4.12: *Valores obfuscados diários de impressões, cliques e receita das estratégias vencedoras do experimento 3.*

Capítulo 5

Conclusão

O mercado de publicidade computacional cresce a cada ano desde que surgiu nos primórdios da Internet. A veiculação de publicidade foi um dos principais financiadores para a disponibilização de de serviços e conteúdo sem custo ao usuário e sem dúvidas um dos principais fatores para o rápido desenvolvimento da rede. Nos últimos anos, a Internet se consolidou como uma plataforma universal, um lugar comum para grande parte da população mundial e, por isso, a grande maioria das campanhas publicitárias hoje incluem alguma exposição online. Além deste fator, há também a possibilidade de veiculação de publicidade baseada em segmentações de usuários feitas sobre dados de suas interações e de seus perfis de navegação. Nesses modelos, o anunciante investe apenas para atingir o público-alvo de interesse, o que aumenta em muito a efetividade da campanha.

A escolha dos anúncios a serem exibidos é de extrema importância, pois determina o que cada usuário irá ver e, consequentemente, afeta diretamente o desempenho global do sistema. Nesse contexto, é primordial que os administradores do sistema de publicidade sejam capazes de, em certa medida, controlar o comportamento do sistema e propor evoluções para o mesmo. Ora, isto se dá principalmente fazendo alterações na fórmula de ordenação. Por isso, o problema que abordamos foi:

Como fazer alterações na função de ordenação de anúncios baseado em dados reais a fim de melhorar o desempenho do sistema de publicidade computacional?

Como vimos, há um tênue equilíbrio de interesses (possivelmente conflitantes) em questão: o do usuário, e do anunciante e e do publicador. Logo, uma boa resposta para essa pergunta é essencial para que os especialistas de domínio possam fazer alterações nas fórmulas de ordenação e evoluir o sistema de publicidade.

5.1 Contribuições

Nossa primeira contribuição foi definir métricas objetivas para representar os interesses de cada um dos participantes (c.f. 2.3.1) e definir a função objetivo que combina os interesses dos participantes conforme os objetivos do administrador do sistema (c.f. 2.3.2). Feito isso, nossa contribuição principal foi descrever um roteiro de análise de dados que avalia as variáveis que compõe as fórmulas tanto individualmente quanto de maneira combinada. Utilizando teste A/B fomos capazes de analisar o comportamento isolado de cada variável na ordenação de anúncios, ao passo que utilizando

a análise de componentes principais tivemos visão de como elas estão correlacionadas entre si. Não só descrevemos um roteiro de análise mas também comprovamos sua efetividade em esclarecer os especialistas de domínio quanto à evolução do sistema de publicidade. Utilizando o ambiente de publicidade do UOL Cliques, executamos nosso roteiro de análise e apresentamos um diagnóstico de como proceder para melhorar os resultados. As conclusões obtidas foram apresentadas e confirmadas junto aos especialistas de domínio e as ações sugeridas foram tomadas (conforme descrito em 4.3.2). Ademais, seguindo uma boa prátcia das conferências de ciência de dados, disponibilizamos o software das análises e uma amostra de dados de exemplo, para que qualquer um possa reproduzir os resultados obtidos.

Nosso estudo sobre as variáveis BID, CTR, CTX e BTU mostrou que elas possuem impacto direto no desempenho do sistema de anúncios, e que o correto entendimento de seu comportamento é a chave para uma manipulação apropriada das estratégias de ordenação. Quando tornamos clara a participação de cada uma nas métricas de interesse dos participantes é muito mais simples para os especialistas de domínio sugerirem mudanças e priorizarem evoluções do sistema. Vimos também que a baixa correlação entre elas sugere que nenhuma delas deve ser descartada e que usá-las em conjunto amplia o contexto capturado e as chances de recomendar anúncios melhores para o usuário. Vale salientar o fato de stCTR desempenhar tão bem em CTR indica que podemos olhar para o problema de exibição de publicidade do ponto de vista de recomendação, no qual usuários recomendam anúncios para outros que tenham comportamento semelhante.

Em todo este trabalho priorizamos uma abordagem aplicada, visando resolver problemas reais enfrentados no mercado de publicidade computacional. Nossa proposta é especialmente útil para as redes de publicidade computacional que estão na transição entre a 2ª e a 3ª geração de sistemas. Pois precisam manter seus sistemas em evolução enquanto ainda não tenham lidado com os desafios que a evolução automática imponha à arquitetura e à inteligência sistêmica. Fomos além de nosso roteiro de análise e apresentamos os desdobramentos que nossa análise inicial obteve. Assim, além da realização do primeiro experimento que validou nossa proposta de análise, uma última contribuição importante foi a realização de outros dois experimentos em ambiente real. No experimento 2 (c.f. 4.5) verificamos que uma boa combinação das variáveis de ordenação pode trazer benefícios aos interesses dos 3 participantes do sistema simultaneamente. Em particular validamos que a fórmula originalmente usada pelo Google ($stTRAD = BID * CTR$) é um bom ponto de partida para composição de fórmulas de ordenação. No experimento 3 (c.f. 4.6) implementamos a veiculação isolada por grupos de usuários, o que nos permitiu tirar conclusões que levassem em conta o tempo de exposição do usuário a uma determinada fórmula de ordenação.

5.2 Publicações

No processo de concepção deste trabalho publicamos um artigo no SBBD 2014 intitulado *"Principais componentes na ordenação de anúncios: um experimento em ambiente real de publicidade computacional"* [CBF14], que ganhou o prêmio de melhor artigo da conferência. Esse resultado nos rendeu ainda um convite para publicar uma versão extendida deste trabalho no periódico eletrônico JIDM (Journal of Information and Data Management). O artigo foi submetido sob o título *"Principal Component Analysis on Ad Ranking: an experiment on a real online computational advertising system"* e ainda encontra-se sob processo de revisão.

5.3 Trabalhos futuros

Este trabalho foi concebido e desenvolvido de maneira aplicada a problemas reais de publicidade computacional, e do mesmo modo devem seguir seus desdobramentos futuros. Como vimos, este trabalho mostrou-se primariamente relevante para os sistemas de publicidade em transição da 2ª para a 3ª geração. Seguindo nessa mesma linha, o primeiro ponto de melhoria deste trabalho deveria ser realizar um estudo mais completo das variáveis utilizadas nos sistemas de publicidade, e não apenas as 4 disponíveis no UOL Cliques e aqui discutidas (BID, CTR, CTX e BTU). Em outros sistemas, há provavelmente outras variáveis disponíveis para estudo, principalmente no tocante a classificação das preferências dos usuários. Vale mencionar ainda que nossos experimentos contemplaram apenas o escopo descrito em 4.1, e a simples reprodução dos mesmos em um outro escopo, como por exemplo o de anúncios de vídeo, pode trazer fatos novos e interessantes. Uma outra opção seria utilizar outras técnicas para a análise das variáveis, como por exemplo trocar PCA por alguma outra técnica de análise multivariada.

A alimentação constante de um proceso de inteligência é uma característica mais acentuada da 3ª geração dos sistemas de publicidade. Portanto, uma outra linha de pesquisa muito relevante seria automatizar a monitoração de métricas para fazer alterações também automáticas nas fórmulas de ordenação. O trabalho de doutorado de Marcos Broinizi, citado no capítulo 3, seguiu esta linha utilizando programação genética como arcabouço para evolução das fórmulas. Uma outra heurística que pode ser explorada seria a geração de variações de uma fórmula oficial baseada no comportamento local das variáveis (por exemplo uma fórmula diferente para cada site, ou para cada tipo de usuário).

Atualmente, a maior tendência do mercado de publicidade digital é a plataforma de *Real-Time Bidding*. Como vimos, nelas a compra de inventário de espaço publicitário se dá por meio de um leilão em tempo real entre várias redes de anunciantes. O veiculador da campanha pode escolher para cada requisição de cada usuário o quanto está disposto a pagar pela impressão de seus anúncios, considerando tudo o que conhece sobre aquele usuário. Nesse cenário, as possibilidades de otimização das campanhas são extremamente ricas e cabe à rede de publicidade o desenvolvimento e manutenção de uma inteligência refinada para fazer as apostas e assim maximizar o resultado do anunciante. Esse seria um cenário perfeito para a monitoração de métricas de interesse segmentadas por inúmeras variáveis de interesse. Um exemplo disso seria o sistemas aumentar as apostas de impressões para um determinado perfil de usuário ou site, devido a uma alta taxa de conversão.

Por fim, poderíamos citar pontos mais simples de melhoria que não foram possíveis de incluir neste trabalho por questões alheias a nós. Tais como a exibição da fórmula oficial de ordenação do UOL Cliques bem os valores absolutos reais de alguns dados, a segmentação das análises por períodos menores de tempo, a disponibilização da plataforma de experimentação, entre outras.

Apêndice A

Implementações das análises

Este apêndice contém os componentes de *software* criados para construir os gráficos necessários nas análises. Todo o material, inclusive exemplos de execução com os dados das análises aqui apresentadas, estão disponíveis no github em:

https://github.com/acasimiro/mestrado

A.1 Implementação da análise de desempenho

A criação dos gráficos de análise de desempenho da seção 4.2.1 foi feita na linguagem Python utilizando bibliotecas comuns para análises de dados. É possível visualizar o resultado de uma execução em:

https://github.com/acasimiro/mestrado/blob/master/analise_desempenho.ipynb

Exemplo de código para execução:

```
from performance_analysis import run_performance_analysis

csv_file = 'data/experiment1-public.csv'
mapping = {
    'a9': ('stRND', 'white'),
    'b9': ('stCTR', 'orange'),
    'c9': ('stCTX', 'darkblue'),
    'd9': ('stBTU', 'darkgreen'),
    'e9': ('stBID', 'darkred'),
}

run_performance_analysis(csv_file, mapping, alpha=0.333, beta=0.333, gamma
    =0.334, delta=1)
```

Arquivo `performance_analysis.py`

```
from __future__ import division

import datetime as dt
import pandas as pd
import matplotlib.pyplot as plt
```

```
6   import numpy as np
7
8   plt.style.use('ggplot')
9
10  OUTPUT_FOLDER = 'output/'
11
12
13  def run_performance_analysis(filepath, mapping, alpha, beta, gamma, delta):
14      if [v for v in (alpha, beta, gamma, delta) if (v < 0) or (v > 1)]:
15          raise RuntimeError('Parameter outside range [0,1]')
16      if alpha + beta + gamma != 1.0:
17          raise RuntimeError('alpha + beta + gamma != 1.0')
18
19      strategies_mapping = {k:v[0] for k, v in mapping.items()}
20      ctr_performance, ecpm_performance = performance_data(filepath,
            strategies_mapping)
21
22      colors_mapping = dict(mapping.values())
23      performance_chart(ctr_performance, colors_mapping, 'CTR')
24      performance_chart(ecpm_performance, colors_mapping, 'ECPM')
25
26      fobj_chart(ctr_performance, ecpm_performance, colors_mapping, alpha, beta,
            gamma, delta)
27
28
29  def run_raw_analysis(filepath, mapping):
30      strategies_mapping = {k:v[0] for k, v in mapping.items()}
31      impressions, clicks, revenue = raw_data(filepath, strategies_mapping)
32
33      colors_mapping = dict(mapping.values())
34      performance_chart(impressions, colors_mapping, 'IMPRESSIONS')
35      performance_chart(clicks, colors_mapping, 'CLICKS')
36      performance_chart(revenue, colors_mapping, 'REVENUE')
37
38
39  def grouped_data(filepath, strategies_mapping):
40      df = pd.read_csv(filepath)
41
42      df = df[df.STRATEGY.isin(strategies_mapping)]
43      df.STRATEGY = df.STRATEGY.apply(strategies_mapping.get)
44
45      make_date = lambda r: dt.date(r['NUM_YEAR'], r['NUM_MONTH'], r['NUM_DAY'])
46      df['DATE'] = df.apply(make_date, axis=1)
47      df['CTR'] = df['CLICKS']/df['IMPRESSIONS']
48      df['ECPM'] = df['REVENUE']/df['IMPRESSIONS']
49      df.set_index(['DATE'], inplace=True)
50
51      grouped = df.groupby(['STRATEGY'])
52      return grouped
53
54
55  def raw_data(filepath, strategies_mapping):
56      grouped = grouped_data(filepath, strategies_mapping)
```

```
57
58      impressions = {}
59      clicks = {}
60      revenue = {}
61      for strategy, group in grouped:
62          impressions[strategy] = group['IMPRESSIONS']
63          clicks[strategy] = group['CLICKS']
64          revenue[strategy] = group['REVENUE']
65
66      return impressions, clicks, revenue
67
68
69  def performance_data(filepath, strategies_mapping):
70      grouped = grouped_data(filepath, strategies_mapping)
71
72      baseline_ctr = grouped.get_group('stRND').CTR
73      baseline_ecpm = grouped.get_group('stRND').ECPM
74
75      ctr_performance = {}
76      ecpm_performance = {}
77      for strategy, group in grouped:
78          if strategy == 'stRND':  continue
79          ctr_performance[strategy] = (group['CTR'] - baseline_ctr) / (
                baseline_ctr)
80          ecpm_performance[strategy] = (group['ECPM'] - baseline_ecpm) / (
                baseline_ecpm)
81
82      return ctr_performance, ecpm_performance
83
84
85  def performance_chart(performance, colors, label):
86      num_days = len(performance.values()[0])
87      too_long = (num_days > 10)
88
89      x_index = np.arange(num_days)
90
91      fig = plt.figure(figsize=(12, 5.5 if too_long else 4))
92      ax  = fig.add_subplot(111)
93
94      if too_long:
95          x_labels = [str(i) for i in range(1, num_days + 1)]
96          for i, (st, series) in enumerate(performance.items(), 1):
97              ax.plot(x_index, series, color=colors[st], label=st, marker='.')
98      else:
99          bar_width = 1/(len(performance) + 2)
100         x_labels = ['Dia ' + str(i) for i in range(1, num_days + 1)]
101         for i, (st, series) in enumerate(performance.items(), 1):
102             ax.bar(x_index + i*bar_width, series, bar_width, color=colors[st],
                    label=st)
103
104     plt.ylabel(label)
105     ax.set_position([0.1, 0, 0.5, 0.8])
```

```
106         ax.legend(loc='lower center', bbox_to_anchor=(0.5, -0.25), ncol=len(x_labels
                ))
107         if too_long:
108             plt.xlabel('Dia')
109             ax.set_xlim([x_index[0], x_index[-1]])
110             plt.xticks(x_index, x_labels, rotation=90)
111         else:
112             plt.xticks(x_index + 0.5, x_labels)
113         plt.tight_layout()
114         fig.savefig(OUTPUT_FOLDER + label + '_performance.pdf')
115         plt.show()
116 
117 
118     def fobj_chart(ctrs, ecpms, colors, alpha, beta, gamma, delta):
119         strategies = ctrs.keys()
120 
121         num_days = len(ctrs.values()[0])
122         too_long = (num_days > 10)
123 
124         x_index = np.arange(num_days)
125 
126         fig = plt.figure(figsize=(12, 5.5 if too_long else 4))
127         ax = fig.add_subplot(111)
128         for st in ctrs:
129             C = ctrs[st]
130             E = ecpms[st]
131             mu = (C + E)/2
132             sigma2 = ((C-mu)**2 + (E-mu)**2)/2
133             fobjs = (alpha + beta)*C + gamma*E - delta*sigma2
134             ax.plot(x_index, fobjs, color=colors[st], label=st, marker='.' if
                    too_long else 'o')
135 
136         if too_long:
137             plt.xlabel('Dia')
138             x_labels = [str(i) for i in range(1, num_days + 1)]
139             plt.xticks(x_index, x_labels, rotation=90)
140         else:
141             x_labels = ['Dia ' + str(i) for i in range(1, num_days + 1)]
142             plt.xticks(x_index, x_labels)
143 
144         plt.ylabel('Fobj')
145         ax.set_position([0.1, 0, 0.5, 0.8])
146         ax.legend(loc='lower center', bbox_to_anchor=(0.5, -0.25), ncol=len(x_labels
                ))
147         ax.set_xlim([x_index[0]-0.1, x_index[-1]+0.1])
148 
149         plt.tight_layout()
150         fig.savefig(OUTPUT_FOLDER + 'Fobj.pdf')
151         plt.show()
```

Exemplo de arquivo CSV de entrada (`experiment1-public.csv`)

```
1  STRATEGY,NUM_YEAR,NUM_MONTH,NUM_DAY,CLICKS,IMPRESSIONS,REVENUE
2  a9,1900,1,1,11702,33871653,2819
```

```
a9,1900,1,2,12670,40771163,3076
a9,1900,1,3,13513,34068057,2887
a9,1900,1,4,11513,37603119,2723
a9,1900,1,5,12524,30228063,2791
a9,1900,1,6,11772,29600088,2534
a9,1900,1,7,10799,34910147,2548
b9,1900,1,1,15433,33787709,3321
b9,1900,1,2,15644,40599288,3582
```

A.2 Implementação da análise de PCA

A criação dos gráficos de análise de PCA da seção 4.2.2 foi feita na linguagem R utilizando bibliotecas comuns para análises de dados. É possível visualizar o resultado de uma execução em:

https://github.com/acasimiro/mestrado/blob/master/analise_pca.ipynb

Exemplo de código para execução:

```
source('pca_analysis.R')
csv_file = 'data/variables-public.csv'
variables = c('BID', 'CTR', 'CTX', 'BTU')

run_pca_analysis(csv_file, variables, components=c(1, 2))
run_pca_analysis(csv_file, variables, components=c(3, 4))
```

Arquivo `loading_plot.R`

```
# Original author: Vincent Q. Vu.
# Modified from: https://github.com/vqv/ggbiplot

loading_plot <- function(pcobj, choices = 1:2, scale = 1, pc.biplot = TRUE,
                         obs.scale = 1 - scale, var.scale = scale,
                         groups = NULL, ellipse = FALSE, ellipse.prob = 0.68,
                         labels = NULL, labels.size = 3, alpha = 1,
                         var.axes = TRUE,
                         circle = FALSE, circle.prob = 0.69,
                         varname.size = 3, varname.adjust = 1.5,
                         varname.abbrev = FALSE, ...)
{
  library(ggplot2)
  library(plyr)
  library(scales)
  library(grid)

  stopifnot(length(choices) == 2)

 if(inherits(pcobj, 'prcomp')){
    nobs.factor <- sqrt(nrow(pcobj$x) - 1)
    d <- pcobj$sdev
    u <- sweep(pcobj$x, 2, 1 / (d * nobs.factor), FUN = '*')
    v <- pcobj$rotation
  } else if(inherits(pcobj, 'princomp')) {
    nobs.factor <- sqrt(pcobj$n.obs)
```

```
27        d <- pcobj$sdev
28        u <- sweep(pcobj$scores, 2, 1 / (d * nobs.factor), FUN = '*')
29        v <- pcobj$loadings
30      } else if(inherits(pcobj, 'PCA')) {
31        nobs.factor <- sqrt(nrow(pcobj$call$X))
32        d <- unlist(sqrt(pcobj$eig)[1])
33        u <- sweep(pcobj$ind$coord, 2, 1 / (d * nobs.factor), FUN = '*')
34        v <- sweep(pcobj$var$coord,2,sqrt(pcobj$eig[1:ncol(pcobj$var$coord),1]),FUN=
              "/")
35      } else if(inherits(pcobj, "lda")) {
36          nobs.factor <- sqrt(pcobj$N)
37          d <- pcobj$svd
38          u <- predict(pcobj)$x/nobs.factor
39          v <- pcobj$scaling
40          d.total <- sum(d^2)
41      } else {
42        stop('Expected a object of class prcomp, princomp, PCA, or lda')
43      }
44
45      choices <- pmin(choices, ncol(u))
46      df.u <- as.data.frame(sweep(u[,choices], 2, d[choices]^obs.scale, FUN='*'))
47
48      v <- sweep(v, 2, d^var.scale, FUN='*')
49      df.v <- as.data.frame(v[, choices])
50
51      names(df.u) <- c('xvar', 'yvar')
52      names(df.v) <- names(df.u)
53
54      if(pc.biplot) {
55        df.u <- df.u * nobs.factor
56      }
57
58      r <- sqrt(qchisq(circle.prob, df = 2)) * prod(colMeans(df.u^2))^(1/4)
59
60      v.scale <- rowSums(v^2)
61      df.v <- r * df.v / sqrt(max(v.scale))
62
63        u.axis.labs <- paste('PC', choices, sep='')
64
65      if(!is.null(labels)) {
66        df.u$labels <- labels
67      }
68
69      if(!is.null(groups)) {
70        df.u$groups <- groups
71      }
72
73      if(varname.abbrev) {
74        df.v$varname <- abbreviate(rownames(v))
75      } else {
76        df.v$varname <- rownames(v)
77      }
78
```

```
79      df.v$angle <- with(df.v, (180/pi) * atan(yvar / xvar))
80      df.v$hjust = with(df.v, (1 - varname.adjust * sign(xvar)) / 2)
81
82      g <- ggplot(data = df.u, aes(x = xvar, y = yvar)) +
83              xlab(u.axis.labs[1]) + ylab(u.axis.labs[2]) + coord_equal()
84
85      if(var.axes) {
86        if(circle)
87        {
88          theta <- c(seq(-pi, pi, length = 50), seq(pi, -pi, length = 50))
89          circle <- data.frame(xvar = r * cos(theta), yvar = r * sin(theta))
90          g <- g + geom_path(data = circle, color = muted('white'),
91                             size = 1/2, alpha = 1/3)
92        }
93
94        g <- g +
95          geom_segment(data = df.v,
96                       aes(x = 0, y = 0, xend = xvar, yend = yvar),
97                       arrow = arrow(length = unit(1/2, 'picas')),
98                       color = muted('red'))
99      }
100
101     if(!is.null(df.u$labels)) {
102       if(!is.null(df.u$groups)) {
103         g <- g + geom_text(aes(label = labels, color = groups),
104                            size = labels.size)
105       } else {
106         g <- g + geom_text(aes(label = labels), size = labels.size)
107       }
108     }
109
110     if(!is.null(df.u$groups) && ellipse) {
111       theta <- c(seq(-pi, pi, length = 50), seq(pi, -pi, length = 50))
112       circle <- cbind(cos(theta), sin(theta))
113
114       ell <- ddply(df.u, 'groups', function(x) {
115         if(nrow(x) < 2) {
116           return(NULL)
117         } else if(nrow(x) == 2) {
118           sigma <- var(cbind(x$xvar, x$yvar))
119         } else {
120           sigma <- diag(c(var(x$xvar), var(x$yvar)))
121         }
122         mu <- c(mean(x$xvar), mean(x$yvar))
123         ed <- sqrt(qchisq(ellipse.prob, df = 2))
124         data.frame(sweep(circle %*% chol(sigma) * ed, 2, mu, FUN = '+'),
125                    groups = x$groups[1])
126       })
127       names(ell)[1:2] <- c('xvar', 'yvar')
128       g <- g + geom_path(data = ell, aes(color = groups, group = groups))
129     }
130
131     if(var.axes) {
```

```
      g <- g +
      geom_text(data = df.v,
                aes(label = varname, x = xvar, y = yvar,
                    angle = angle, hjust = hjust),
                color = 'darkred', size = varname.size)
    }

    return(g)
  }
```

Exemplo de arquivo CSV de entrada (`variables-public.csv`)

```
st, fmt, uuid, ad, cr, fl, BID, mibid, avbid, mxbid, nmbid, CTR, mictr, avctr, mxctr, nmctr, CTX, mictx, avctx, mxctx, nmctx, qlt, miqlt, avqlt, mxqlt, nmqlt, CTU, mictu, avctu, mxctu, nmctu, BTU, mibtu, avbtu, mxbtu, nmbtu, hour, comb, order, adnum, cpc, click
b9, ci_n_ads.js, dd17cd73a5c748e292e5e2156bf2c48e, 1010666, 1324357, 3173001, 0.60, 0.05, 0.43, 2.50, 0.06939, 5655, 259, 2396, 7723, 0.4366, 1.00, 1.00, 1.00, 1.00, 1000, 1.00, 1.00, 1.00, 1.00, 1000, 0.58, 0.26, 0.74, 3.79, −0.04533, 0.00, 0.00, 0.00, 0.00, 0, 0, 729014:980277:1010812:1010666:985428:1010837:1008306:1010873:1010130:, 4, 9, 0.00, 0
auctionexpr, ci_n_ads.js, 0b4808bc4e934c6e9d87126c00fec584, 1010557, 1324109, 3172925, 1.00, 0.05, 0.56, 1.10, 0.4190, 1663, 615, 2633, 11508, −0.08905, 0.53, 0.39, 0.51, 0.95, 0.03571, 4.40, 3.31, 4.60, 7.04, −0.05362, 0.00, 0.00, 0.00, 0.00, 0, 0.00, 0.00, 0.00, 0.00, 0, 0, 1010520:1009614:1010557, 3, 3, 0.00, 0
auctionexpr, ci_n_ads.js, 3cdfdfd6c1ba4647821b287c3537d1f6, 1008306, 1320140, 3170962, 0.80, 0.05, 0.44, 0.80, 0.4800, 4118, 414, 2435, 6574, 0.2732, 2.51, 2.20, 3.74, 5.78, −0.3436, 6.44, 4.29, 6.26, 8.38, 0.04401, 0.00, 0.00, 0.00, 0.00, 0, 0.00, 0.00, 0.00, 0.00, 0, 1, 1008306:1010127, 1, 2, 0.00, 0
b9, ci_n_ads.js, 88dc05211f2f4faf9a2d986d8bfd2842, 980277, 1298145, 3115793, 0.05, 0.05, 0.41, 0.90, −0.4235, 5692, 442, 1958, 5692, 0.7112, 1.00, 1.00, 1.00, 1.00, 1000, 1.00, 1.00, 1.00, 1.00, 1000, 1.81, 1.81, 2.88, 4.17, −0.4534, 0.00, 0.00, 0.00, 0.00, 0, 2, 980277:1008306:, 1, 2, 0.00, 0
auctionexpr, ci_n_ads.js, f38069b211d545c09aa8a96ee9744142, 1008306, 1320139, 3170962, 0.80, 0.05, 0.42, 0.80, 0.5067, 4113, 405, 1935, 6573, 0.3531, 3.68, 1.34, 3.29, 4.14, 0.1393, 8.16, 4.82, 6.44, 8.16, 0.5150, 0.00, 0.00, 0.00, 0.00, 0, 0.00, 0.00, 0.00, 0.00, 0, 5, 1008306:1010105, 1, 2, 0.8, 1
auctionexpr, ci_n_ads.js, a93e0911fd744456b955c0baab7d0784, 1008306, 1320140, 3170960, 0.80, 0.05, 0.55, 1.00, 0.2632, 4421, 570, 3016, 7722, 0.1964, 0.89, 0.89, 1.08, 1.99, −0.1727, 5.79, 3.79, 5.52, 8.12, 0.06236, 0.00, 0.00, 0.00, 0.00, 0, 0.00, 0.00, 0.00, 0.00, 0, 7, 1008306:1008526:1007922, 1, 3, 0.00, 0
auctionexpr, ci_n_ads.js, 37a3b66c7f164f238681254923338b54, 1008306, 1320140, 3170962, 0.80, 0.05, 0.32, 0.80, 0.6400, 4109, 227, 1745, 4112, 0.6085, 3.25, 2.42, 2.79, 3.25, 0.5542, 10.00, 5.68, 7.20, 10.00, 0.6481, 0.00, 0.00, 0.00, 0.00, 0, 0.00, 0.00, 0.00, 0.00, 0, 7, 1008306:1010127, 1, 2, 0.00, 0
c9, ci_n_ads.js, 6333f39154b645d0862e661bf99c1b16, 1009764, 1322795, 3172127, 0.39, 0.05, 0.38, 0.80, 0.01333, 2267, 434, 2141, 5679, 0.02402, 1.00, 1.00, 1.00, 1.00, 1000, 1.00, 1.00, 1.00, 1.00, 1000, 4.00, 2.80, 4.52, 5.82, −0.1722, 0.00, 0.00, 0.00, 0.00, 0, 7, 1009764:1008525, 1, 2, 0.39, 1
auctionexpr, ci_n_ads.js, 6cb0ce16bd43458db390ce407e95b781, 1010105, 1323512, 3172448, 0.80, 0.05, 0.42, 0.80, 0.5067, 1562, 403, 1993, 5677, −0.08172,
```

```
6.20, 2.20, 5.76, 6.77, 0.09628, 6.88, 5.25, 6.93, 8.67, −0.01462, 0.00,
0.00, 0.00, 0.00, 0, 0.00, 0.00, 0.00, 0.00, 0, 7, 1008306:1010105, 2, 2,
0.72, 1
```

Referências Bibliográficas

[AC14] Azin Ashkan e Charles L. A. Clarke. Location- and query-aware modeling of browsing and click behavior in sponsored search. *ACM Trans. Intell. Syst. Technol.*, 5(4):59:1–59:31, Dezembro 2014. 32

[AG12] Deepak Agarwal e Maxim Gurevich. Fast top-k retrieval for model based recommendation. Em *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*, WSDM '12, páginas 483–492, New York, NY, USA, 2012. ACM. 32, 33

[AGH+09] Deepak Agarwal, Evgeniy Gabrilovich, Robert Hall, Vanja Josifovski e Rajiv Khanna. Translating relevance scores to probabilities for contextual advertising. Em David Wai-Lok Cheung, Il-Yeol Song, Wesley W. Chu, Xiaohua Hu e Jimmy J. Lin, editors, *CIKM*, páginas 1899–1902. ACM, 2009. 17

[ALT+14] Deepak Agarwal, Bo Long, Jonathan Traupman, Doris Xin e Liang Zhang. Laser: A scalable response prediction platform for online advertising. Em *Proceedings of the 7th ACM International Conference on Web Search and Data Mining*, WSDM '14, páginas 173–182, New York, NY, USA, 2014. ACM. 34

[Ash08] Tim Ash. *Landing page optimization : the definitive guide to testing and tuning for conversions*. Serious skills. Indianapolis, Ind. Sybex/Wiley, 2008. Index. 27

[AZZ+12] Javad Azimi, Ruofei Zhang, Yang Zhou, Vidhya Navalpakkam, Jianchang Mao e Xiaoli Fern. Visual appearance of display ads and its effect on click through rate. Em *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, páginas 495–504, New York, NY, USA, 2012. ACM. 13, 17, 32

[BCF+08a] Andrei Broder, Massimiliano Ciaramita, Marcus Fontoura, Evgeniy Gabrilovich, Vanja Josifovski, Donald Metzler, Vanessa Murdock e Vassilis Plachouras. To swing or not to swing: Learning when (not) to advertise. Em *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, CIKM '08, páginas 1003–1012, New York, NY, USA, 2008. ACM. 17, 32

[BCF+08b] Andrei Z. Broder, Peter Ciccolo, Marcus Fontoura, Evgeniy Gabrilovich, Vanja Josifovski e Lance Riedel. Search advertising using web relevance feedback. Em *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, CIKM '08, páginas 1013–1022, New York, NY, USA, 2008. ACM. 32

[BCK+14] Paul Barford, Igor Canadi, Darja Krushevskaja, Qiang Ma e S. Muthukrishnan. Adscape: Harvesting and analyzing online display ads. Em *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14, páginas 597–608, New York, NY, USA, 2014. ACM. 34

[BF15] Marcos Broinizi e João Eduardo Ferreira. *Ordenação evolutiva de anúncios em publicidade computacional*. Tese de Doutorado, IME-USP, fevereiro 2015. 25, 35

[BFJR07] Andrei Broder, Marcus Fontoura, Vanja Josifovski e Lance Riedel. A semantic approach to contextual advertising. Em *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '07, páginas 559–566, New York, NY, USA, 2007. ACM. 5, 32

[BGJM10] Michael Bendersky, Evgeniy Gabrilovich, Vanja Josifovski e Donald Metzler. The anatomy of an ad: structured indexing and retrieval for sponsored search. Em *Proceedings of the 19th international conference on World wide web*, WWW '10, páginas 101–110, New York, NY, USA, 2010. ACM. 17, 21

[BHR10] Abraham Bagherjeiran, Andrew O. Hatch e Adwait Ratnaparkhi. Ranking for the conversion funnel. Em *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '10, páginas 146–153, New York, NY, USA, 2010. ACM. 15, 17, 33

[BJ11] Andrei Broder e Vanja Josifovski. *Introduction to Computational Advertising*. http://www.stanford.edu/class/msande239/lectures-2011/Lecture 01 Intro.pdf. Presented in Stanford University, Autumn 2011. 2

[BMF14] Marcos Broinizi, Danilo Mutti e João Eduardo Ferreira. Application configuration repository for adaptive service-based systems: Overcoming challenges in an evolutionary online advertising environment. Em *Proceedings of the 21th International Conference on Web Services*, ICWS '14, 2014. 34, 41

[CAJ08] Deepayan Chakrabarti, Deepak Agarwal e Vanja Josifovski. Contextual advertising by combining relevance with click feedback. Em *Proceedings of the 17th International Conference on World Wide Web*, WWW '08, páginas 417–426, New York, NY, USA, 2008. ACM. 33

[CBF14] André Casimiro, Marcos Broinizi e João Eduardo Ferreira. Principais componentes na ordenação de anúncios: um experimento em ambiente real de publicidade computacional. Em *Proceedings of the 29th Brazilian Symposium on Databases*, SBBD '14, 2014. 56

[CDC87] C.A. Callioli, H.H. Domingues e R.C.F. Costa. *Algebra linear e aplicações*. Atual, 1987. 29

[CZA+12] Haibin Cheng, Roelof van Zwol, Javad Azimi, Eren Manavoglu, Ruofei Zhang, Yang Zhou e Vidhya Navalpakkam. Multimedia features for click prediction of new ads in display advertising. Em *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '12, páginas 777–785, New York, NY, USA, 2012. ACM. 13, 32

[DCR+14] Brian Dalessandro, Daizhuo Chen, Troy Raeder, Claudia Perlich, Melinda Han Williams e Foster Provost. Scalable hands-free transfer learning for online advertising. Em *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, páginas 1573–1582, New York, NY, USA, 2014. ACM. 34

[Ema14] Emarketer.com. Global ad spending growth to double this year, Julho 2014. http://www.emarketer.com/Article/Global-Ad-Spending-Growth-Double-This-Year/1010997. 1

[EQvDC08] B. Eisenberg, J. Quarto-vonTivadar, L.T. Davis e B. Crosby. *Always Be Testing: The Complete Guide to Google Website Optimizer*. Serious skills. Wiley, 2008. 27

[FKLT12] Ariel Fuxman, Anitha Kannan, Zhenhui Li e Panayiotis Tsaparas. Enabling direct interest-aware audience selection. Em *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, páginas 575–584, New York, NY, USA, 2012. ACM. 32, 44

[GMKP11] Hector Garcia-Molina, Georgia Koutrika e Aditya Parameswaran. Information seeking: Convergence of search, recommendations, and advertising. *Commun. ACM*, 54(11):121–130, Novembro 2011. 44

[Goo02] Google. Google introduces new pricing for popular self-service online advertising program, 2002. http://googlepress.blogspot.com.br/2002/02/google-introduces-new-pricing-for.html, Fevereiro. 10

[Goo03] Google. Google builds world's largest advertising and search monetization program, 2003. 11

[IAB13] IAB. Internet advertising revenue report 2012 full year results, Abril 2013. http://www.iab.net/media/file/IAB_Internet_Advertising_Revenue_Report_FY_2012_rev.pdf, Abril. 11

[IAB14] IAB. Internet advertising revenue report 2013 full year results. Relatório técnico, Abril 2014. http://www.iab.net/media/file/IAB_Internet_Advertising_Revenue_Report_FY_2013.pdf, Abril. 1

[Jol02] I.T. Jolliffe. *Principal Component Analysis*. Springer Series in Statistics. Springer, 2002. 28

[KLSH09] Ron Kohavi, Roger Longbotham, Dan Sommerfield e RandalM. Henne. Controlled experiments on the web: survey and practical guide. *Data Mining and Knowledge Discovery*, 18(1):140–181, 2009. 27

[KSSS15] Komal Kapoor, Karthik Subbian, Jaideep Srivastava e Paul Schrater. Just in time recommendations: Modeling the dynamics of boredom in activity streams. Em *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, WSDM '15, páginas 233–242, New York, NY, USA, 2015. ACM. 33

[LCG+06] Anísio Lacerda, Marco Cristo, Marcos André Gonçalves, Weiguo Fan, Nivio Ziviani e Berthier Ribeiro-Neto. Learning to advertise. Em *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '06, páginas 549–556, New York, NY, USA, 2006. ACM. 34

[LODL12] Kuang-chih Lee, Burkay Orten, Ali Dasdan e Wentong Li. Estimating conversion rate in display advertising from past erformance data. Em *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '12, páginas 768–776, New York, NY, USA, 2012. ACM. 35

[MCP07] Vanessa Murdock, Massimiliano Ciaramita e Vassilis Plachouras. A noisy-channel approach to contextual advertising. Em *Proceedings of the 1st International Workshop on Data Mining and Audience Intelligence for Advertising*, ADKDD '07, páginas 21–27, New York, NY, USA, 2007. ACM. 21, 33

[MRS08] Christopher D. Manning, Prabhakar Raghavan e Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA, 2008. 21

[Obe13] Ankit Oberoi. The history of online advertising, 2013. http://www.adpushup.com/blog/the-history-of-online-advertising, Julho. 10

[RH12] Ana Radovanovic e William D. Heavlin. Risk-aware revenue maximization in display advertising. Em *Proceedings of the 21st International Conference on World Wide Web*, WWW '12, páginas 91–100, New York, NY, USA, 2012. ACM. 15

[RNCGSdM05] Berthier Ribeiro-Neto, Marco Cristo, Paulo B. Golgher e Edleno Silva de Moura. Impedance coupling in content-targeted advertising. Em *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '05, páginas 496–503, New York, NY, USA, 2005. ACM. 21, 32, 33

[SK11] JamesG. Shanahan e Goutham Kurra. Digital advertising: An information scientist's perspective. Em Massimo Melucci e Ricardo Baeza-Yates, editors, *Advanced Topics in Information Retrieval*, volume 33 of *The Information Retrieval Series*, páginas 209–237. Springer Berlin Heidelberg, 2011. 1, 2

[SK13] Dan Siroker e Pete Koomen. *A/B Testing: The Most Powerful Way to Turn Clicks Into Customers*. Wiley Publishing, 1st edição, 2013. 27

[SLCC14] Fanhua Shang, Yuanyuan Liu, James Cheng e Hong Cheng. Robust principal component analysis with missing data. Em *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, CIKM '14, páginas 1149–1158, New York, NY, USA, 2014. ACM. 29

[SLY12] Konstantin Salomatin, Tie-Yan Liu e Yiming Yang. A unified optimization framework for auction and guaranteed delivery in online advertising. Em *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, páginas 2005–2009, New York, NY, USA, 2012. ACM. 15

[SM12a] Maad Shatnawi e Nader Mohamed. Statistical techniques for online personalized advertising: a survey. Em *Proceedings of the 27th Annual ACM Symposium on Applied Computing*, SAC '12, páginas 680–687, New York, NY, USA, 2012. ACM. 9, 34

[SM12b] Maad Shatnawi e Nader Mohamed. Statistical techniques for online personalized advertising: A survey. Em *Proceedings of the 27th Annual ACM Symposium on Applied Computing*, SAC '12, páginas 680–687, New York, NY, USA, 2012. ACM. 34

[SXY13] Qian Sun, Shuo Xiang e Jieping Ye. Robust principal component analysis via capped norms. Em *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, páginas 311–319, New York, NY, USA, 2013. ACM. 29

[TLY+11] Jian Tang, Ning Liu, Jun Yan, Yelong Shen, Shaodan Guo, Bin Gao, Shuicheng Yan e Ming Zhang. Learning to rank audience for behavioral targeting in display ads. Em *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, páginas 605–610, New York, NY, USA, 2011. ACM. 5, 24, 44

[TPGJ11] Sarah K. Tyler, Sandeep Pandey, Evgeniy Gabrilovich e Vanja Josifovski. Retrieval models for audience selection in display advertising. Em *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, páginas 593–598, New York, NY, USA, 2011. ACM. 5, 32

[VdMH08] Laurens Van der Maaten e Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(2579-2605):85, 2008. 29

[WLF+09] Haofen Wang, Yan Liang, Linyun Fu, Gui-Rong Xue e Yong Yu. Efficient query expansion for advertisement search. Em *Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, páginas 51–58, New York, NY, USA, 2009. ACM. 32

[WYZ11] Lei Wang, Mingjiang Ye e Yu Zou. A language model approach to capture commercial intent and information relevance for sponsored search. Em *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, páginas 599–604, New York, NY, USA, 2011. ACM. 5

[YGC06] Wen-tau Yih, Joshua Goodman e Vitor R. Carvalho. Finding advertising keywords on web pages. Em *Proceedings of the 15th International Conference on World Wide Web*, WWW '06, páginas 213–222, New York, NY, USA, 2006. ACM. 32

[ZYW14] Weinan Zhang, Shuai Yuan e Jun Wang. Optimal real-time bidding for display advertising. Em *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, páginas 1077–1086, New York, NY, USA, 2014. ACM. 35