Sugestão de palavras chave para campanhas em motores de busca em arranque

AdWors, Palavras Chave, Regressão Logística, Análise de Campanhas

Keyword suggestion for bootstrapping search engine campaigns

AdWords, Keywords, Logistic Regression, Campaing Analysis

João Albuquerque e Rui Campos (AdClick), Ricardo Morla e Gabriel David (INESC TEC e FEUP)

Resumo

As campanhas de publicidade online sobre pesquisas em motores de busca têm um grande peso nas receitas das empresas de marketing digital. Cada campanha é definida por um conjunto de palavras chave que são utilizadas pelo motor de busca para escolher a melhor publicidade candidata a apresentar juntamente com os resultados da pesquisa. Um problema com que se deparam estas campanhas é a da sugestão das melhores palavras chave para uma determinada campanha. Neste artigo é apresentado um modelo para a análise de relações entre variáveis de campanhas de publicidade, bem como um algoritmo para produzir sugestões de palavras chave. Uma vez que este algoritmo se baseia no histórico da campanha, não é aplicável diretamente a campanhas em arranque que não têm histórico. Este artigo apresenta também uma abordagem baseada no histórico de campanhas semelhantes e que permite ultrapassar esta dificuldade. Finalmente é apresentada uma caracterização estatística do histórico de campanhas AdWords reais e da relação entre estas campanhas.

Abstract

Search engine advertisement campaigns have a huge weight on the income of digital marketing companies. Each campaign is defined by a set of keywords that are used by the search engine to choose the best candidate ad to show together with search results. A problem these campaigns face is how to suggest the best keywords for a given campaign. In this paper we present a model for analyzing the relations between the ad campaign variables, as well as an algorithm to suggest keywords. As this algorithm is based on the history of the campaign, it is not directly applicable to bootstrapping campaigns that do not have history. This paper presents an approach based on the history of similar campaigns that can be used to overcome this bootstrapping problem. Finally we present a statistical characterization of the history of real AdWord campaigns and of the relation between these campaigns.

1. Introdução

Um termo de pesquisa é a expressão que o utilizador introduz num do motor de busca para obter os resultados que procura. A publicidade online em motores de busca tem como um dos seus objetivos a conversão, isto é, o registo de dados pessoais no site da empresa de publicidade ou do vendedor, a compra online, ou uma outra métrica definida caso a caso com o cliente da empresa de publicidade. A taxa de conversão é o rácio entre o número de conversões e o número de cliques sobre áreas de impressão pagas (ou não) no motor de pesquisa, e é normalmente utilizada para validar o sucesso ou insucesso da publicidade. Certas palavras quando incluídas nos termos de pesquisa dos motores de busca parecem influenciar a taxa de conversão de publicidade nestes motores. Uma análise mais profunda deste assunto poderia contribuir para reduzir os custos e aumentar a performance das campanhas de Adwords, ou seja do conjunto de produtos a anunciar do mesmo vendedor, sendo muitas vezes variações do mesmo produto. Uma campanha é constituída por um conjunto de grupos, cada um identificado por uma *landing page* e um conjunto de palavras chave. Um grupo é o elemento base sobre o qual se pode raciocinar sobre a relação entre palavras chave, termos de pesquisa, e conversões. Uma palavra chave (*keyword*) é a palavra ou expressão que o gestor de campanha introduz no AdWord para ajudar a definir, através de um mecanismo de leilão automático, se e onde a URL para a *landing page* do seu grupo será impressa para um determinado termo de pesquisa.

Pretende-se implementar um processo automático que auxilie os gestores de campanha a maximizar a performance das campanhas do Google Adwords com recurso ao relatório de termos de pesquisa do Adwords. Este relatório permite visualizar todos os termos de pesquisa para os quais foram exibidos anúncios e fornece valores de algumas medidas, nomeadamente número de cliques e número de conversões associados a cada termo. Analisando periodicamente este relatório é possível otimizar as campanhas adicionando como *keywords* termos de pesquisa que produzem bons resultados e excluindo os termos que não geram conversões.

No entanto, a análise dos relatórios periódicos não nos permite identificar a influência individual das palavras presentes em cada uma das expressões. Por exemplo, em campanhas relacionadas com crédito, a presença da palavra "ganhar" num termo de pesquisa resulta habitualmente em custos elevados e taxas de conversão baixas. Isto no entanto pode não ser fácil de detetar. Individualmente um termo pode não ter grande custo mas o conjunto dos termos que contêm essa palavra têm um custo significativo. Pretende-se então desenvolver um processo que permita identificar as palavras mais influentes e otimizar as campanhas com base nesses resultados.

Inicia-se este processo identificando as palavras mais frequentes nos termos de pesquisa. Para tal recorre-se a uma ferramenta de análise de texto que conta o número de vezes que determinada palavra apareceu na totalidade dos termos de pesquisa. Para cada uma destas palavras é criada uma variável binária onde se atribui o valor 1 se o termo de pesquisa contem a palavra em análise e o valor 0 caso contrário. Assim sendo, cada um dos termos de pesquisa é codificado com base nestas variáveis binárias.

Um termo de pesquisa com um elevado volume de cliques é sempre mais influente que um com um número de cliques reduzido. Para que este tipo de ponderação esteja presente na análise cada um dos termos de pesquisa é replicado tantas vezes como foi clicado. Desta forma, se um termo obteve 5 cliques e 2 conversões irá ser considerado 5 vezes, 2 com conversão e 3 sem conversão.

Com estes dados é possível fazer uma análise detalhada da influência de cada uma das palavras e de combinações de palavras na conversão. Começa-se por calcular taxas de conversão médias dos termos com e sem cada uma das palavras em análise. Esta pode ser uma forma simples e rápida de identificar palavras influentes mas não tem em consideração a influência das restantes palavras presentes no termo. Para obter resultados mais fieis é necessário recorrer a outros métodos, tais como a modelação estatística. Este artigo apresenta uma abordagem baseada em regressão logística e no histórico das campanhas que se propõe a identificar o efeito positivo ou negativo de uma palavra chave na conversão. Este artigo apresenta também uma abordagem baseada na distância das campanhas para identificar campanhas semelhantes e assim permitir a utilização da aprendizagem estatística em campanhas em arranque.

O resto do artigo desenvolve-se da seguinte forma. São primeiro apresentados casos de utilização das duas abordagens do artigo. De seguinte apresenta-se o modelo de regressão logística e a sua aplicação ao problema de identificação do efeito positivo ou negativo das palavras-chave. Finalmente é apresentada a métrica de distância entre campanhas e são mostradas visualizações de grafos de semelhança entre campanhas e grupos de campanhas para dados de campanhas reais AdWords.

2. Casos de utilização

Na figura 1 está representado o diagrama de utilização dos módulos de análise. O gestor de campanha cria uma nova campanha, que pode posteriormente editar, atualizar, ou eliminar. A criação da campanha é um passo simples que se segue à criação da campanha no AdWords e que requer apenas a importação de um ficheiro fornecido pelo AdWords. Para essa campanha o gestor pretende obter recomendações adicionais às que consegue obter através do AdWords. Para tal utiliza os

módulos de análise de peso das palavras e de análise por comparação. Com um conjunto novo de recomendações, o gestor de campanha pode retornar ao Adwords e inserir as novas Keywords que achar convenientes sugeridas pela ferramenta de análise.

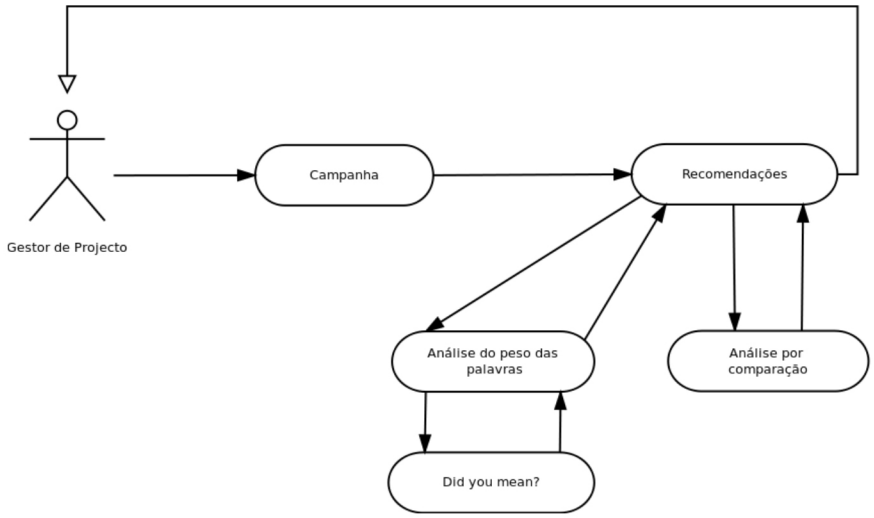

Figura 1. Diagrama de utilização dos módulos de análise.

A figura 2 representa os objetos de alto nível que a ferramenta de análise considera: campanhas, categorias, e sub-categorias.

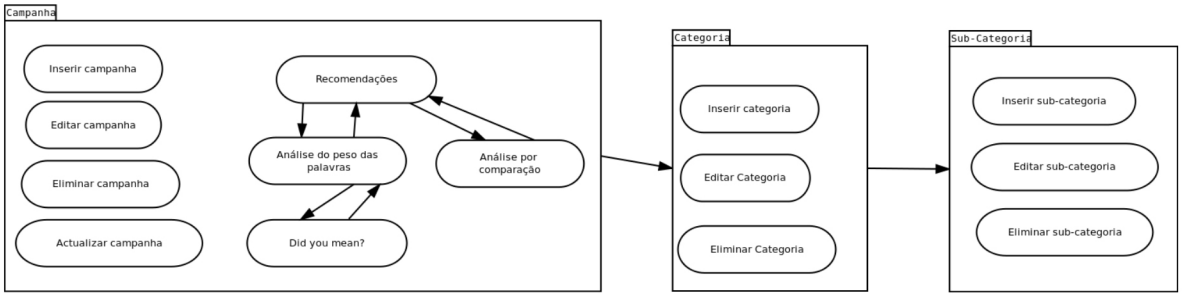

Figura 2. Objetos de alto nível da ferramenta de análise.

Criou-se um sistema onde é possível inserir campanhas e efectuar-se as operações de inserção, atualização, e de esconder/activar a campanha. Estes são armazenados numa base de dados de forma incremental. Criou-se também um sistema de categorização para relacionar as várias campanhas.

É possível efetuar dois tipos de análise:

Análise do peso das palavras
Aqui usamos os "Search Terms" que são frases e verificamos o peso de cada palavra em todas as frases. Manipulamos os dados para poderem entrar como input na regressão logística e retornamos Odds Ratio (que são os indicadores de efeito positivo ou negativo referidos na secção 3). As palavras com um volume baixo de conversão vão ser definidas como palavras negativas. Utilizamos as restantes palavras para tentarmos encontrar frases com uma boa taxa de conversão. Devolvemos o resultado ao gestor de campanha.

Análise por comparação
Efetua-se uma análise de comparação entre a campanha e o histórico das campanhas. Vemos agora as campanhas que são próximas da nossa nova campanha e vamos buscar as palavras negativas assim como as *keywords* que tem o melhor nível de conversão. Devolvemos o resultado ao gestor de campanha.

3. Modelo de regressão logística

Para cada grupo de uma campanha h aplicam-se as seguintes variáveis:

- n termos de pesquisa, com $i = 1, \dots, n$
- m palavras distintas que compõe todos os termos de pesquisa ou universo de palavras, com $j = 1, \dots, m$
- l Keywords, com $k = 1, \dots, l$
- Observação z_{kj}, onde $z_{kj} = 1$ se a Keyword k contém a palavra j. Podemos escrever a matriz Z com dimensões l por m.
- Observações binárias x_{ij}, onde $x_{ij} = 1$ caso o termo de pesquisa i contenha a palavra j, e $x_{ij} = 0$ caso não contenha. Podemos escrever x_{ij} como a matriz X com dimensões n por m.
- Observação binária y_i, onde $y_i = 1$ caso o termo de pesquisa $i = 1, \dots, n$ tenha obtido uma conversão e $y_i = 0$ caso não tenha obtido conversão. Podemos escrever y_i como o vetor Y com dimensão n.
- Variável binária $X = (x_1, \dots, x_m)$ que representa a distribuição de palavras distintas por termos de pesquisa. X é a variável independente.
- Variável binária y que representa a conversão ou não que resulta de um termo de pesquisa. Y é a variável dependente que queremos prever. Temos uma instancia $X_i = (x_{i1}, \dots x_{im})$ e y_i para cada termo de pesquisa i.

Consideramos a existência de c grupos de campanhas. Quando é necessário distinguir entre grupos de campanhas aplicamos o índice $h = 1..c$ a cada variável. Por exemplo utilizamos n^h para representar o número de termos de pesquisa da campanha h.

É possível ajustar estes dados a um modelo de regressão logística onde Y é a variável de conversão de interesse (dependente) e $X = (x_1, x_2, \dots, x_k)$ são os preditores de palavras nos termos de pesquisa (variáveis independentes).

Uma regressão linear pode ser expressa como $Y = \boldsymbol{\beta} \cdot \boldsymbol{X} + \varepsilon$, onde Y é a variável dependente, $\boldsymbol{X}$ é um vetor com variáveis independentes, $\boldsymbol{\beta}$ é um vetor com os coeficientes da regressão que determinam o efeito de cada componente de $\boldsymbol{X}$ em Y e que inclui o valor da interseção (i.e. o valor de Y quando as variáveis independentes são zero), e ε um termo de erro. Uma regressão logística [*Hosmer00*] pode ser vista como uma regressão linear em que os valores de Y estão sujeitos à função logística $1/(e^{-Y} + 1)$, que varia entre 0 e 1 como uma probabilidade.

Fornecendo a matriz $X = \{x_{ij}\}$ e o vetor $Y = \{y_i\}$ ao algoritmo de aprendizagem de parâmetros da regressão logística, é possível obter o vetor de coeficientes $\boldsymbol{\beta}$. Tomando a exponencial dos coeficientes obtemos estimativas do odds ratio (OR), que nos permitem interpretar o peso de cada uma das variáveis relativamente à conversão.

Esta medida interpreta-se da seguinte forma:
OR > 1: efeito positivo.
OR < 1: efeito negativo.
OR ≈ 1: sem efeito, a palavra em análise não está relacionada com a conversão.

É ainda possível identificar associações de palavras influentes com recurso a algoritmos que detetam e introduzem interações importantes no modelo. Para terminar será necessário cruzar os resultados obtidos com o relatório de termos de pesquisa analisado e selecionar os termos a introduzir e os termos a excluir da campanha.

4. Métrica de distância

No arranque de uma campanha o gestor de campanha utiliza a informação de que dispõe sobre a campanha e as ferramentas disponibilizadas pelo Adwords para escolher um conjunto de Keywords para cada grupo da campanha. Após este passo o gestor dispõe de um conjunto de Keywords para a campanha em arranque, mas não dispõe de histórico.

Um aspeto importante na sugestão de Keywords para uma campanha sem histórico (i.e. em arranque) é conseguir identificar outras campanhas com histórico que sejam semelhantes à campanha para a qual queremos gerar sugestões. Para tal utilizamos a métrica de distância de Jaro-Winkler.

A distância de Jaro-Winkler é uma medida da semelhança entre duas *strings* que retorna 1 caso as *strings* sejam idênticas e 0 caso não haja semelhança alguma [*Winkler90*]. Definimos o operador $JW\left(z_{k_1}^{h_1}, z_{k_2}^{h_2}\right)$ como esta medida de semelhança entre a Keyword k_1 da campanha h_1 e a Keyword k_2 da campanha h_2. A distância entre duas campanhas é então definida como:

$$D_{JW}\ (h_1, h_2) = \frac{1}{l^{h_1}} \sum_{k_1=1}^{l^{h_1}} \max_{k_2} JW\left(z_{k_1}^{h_1}, z_{k_2}^{h_2}\right)$$

5. Semelhança entre campanhas e grupos

Para perceber melhor o sentido da métrica de distância aplicámo-la a 362 campanhas AdWords, resultando na seguinte figura. É possível identificar algumas campanhas isoladas, mas a grande maioria tem alguma semelhança com outras campanhas. Assim sendo, seria possível sugerir ao gestor de campanha o conjunto de palavras chave das campanhas semelhantes.

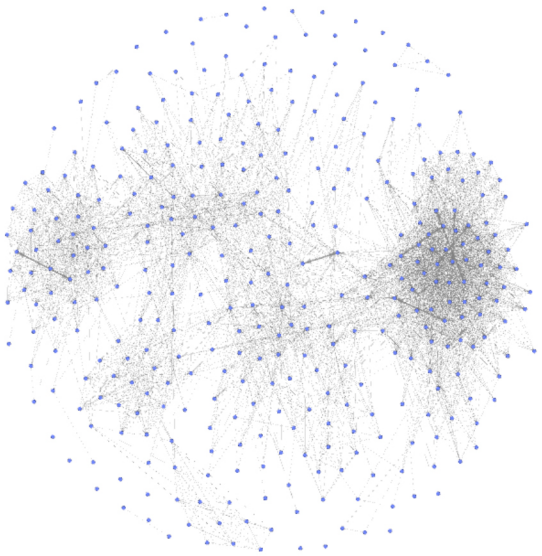

Figura 3. Grafo de semelhança entre campanhas.

Aplicou-se aos grupos individuais a mesma métrica de distância utilizada para as campanhas, resultando na figura seguinte. A maior resolução no grafo indica um maior número de grupos do que campanhas, uma vez que cada campanha tem necessariamente um ou mais grupos. Este grafo revela que existem relações mais fortes entre grupos resultando em 6 núcleos visíveis e um grande número de grupos isolados, ao contrário do que acontece com as campanhas. Concluímos que apesar de o conjunto de palavras-chave dos grupos que constituem as campanhas seja suficiente para identificar campanhas semelhantes, tal não acontece para os grupos e a sugestão de palavras-chave por grupo torna-se pouco viável.

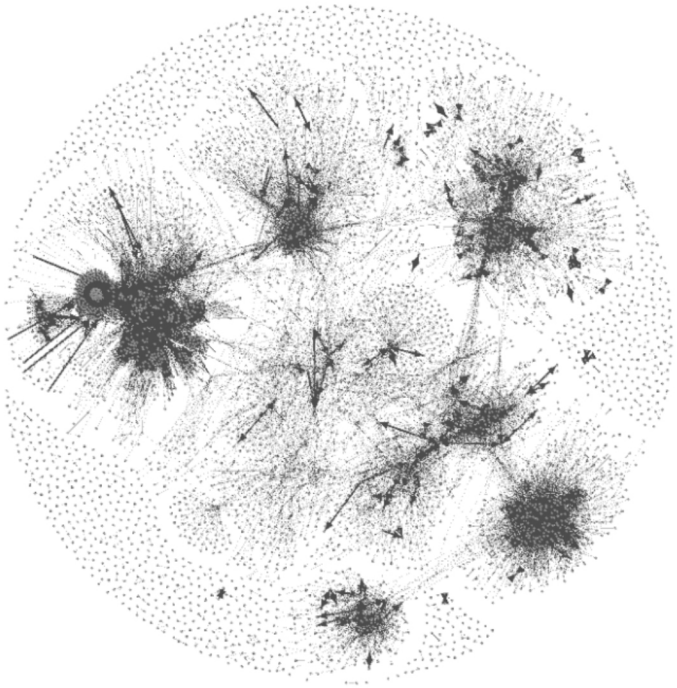

Figura 4. Grafo de semelhança entre grupos.

6. Conclusões

Este artigo explora a análise estatística de campanhas de publicidade em motores de busca para ajudar a identificar palavras chave com efeito negativo na taxa de conversão. Para isso utilizamos uma regressão logística. De modo a ultrapassar a impossibilidade de aplicação deste método em campanhas em arranque, propomos uma abordagem baseada na semelhança de campanhas e propomos utilizar as campanhas semelhantes para a regressão logística. Utilizámos dados reais de centenas de campanhas de Google AdWords para mostrar que para a maioria das campanhas é possível recomendar outras campanhas semelhantes, mas que o mesmo já não é verdade para os grupos. Neste caso a semelhança entre grupos resultou em meia dúzia de núcleos muito fortes e um grande número de grupos isolados.

Referências / *References*

[Winkler90]. Winkler, W. E. (1990). "String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage". Proceedings of the Section on Survey Research Methods (American Statistical Association): 354–359.

[*Hosmer00*]. Hosmer, David W.; Lemeshow, Stanley (2000). Applied Logistic Regression (2nd ed.). Wiley.