Noname manuscript No.
(will be inserted by the editor)

Profitable Keyword Selection: Paid Search Advertising for Online Airplane Ticket Booking

Minghua Jiang · Xuefeng Li · Chih-Ling Tsai · Hansheng Wang

Received: date / Accepted: date

Abstract The selection of key words for bidding is a critical component of paid search advertising. When the number of possible keywords is enormous, it becomes difficult to choose the best keywords for advertising and then subsequently to assess their effect. To this end, we propose a profitable keyword selection algorithm that not only reduces the dimension for selections, but also generates the top listed keywords for profits. An empirical example on paid search advertising for online airplane ticket booking is presented to illustrate the usefulness of this algorithm.

Keywords Keyword Selection · Online Airplane Ticket Booking · Paid Search Advertising · Proxy Variable · Search Engine

1 Introduction

Paid search advertising, also referred to as sponsored search advertisement and paid placement, is particularly attractive to firms for two reasons. First, it targets consumers who have chosen to search for relevant keywords so it generally reaches consumers who are already interested in the product or service offered by the firm. Second, a firm pays for the advertising based on the number of times a consumer clicks on its ad to enter its site and can track what the consumer does subsequently, so firms are

Minghua Jiang
Guanghua School of Management, Peking University, 100871, Beijing, P. R. China
E-mail: jmh@gsm.pku.edu.cn

Xuefeng Li
CubeAD Online Marketing Consulting Co. Ltd., 100022, Beijing, P.R.China
E-mail: lixuefeng_06p@gsm.pku.edu.cn

Chih-Ling, Tsai
Graduate School of Management, University of California–Davis, Davis, CA, 95616, USA
E-mail: cltucd@gmail.com

Hansheng Wang
Guanghua School of Management, Peking University, 100871, Beijing, P. R. China
Tel.: 011-86-10-62757915
E-mail: hansheng@gsm.pku.edu.cn

able to track the effectiveness and profitability of their advertising. It is therefore not surprising that paid search advertising has become so prevalent; revenues exceeded $8 billion in 2007 and are expected to increase to $15.4 billion by 2012 ("US Online Advertising Forecast, 2007 to 2012," *jupiterresearch.com* 2007). Detailed discussions of the use of search engines for advertising can be found in [2], [12], [3], [7], [4], [6], and [9].

In paid search advertising, firms purchase specific keywords (e.g., *Beijing to Shanghai roundtrip discount ticket*, directly translated from Chinese) and create an advertisement that will be appear in the sponsored section of the search results page when a consumer searches for those keywords. Unlike in traditional advertising formats, advertisers do not pay a fixed amount to place a paid search advertisement; instead they bid what they are willing to pay for a single click on a paid search ad. The search engine then used an algorithm to determine the order in which the paid search advertisements will appear. Recently, [10] and [5] both offer excellent discussions of how paid search advertising works and some of the theoretical and technical issues it raises.

To effectively make use of paid search advertising, selecting the keywords on which to bid is one of the most important procedures. The most powerful and useful keywords are able to yield a greater number of clicks and improve the pay-per-click (PPC) conversion rate. As a result, advertiser's (or seller's) profits increase. When the total number of possible keywords is enormous, however, the practice of identifying the most profitable keywords is not a trivial task. For example, suppose that a candidate of keywords consists of S different semantic structures and the j-th semantic component has n_j categories for $j = 1, \cdots, S$. Accordingly, there are $n_1 \times n_2 \cdots \times n_S$ possible keywords being generated. If some of n_j is large (e.g., the number of cities or brands or services), then keyword selection becomes quite challenging. This motivates us to substitute those large categorical semantic components with a more manageable number of proxy variables.

To illustrate the usefulness of proxy variables, we study an aforementioned airline travel example in Mainland China. Consider the keyword *"Beijing to Shanghai roundtrip discount ticket"*. This keyword consists of five semantic variables: the departure city (*Beijing*), the link preference (*to*), the destination city (*Shanghai*), the trip preference (*round-trip*), and the price preference (*discount ticket*). According to the statistics of Civil Aviation Administration of China (CAAC, *www.caac.gov.cn*), as of 2008, more than 150 cities have major airports. Furthermore, there are at least 6 different expressions in Chinese that yield exactly the same meaning as "*to*[1]" and 10 different representations for "*discount ticket*[2]". Moreover, the type of trip has two categories (i.e., "round-trip" and "no word or space").In sum, a total of $150^2 \times 6 \times 10 \times 2 \approx 2.70$ million candidate keywords can be easily constructed. Because working with this many keywords in practice is impractical, we substitute the departure and destination cities with their corresponding economic or geographic proxy variables (e.g., population, gross domestic product, etc). As a result, we not only resolve the problem of the large number of combinations created by departure and destination cities but also link those proxy variables to the advising effect (e.g., the number of clicks and conversion

[1] For example, no word or space, BLANK (an empty space), DAO, FEI, QU, ZHI.

[2] For example, no word or space, PIAN YI, DA ZHE, DI JIA, JIA GE, JIA QIAN, PIAO JIA, TE JIA, YOU HUI, and ZHE KOU.

rates), which can be used for the selection of profitable keywords.

The aim of this paper is to propose a profitable keyword selection algorithm which first replaces the large categories in semantic components with proxy variables. We then estimate regression parameters by linking the mean of the number of clicks (or conversion rates) to semantic variables (induced by semantic components of keywords) and control variables (affecting the performance of semantic variables on the responses, e.g., the search engine provider, the cost of advertising, and the length of keyword). Subsequently, we employ the estimated clicks and conversion rates to identify keywords via their predicted profitability. The rest of the article is organized as follows. Section 2 presents a novel algorithm to select profit keywords. Section 3 employes an empirical analysis on online airplane ticket booking, and Section 4 concludes the article with a short discussion.

2 Profitable Keyword Selection

In the context of search engine advertising, keyword selection plays an important role for paid search advertisement. To improve profits for advertisers, we propose an algorithm that encompasses both dimension reduction and regression estimations. For the sake of illustration, we first define variables and notations. Let the response variable Y be an $n \times 1$ vector and let the semantic variables $\tilde{X}$ generated from the m semantic components of keywords be an $n \times \tilde{p}$ matrix. In addition, let control variables Z that affect the performance of keywords to the response variable be an $n \times q$ matrix. Without loss of generality, we assume that the j-th semantic component has large categories, and we label its resulting high dimension submatrix of $\tilde{X}$ as $\tilde{X}_{(j)}$. Next, we propose the profitable keyword selection (PKS) algorithm.

Step I: Substitute $\tilde{X}_{(j)}$ with the low dimension matrix $X_{(j)}$ that is generated by the proxy variables of the j-th semantic component. Call the resulting explanatory matrix X with dimension $n \times p$.

Step II: Randomly split the data into learning sample and testing sample with equal size. Using the learning sample, estimate regression parameters by linking the mean functions of Y_c (the number of clicks) and Y_r (the conversion rate) to the learning sample variables X and Z. Then, apply the fitted regression equations to the testing sample data (X^*) to compute the predicted clicks ($\widehat{click}^*$) and conversion rates ($\widehat{rate}^*$).

Step III: Compute the profits for the i-th testing sample,

$$\widehat{profit}^*_{(i)} = profit\ margin \times \widehat{click}^*_i \times \widehat{rate}^*_i - \overline{cost} \times \widehat{click}^*_i \quad (i = 1, \cdots, n^*),$$

where n^* is the size of training sample, $profit\ margin$ is chosen by the advertiser *a priori*, and $\overline{cost}$ is the sample mean of the *cost* variable in the training sample.

Step IV: Sort those keywords in the testing sample according to their predicted profitability (i.e., $\widehat{profit}^*_i$) in descending order. Subsequently, select m out of n^* keywords

with the highest predicted profitability, i.e., $\widehat{profit}^*_{(i)}$ with $1 \leq i \leq m$, whose performance can then be evaluated as

$$Average\ Profit = \frac{1}{m}\sum_{i=1}^{m} \widehat{profit}^*_{(i)}.$$

To better understand the out-of-sample performance of this algorithm, we further define a relative measure,

$$Relative\ Profit = \frac{Average\ Profit}{Baseline\ Profit},$$

where the $Baseline\ Profit$ is simply the sample mean of $\widehat{profit}^*_i$ over the entire testing sample. If the $Relative\ Profit$ is larger than 1, then PKS is superior to the random guessing selection.

Step V: To mitigate sampling errors, repeat Steps II through IV R times. Afterwards, compute the mean of $Relative\ Profit$ across R realizations. Then, for each level of $Profit\ Margin$, advertisers are able to choose the best size, m^*, of the selected set of keywords to meet their own interest.

Step VI: Use the whole sample data to obtain the estimates $\widehat{click}$ and $\widehat{rate}$, and then compute $\widehat{profit} = profit\ margin \times \widehat{click} \times \widehat{rate} - cost \times \widehat{click}$. Subsequently, sort keywords via their predicted profitability and select the top m^* keywords for any given $profit\ margin$. Furthermore, the above equation is applicable for the out-of-sample predictions.

To illustrate the application of PKS in practice, we next present an empirical example on online airplane ticket booking.

3 Empirical Analysis

3.1 Data and Variables

We consider online airplane ticket booking data collected by one of the major online travel planners in mainland China. It contains a total of 27,663 keywords and their corresponding performance measures on Google and/or Baidu during the period of 10/2008–03/2009. As mentioned in Section 1, the semantic structure of those keywords is formulated as follows:

$$KeyWord = Departure\ City + Link\ Preference + Destination\ City$$

$$+Trip\ Preference + Price\ Preference.$$

Because the *Departure City* and *Destination City* consist of more than 150×150 combinations, we follow Step I of the PKS algorithm to replace them by their proxy variables, population size (POP) and gross domestic product (GDP). This together with those semantic components discussed in Section 1 yields two proxy variables X_1 (POP_1) and X_2 (GDP_1) for *Departure City*, two proxy variables X_3 (POP_2) and X_4 (GDP_2) for *Destination City*, five dummy variables (X_5 to X_9) for the six categories of $Link\ Preference$, one dummy variable (X_{10}) for the two levels of $Trip\ Preference$,

and nine variables (X_{11} to X_{19}) for the ten categories of $Price\ Preference$. The baseline for creating those dummy variables used in semantic components, $Link\ Preference$, $Trip\ Preference$, and $Price\ Preference$ is the category of "No Word or Space". In sum, this study consists of 19 semantic variables, X.

Since the search engine provider (SEARCH ENGINE), the cost of advertisement (COST), and the length of keyword (LENGTH) can affect the performance of semantic variables on response variables Y_c (the number of clicks) and Y_r (the conversion rate), we denote these as variables, Z_1 to Z_3, which constitute control variables Z. Furthermore, we define $Z_1 = 1$ when the search engine is Google and $Z_1 = 0$ when the search engine is Baidu. Finally, we transform Y_c and Y_r to $\log Y_c$ and $\log(Y_r + 1)$, respectively, so that the resulting regression on X and Z can yield better results for forecasting profits.

3.2 Keyword Selection for Paid Search Airplane Ticket Advertising

Following Steps II through V of the PKS algorithm, we compute relative profits across $R = 100$ realizations for four levels of $Profit\ Margin$, 5, 10, 15 and 20 RMB (Renminbi, Chinese paper currency) and for four sizes of the selected keyword set, $m =$ 100, 200, 500, and 1000. To enable us to assess the effect of the control variables, we fit the data to two ordinary linear regression models: Model A regresses $\log Y_c$ (and $\log(Y_r + 1)$) on X and Z, while Model B regresses $\log Y_c$ (and $\log(Y_r + 1)$) on X.

Table 1 Relative Profit. Model A: the full model with variables X and Z; Model B: the sub-model with variable X only.

		Profit Margin			
Size (m)	Model	5	10	15	20
100	A	8.99	8.95	9.69	8.53
	B	5.94	5.00	4.90	4.85
200	A	7.28	6.96	6.77	6.71
	B	4.88	4.98	5.05	5.07
500	A	4.96	4.88	4.83	4.81
	B	4.03	4.16	4.18	4.16
1000	A	3.74	3.67	3.61	3.58
	B	3.34	3.23	3.18	3.16

Table 1 reports relative average profits across the levels of profit margins and the sizes of selected keyword sets, and they are all larger than 1. This indicates that PKS yields greater profits than those generated by the random selection of a keyword set. It is not surprising that the relative profits deceases as the size increases. It is also interesting to note that all four levels of $Profit\ Margin$ yield similar relative profits for a given size. Overall, the combination of $Profit\ Margin = 5$ and $m = 100$ generates the highest relative profit, which is almost nine times the average profit obtained

from random selections. Moreover, Model B's relative profits are substantially less than those of Model A when $m = 100$ and 200, so we conclude that control variables are valuable for keyword selections.

Table 2 Parameter estimation and testing results for the number of clicks

Semantic Component	Explanatory Variable	Parameter Estimate	Standard Error	Test Statistic	p Value
	INTERCEPT	-0.681	0.188	-3.612	0.000
Departure City	POP_1 (X_1)	0.209	0.008	26.138	0.000
	GDP_1 (X_2)	0.082	0.012	6.797	0.000
Destination City	POP_2 (X_3)	0.176	0.008	22.791	0.000
	GDP_2 (X_4)	0.051	0.012	4.305	0.000
Link Preference	BLANK (X_5)	0.552	0.023	24.169	0.000
	DAO (X_6)	1.213	0.023	53.354	0.000
	FEI (X_7)	0.205	0.026	7.928	0.000
	QU (X_8)	-0.120	0.034	-3.519	0.000
	ZHI (X_9)	1.078	0.026	41.070	0.000
Trip Preference	WANG FAN (X_{10})	-0.099	0.110	-0.899	0.368
Price Preference	PIAN YI (X_{11})	-0.922	0.080	-11.565	0.000
	DA ZHE (X_{12})	-0.091	0.022	-4.071	0.000
	DI JIA (X_{13})	-0.977	0.123	-7.953	0.000
	JIA GE (X_{14})	-0.095	0.048	-1.978	0.048
	JIA QIAN (X_{15})	-0.875	0.216	-4.052	0.000
	PIAO JIA (X_{16})	-0.548	0.051	-10.653	0.000
	TE JIA (X_{17})	0.249	0.021	11.669	0.000
	YOU HUI (X_{18})	-1.336	0.146	-9.135	0.000
	ZHE KOU (X_{19})	-0.762	0.031	-24.780	0.000
	SEARCH ENGINE (Z_1)	-1.296	0.048	-27.235	0.000
	COST (Z_2)	-0.205	0.011	-18.346	0.000
	LENGTH (Z_3)	-0.270	0.008	-31.810	0.000

Since relative average profits are greater than 1, we next follow Step VI of the PKS algorithm to fit the whole data using Model A. Tables 2 and 3 report detailed parameter estimations and their corresponding standard errors, test statistics, and p values. We summarize our key findings below.

– Table 2 indicates that Google is less effective in producing a large number of clicks than Baidu in the Chinese market. However, Google is very comparable to Baidu in terms of conversion rate (see the non-significant result in Table 3).

– Table 3 shows that higher cost keywords are typically associated with better conversion rates. Accordingly, the marketing competition for such keywords is keen, which results in the smaller number of clicks (see the negative and significant estimation result for COST in Table 2).

Table 3 Parameter estimation and testing results for the number of conversion rates

Semantic Component	Explanatory Variable	Parameter Estimate	Standard Error	Test Statistic	p Value
	INTERCEPT	-0.163	0.059	-2.761	0.006
Departure City	POP_1 (X_1)	0.009	0.003	3.703	0.000
	GDP_1 (X_2)	0.011	0.004	2.994	0.003
Destination City	POP_2 (X_3)	0.018	0.002	7.445	0.000
	GDP_2 (X_4)	0.009	0.004	2.425	0.015
Link Preference	BLANK (X_5)	0.021	0.007	2.929	0.003
	DAO (X_6)	0.002	0.007	0.285	0.776
	FEI (X_7)	0.008	0.008	1.009	0.313
	QU (X_8)	-0.014	0.011	-1.323	0.186
	ZHI (X_9)	-0.007	0.008	-0.848	0.397
Trip Preference	WANG FAN (X_{10})	0.020	0.034	0.594	0.553
Price Preference	PIAN YI (X_{11})	0.050	0.025	1.980	0.048
	DA ZHE (X_{12})	0.020	0.007	2.887	0.004
	DI JIA (X_{13})	0.035	0.039	0.900	0.368
	JIA GE (X_{14})	-0.079	0.015	-5.209	0.000
	JIA QIAN (X_{15})	-0.096	0.068	-1.420	0.156
	PIAO JIA (X_{16})	-0.095	0.016	-5.901	0.000
	TE JIA (X_{17})	0.058	0.007	8.622	0.000
	YOU HUI (X_{18})	0.071	0.046	1.554	0.120
	ZHE KOU (X_{19})	0.029	0.010	3.040	0.002
	SEARCH ENGINE (Z_1)	0.000	0.015	-0.012	0.990
	COST (Z_2)	0.006	0.004	1.806	0.071
	LENGTH (Z_3)	-0.006	0.003	-2.372	0.018

– Both tables demonstrate that longer keywords generate fewer clicks and lower conversion rates.

– Table 2 shows that different choices of *Link Preference* and *Price Preference* do affect the resulting keyword's performance in the number of clicks, while Table 3 indicates those choices have less effect on the conversion rate. Because the profit depends on both estimations of clicks and rates, the semantic components *Link Preference* and *Price Preference* should be included in keyword selections. In contrast, the semantic component, *Trip Preference*, does not play a significant role in clicks and conversion rates. In particular, its baseline is "No Word or Space," which suggests this component can be removed.

– Tables 2 and 3 indicate that both the departure and destination cities' size (POP) and income (GDP) are significant and positively related to the number of clicks and conversion rates. Hence, these two proxy variables play an important role in keyword selections.

In sum, the control variables, SEARCH ENGINE, COST, and LENGTH, are useful for keyword selection since their estimations are significant in either explaining clicks or conversion rates (see Tables 2 and 3). In addition, the semantic variables in-

duced by the semantic components, $Link\ Preference$ and $Price\ Preference$, as well as those replaced by the proxy variables POP and GDP are essential for identifying profitable keywords; in contrast the semantic component $Trip\ Preference$ is not useful. Moreover, variables X_5 and X_6 lead to the largest rates and clicks, respectively, in $Link\ Preference$, and variables X_{16} yields both the largest clicks and rates in $Price\ Preference$. Accordingly, those variables should be considered as potential candidates for future keyword selections. Finally, advertisers can use the relative profits reported in Table 1 to determine the best size for a given $Profit\ Margin$ and then follow Step VI to select the top m^* keywords.

4 Concluding Remarks

In paid search advertising, we propose a novel algorithm to select keywords for their profitability. We substitute the semantic component with high dimensional categories by proxy variables, and then split the entire data into learning and testing samples for estimations and predictions, respectively, to construct the relative profit index. This allows us to select the most profitable keywords from the whole sample. To facilitate the use of PKS algorithm, we need to further investigate model structures, variables, and estimations. For example, we may employ the generalized linear model or multivariate regression model to model online air ticket data, and subsequently employ the Akaike Information criterion [1] or Bayesian information criterion [11] to select relevant semantic and control variables. Detailed illustrations regarding those models and selections can be found in [8]. Finally, identifying appropriate proxies for semantic components of the keyword is a very interesting and challenging topic that warrants further study.

Although we only present an empirical study of online airplane ticket booking, the PKS algorithm is easily applied or extended to many paid search advertising subjects (e.g., hotel booking, online sales, and online services). Accordingly, this algorithm will allow advertisers to improve the profitability of online advertisement. We believe that our algorithm could also be used by search engine providers to improve their algorithms assigning keywords and positions for paid search advertising and so improve both their market share and their profitability. Furthermore, the use of our PKS algorithm could contribute to addressing two significant problems in search engine advertising, and paid search advertising in particular: search engine optimization (SEO) and the mitigating click fraud; see [13] for a discussion of this problem. Consequently, we believe that our proposed algorithm will strengthen the field of online advertising from a variety of perspectives.

References

1. Akaike, H. (1973), Information theory and an extension of the maximum likelihood principle, *In 2nd International Symposium on Information Theory,* Ed. B. N. Petrov & F. Csaki, 267-281. Budapest: Akademia Kiado.
2. Bradlow, E. and Schmittlein, D. (2000), The little engine that could: modeling the performance of word wide web search engines, *Marketing Science*, 19(1), 43–62.
3. Chen, Y. and He, C. (2006), Paid placement: advertising and search on the Internet, *Working Paper*, University of Colorado, Boulder, CO.

4. Edelman, B., Ostrovsky, M., and Schwarz, M. (2007), Internet advertising and the generalized 2nd-price auction: selling billions of dollars owrth of keywords, *American Economics Review*, 97(1), 242–259.
5. Ghose, A. and Yang, S. (2009), An empirical analysis of search engine advertising: sponsored search in electronic markets, *Management Science*, Forthcoming.
6. Goldfarb, A. and Tucker, C. (2007), Search engine advertising: pricing ads to context, *Working Paper*.
7. Manchanda, P., Dube, J., Goh, K., and Chintagunta, P. (2006), The effect of banner advertising on internet purchasing, *Journal of Marketing Research*, 43(1), 98–108.
8. McQuarrie, D. R. and Tsai, C. L. (1998), *Regression and Time Series Model Selection.* World Scientific: Singapore.
9. Plummer, J., Rappaport, S., Hall, T., and Barocci, R. (2007), *The Online Advertising Playbook.* Wiley & Sons: New Jersey.
10. Rutz, O. J. and Bucklin, R. E. (2007), A model for individual keyword performance in paid search advertising, *Working Paper*.
11. Schwarz, G. (1978), Estimating the Dimension of a Model, *The Annals of Statistics*, 19(2), 461–464.
12. Telang, R., Boatwright, P., and Mukhopadhyay, T. (2004), A mixture model for Internet search engine visits, *Journal of Marketing Research*, XLI (May), 206–241.
13. Wilbur, K. C. and Zhu, Y. (2009), Click Fraud, *Marketing Science*, 28(2), 293–308.