RESEARCH ON OPTIMAL SELECTION STRATEGY OF SEARCH ENGINE KEYWORDS BASED ON 0-1 KNAPSACK

Han NIE[1]

Keyword advertising business is an extremely important source of profit for search engine companies. Search engine companies have designed a complete set of keyword bidding schemes for search advertising, in which advertisers participate in keyword bidding. Advertisers need to select a certain number of keywords, hoping to gain maximum profit. With limited advertising budget resources, it is impossible for advertisers to evenly allocate resources to each keyword. How to reasonably select keywords from the existing set of candidate keywords to maximize overall revenue is a problem faced by every advertiser. This paper will set up a keyword optimization model based on 0-1 knapsack. In this paper dynamic programming is used to attain the result. The model proposed in this article helps advertisers select keywords through their performance parameters (such as search demand, click rate, cost per click, etc.). This paper designs an experimental scenario. Firstly, an initial set of keywords is obtained through keyword mining tools, and then the model in this paper is used to solve and provide keyword selection results. The analysis of experimental results shows that the model proposed in this paper can recommend a keyword set with a total revenue superior to other methods within the budget range.

Keywords: Search Advertisement; Keyword Strategy; Keyword Selection; Knapsack Problem; Dynamic Programming

1. Introduction

As an efficient means of business promotion, search advertising has become the right-hand man for many enterprises to explore the market. The keyword advertisements provided by the existing mainstream search engines (such as Google AdWords) usually use the generalized two-price "keyword bidding" mechanism (GSP)[1]. In each search promotion campaign, advertisers will need to select a certain number of keywords for bidding. And search users with specific needs will input keywords on the search engine to search for the products they are interested in[2]. Search engine results pages typically display both organic and paid results. If consumers click on the paid search results, they can enter the website expected by the advertiser. At this time, the advertiser must pay for the click according to a certain price mechanism.

In the whole process of carrying out search engine advertisements, advertisers are faced with multiple keyword selection decisions [3, 4]. Before the

[1] School of Management, North Sichuan Medical College, Nanchong, Sichuan, China, e-mail: niehan@nsmc.edu.cn

start of the advertising campaign, the advertiser needs to select several keywords within the budget to form an initial ad group for bidding. After the start of an advertising campaign, advertisers have to constantly adjust keywords through market feedback performance on the initial advertising group, and constantly observe the adjusted market effect [5]. Due to the diverse search habits of consumers and the large number of keywords related to advertisers' products or services, the number of candidate keywords will be large. With limited advertising budget resources, it is impossible for advertisers to evenly allocate resources to each keyword. How to reasonably select keywords from the existing set of candidate keywords to maximize the overall revenue is a problem faced by every advertiser. On the one hand, advertisers hope that the keywords they choose can cover various target groups as much as possible, so that the search keywords used by potential consumers can basically be included in the advertiser's bidding advertising plan; on the other hand, advertisers need to consider the advertising performance of selected keywords, especially if the budget is limited. Search keywords can generally be divided into two categories: popular keywords and long-tail keywords [6]. Popular keywords refer to keywords with a large search volume, but such keywords are usually very expensive, resulting in excessive budget costs for advertisers. So they are not cost-effective in terms of comprehensive revenue. Some keywords are long-tail keywords, which contain more precise and targeted words. Long-tail keywords have non-volume and non-obvious characteristics [7], but the probability of customers brought by them converting to website product customers is higher than that of target keywords, and cheaper than popular keywords, so such keywords can help advertisers achieve the goal of maximizing total revenue.

As a result, choosing suitable keywords has become one of the difficult problems for advertisers to carry out search engine advertisements. Aiming at this problem, this paper proposes a keyword selection model for advertisers to choose keywords based on 0-1 knapsack, which helps keyword selection through the performance parameters of keywords (such as search demand, click rate, cost per click, etc.). The 0-1 multiple knapsack problem is a valuable NP-hard problem involved in many science-and-engineering applications [8]. In the optimization of advertising strategy, the keyword selection optimization problem can be transformed into a 0-1 Knapsack problem [9]. This paper designs an experimental scenario. First, an initial keyword set is obtained through a keyword mining tool, and then the model is used to solve the problem and the keyword selection results are given. Finally, the experimental results are verified and evaluated. The analysis of experimental results shows that the 0-1 knapsack keyword selection model proposed in this paper can recommend a keyword set whose total revenue is better than other methods within the budget.

The follow-up content of this article will be organized as follows: the second part will introduce the research status of keyword selection strategies at home and

abroad; the third and fourth parts will introduce the 0-1 knapsack algorithm model and introduce the solution process in detail; The experiment was carried out using word sets, the optimal solution was calculated by using the model, and finally the results were analyzed and summarized; the sixth part is the work summary and future prospects.

2. Literature review

As an important part of search bidding advertisement, keyword selection has attracted the interest of many researchers at home and abroad. Current research first focuses on how to extract and generate keywords from various original pages and query logs. However, the keyword groups obtained in this way belong to the popular keyword set to a large extent [10], as analyzed above, this does not bring better comprehensive benefits for advertisers. Therefore, further research often focuses on recommending keywords through semantic similarity at the concept level or other feature weights [11].

There are also some studies that focus on keyword selection to measure benefits. In the work of [12, 13], heuristics-based method is used to design comparisons under the assumption that the cost-profit ratio of each keyword is known. Although many scholars at home and abroad try to improve the effect of keyword selection through research from different angles, the ultimate measure is still the total revenue obtained by advertisers [14, 15]. Recently, there have been some related studies conducted from the perspective of advertising structure or advertising matching [16, 17]. In summary, these studies are not valuable to build domain-specific keyword pools. In this sense, none of keyword generation methods based on a single information source can provide the perfect solution because each of them has its own advantages and shortcomings. The research focus of this paper is how to maximize the total revenue of the selected keyword set. By establishing an optimization model with strong versatility and high accuracy, the complexity of keyword selection is reduced, and the solution is given and tested and evaluated. The method provided in this article complements these branches of works, and thus it calls for a benchmark study integrating the sources of information to generate relevant keywords.

3. Problem statement and model

Search bidding promotion is a continuous process, and keyword decision-making runs through the entire life cycle of search advertising. First of all, the advertiser needs to determine the initial keyword set based on the product characteristics and business of the main product; then the advertiser conducts a detailed analysis on the basis of the initial keyword set, combined with the budget and competitors, combined with different search engine bidding systems. The

promotion mode adjusts the keyword set according to different advertising structures; finally, the search engine bidding platform will require advertisers to assign keywords according to certain rules to form ad groups and cooperate with text advertisements (ad copy) to participate in search bidding. In the later stage, advertisers will also constantly adjust the keyword combination through market feedback performance in order to obtain better promotion effects.

Advertisers measure the effect of promotion mainly on advertising revenue, specifically, they expect to maximize the total revenue of selected keywords. The acquisition of income is based on the pre-investment of the advertising budget, and the decision-making space for keyword selection is limited by the advertising budget. An unlimited budget will greatly simplify the keyword selection process, that is, select all keywords that may generate positive revenue. However, advertisers usually have limited budgets in practice. Therefore, this paper considers the keyword selection strategy in the case of limited budget.

The goal of keyword selection decision-making is to maximize the total revenue that all selected keywords can bring under the budget constraint. Thus, this paper defines keyword decision-making as the following integer programming model:

$$\begin{aligned} &Max\sum_{i=1}^{N}(v(i)-p(i))\times c(i)\times d(i) \\ &s.t.\sum_{i=1}^{N}p(i)\times c(i)\times d(i)\le B \end{aligned} \tag{1}$$

Where v(i) represents the value per click (value-per-click) of keyword i in the keyword set; p(i) represents the cost-per-click (cost-per-click) of keyword i; c(i) represents The click-through rate (click-through-rate) of keyword i, d(i) represents the search volume of keyword i, and B represents the advertiser's budget.

This paper solves the 0-1 knapsack model in the following scenario: Advertisers need to choose among the initial keywords, each keyword can only be selected once at most, and a single keyword is either selected or discarded, the ultimate goal of the model solution is to maximize the total revenue of the selected keyword set.

There are two types of solutions to the 0-1 knapsack problem, exact and approximate. The former is represented by dynamic programming, which can obtain the optimal solution; the latter is represented by Greedy Algorithm, which has great advantages in coding and execution speed, but often only gets an approximate solution to the problem. For optimal results, this paper will use dynamic programming to solve the model.

A feasible algorithm for dynamic programming solution is given below:

Let B be the budget of the advertiser, $V(i, j)$ represents the maximum revenue of all keywords that can be loaded into the backpack with budget j among

the first i keywords, C_i is the cost of keyword i, $C_i = p(i) \times c(i,k) \times d(i)$; V_i is the income of the keyword i, $V_i = (v(i) - p(i)) \times c(i,k) \times d(i)$; the array X[i] is used to store the state of the keyword loaded into the backpack.

4. Experimental preparation

The experiment collected the data of a certain sports brand's main product in the bidding system of a certain search engine and designed a corresponding keyword optimization experiment scenario. The experiment used the keyword planning tool to generate more than 2,000 keywords after recommendation. Taking into account relevant factors such as search volume and browsing volume, deleting keywords with extremely low search volume, and taking into account their relevance to the brand's main products, an initial set of 207 keywords was ultimately obtained.

In the following experiments, the experiment will use the performance parameters of the keyword set, such as search demand (SD), the average monthly search volume, click-through rate (CTR), and cost-per-click (CPC), all of which can be obtained from the Google Keyword Planning Tool. The value per click v(i) of the website needs to be calculated from past data, but because it is impossible to evaluate the actual activities of advertisers at this stage, and it is difficult to obtain data from search engine operators, it is difficult to obtain data from advertisers. It involves commercial secrets, for the convenience of calculation, the value of v(i) is set by us according to the normal distribution (set =30, =1). The revenue of a keyword is affected by many small independent random factors, so it is reasonable to assume that the revenue of a keyword obeys a normal distribution.

Table 1

Total revenue of the selected keywords

Advertiser's budget	50	70	100	150	200	300
The number of selected keywords	93	95	96	98	100	103
Total revenue	2796	2837	2882	2957	3023	3108

5. Results evaluation

From the experimental results in Table 1, it is obvious that with the increase of the advertiser's budget, the number of selected keywords will gradually increase, but the increase rate is relatively stable, which shows that this experimental model can better select the cost-biased keywords. Low-cost but high-profit keywords, those popular keywords with high cost and low comprehensive income can be well filtered out, so as to ensure the optimal total income of the selected keywords to the greatest extent.

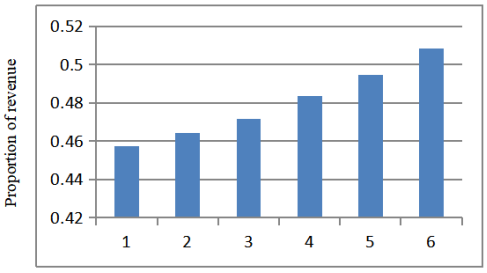

Fig. 1. Total income of all candidate keywords under 6 budget schemes

Fig. 1 shows the ratio of the total income of selected keywords to the total income of all candidate keywords under the six budget schemes set in the experiment.

The experimental results are evaluated as follows. According to the formula (1), calculate the income V_i of each selected keyword:

$$V_i = (v(i) - p(i)) \times c(i,k) \times d(i) \tag{2}$$

The purpose of the model recommendation in this paper is to maximize the total revenue of all selected keywords, namely:

$$\max \Sigma V_i \tag{3}$$

Suppose H is the initial keyword set, k is the selected keyword set, i is the selected keyword, and m is the unselected keyword, then:

$$i \in k$$

$$m \notin k$$

$$m, \quad \mathrm{k} \in \mathrm{H}$$

Because the value of v(i) in formula (2) is set by us according to the normal distribution, the problem of maximizing the total keyword revenue is actually transformed into minimizing the percentage of the total cost of the selected keywords in the total cost of all initial keywords question:

$$\min \frac{\Sigma C_i}{\Sigma C_{\mathrm{i}} + \Sigma \mathrm{C}_{\mathrm{m}}} \tag{4}$$

C_m in formula (4) is the cost of each unselected keyword.

Through the comparative analysis in the keyword tool, the experiment found that when the advertiser's budget changes, the keyword's performance parameters will also change accordingly, mainly because the advertiser's budget will affect the bidding ranking. Taking the budgets of 50, 100, 150, 500, 1000, 1500, 2000, and 5000 as numerical values, the performance parameters of the keywords can be obtained. The experiment took the 98 keywords selected when the budget was 150 as an example. It can obviously find that as the advertiser's daily budget increases, the proportion of the cost of the keywords selected in this experiment to

the total cost of all 207 keywords shows a downward trend, as shown in Fig. 2 for details:

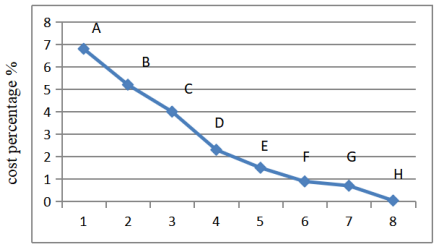

Fig. 2. Cost in the advertiser’s daily budget

Points A, B, C, D, E, F, G, and H in Fig. 2 represent the selected keywords cost percentage when the advertiser’s daily budget is 50, 100, 150, 500, 1000, 1500, 2000, and 5000, respectively.

The experiment compare the results of keyword recommendation with the baseline algorithm under the same budget conditions and the results of this experiment. The Baseline algorithm first calculates the CTR/CPC values of all initial keywords, and then sorts them in descending order. Next, the experiment start to select keywords with the smallest CTR/CPC value until reach the same budget limit as this experiment. The results of this algorithm show that the number of keywords finally selected is less than the result of this experimental model, and the total income is also lower than this experimental result (see Table 2).

Table 2

Total revenue

	The number of selected keywords	Total revenue
baseline	86	2546
KP	98	2957

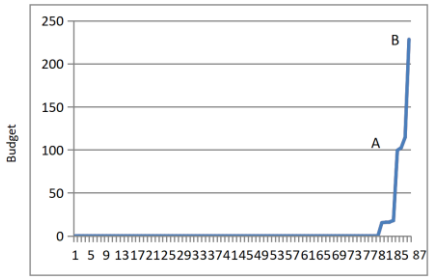

Fig. 3. Cost accumulation curve

Fig. 3 shows the cost accumulation curve. It can be observed that under the constraint of a budget of 150, only point A it can be selected at most (the cost accumulation of the 86th keyword, and point B is the 87th keyword).

6. Analysis of experimental results

According to the above experiments and results, here are the following conclusions:

(1) From table 1, with the increase of the advertiser's budget, this model can better select keywords with low cost but high profit, and the number of selected keywords will not increase blindly, those popular keywords that are not cost-effective can be well filtered out, so as to ensure the optimal total revenue of the selected keywords to the greatest extent and make advertisers' budgets worthwhile. Moreover, as the advertiser's budget increases, the proportion of the total income of the selected keywords to the total income of all candidate keywords gradually increases.

(2) From Fig. 2, in the case of a normal distribution of keyword revenue, as the advertiser's budget increases, the proportion of the cost of the keywords selected in this experiment to the total cost will increase. The smaller the value, it means that as long as the overall trend of the revenue of each keyword does not change, the advertiser can obtain a larger total revenue by choosing the keywords recommended in this experiment than by choosing other keywords, and the increase in this revenue will increase with the Expanded with an increase in the budget. This reflects the emphasis of the 0-1 knapsack model on the weight of total revenue.

(3) By comparing with the results recommended by the baseline algorithm, it can found that 83 words in the recommended results of this experiment are the same as those recommended by the baseline method. The different keywords in the recommended results of the two methods are shown in Table 3:

Table 3

Recommended results

baseline	air rift ,casual,basketball
KP	air max 95,elite,lebronjames,drift,air force one,high top,men, football boots,kobe,dart,discount,air rift,clubs,black,retro

In Table 3, by analyzing the keyword performance parameters, it can find that air rift, casual, and basketball are all keywords with high $p(i)\times c(i,k)\times d(i)$ values, so they were not selected in the dynamic programming solution, and other 15 keywords with lower costs were selected instead. Therefore, the number of keywords selected in the final result of this experiment is more, and the total income is also higher than that of the baseline method.

(4) The keyword performance parameters required for the calculation of this experimental model can be easily obtained from many search engine tools or third-party tools, and advertisers can also derive them from the past data of the Web-log, without introducing complex parameters such as semantic similarity , which makes

the experimental model very convenient to use, has few restrictions, reduces the complexity of keyword selection, and is suitable for use in a variety of situations.

(5) This experiment is based on the premise that the keyword value obeys the normal distribution. If there are more assumptions about the keyword revenue distribution, just modify v(i) in the formula (1), and the model can still The optimal solution is given; if there is an uncertain revenue distribution, then the mean and variance of unknown keyword revenue can be estimated based on the distribution of known keyword revenue, and then the model is used to obtain the result.

7. Conclusions

This paper proposes a keyword selection model based on 0-1 knapsack to help advertisers choose keywords through keyword performance parameters. This paper designs an experimental scenario. First, an initial keyword set is obtained through a keyword mining tool, and then the model is used to solve the problem and the keyword selection results are given. Finally, the experimental results are verified and evaluated. The experimental results show that the total revenue of the keywords recommended is better than other methods within the budget range, and the revenue increase will become larger as the budget increases.

In the follow-up work, further work can be improved: (1) collect more advertisers' actual keyword data in search promotion for experimental analysis and comparison; (2) the estimation of word performance parameters makes the recommendation of the model more accurate.

Acknowledgment

This work is partially supported by North Sichuan Medical College 2021 Doctoral Initiation Fund (CBY21-QD36), Project of Nanchong City and University Cooperation (22SXQT0259). The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

R E F E R E N C E S

[1]. *Liu-Thompkins, Yuping.* "A decade of online advertising research: What we learned and what we need to know." Journal of advertising 48.1 (2019): 1-13.

[2]. *Olbrich, Rainer, Patrick Mark Bormann, and Michael Hundt.* "Analyzing the Click Path Of Affiliate-Marketing Campaigns: Interacting Effects of Affiliates' Design Parameters With Merchants' Search-Engine Advertising." Journal of Advertising Research 59.3 (2019): 342-356.

[3]. *Symitsi, Efthymia, Raphael N. Markellos, and Murali K. Mantrala.* "Keyword portfolio optimization in paid search advertising." European Journal of Operational Research 303.2 (2022): 767-778.

[4]. *Arroyo-Cañada, Francisco-Javier, and Jaime Gil-Lafuente.* "A fuzzy asymmetric TOPSIS model for optimizing investment in online advertising campaigns." Operational Research 19.3 (2019): 701-716.

[5]. *Polato, M., Demchenko, D., Kuanyshkereyev, A., and Navarin, N.* "Efficient Multilingual Deep Learning Model for Keyword Categorization." 2021 IEEE Symposium Series on Computational Intelligence (SSCI). IEEE, 2021.[6].

[6]. *Li, Huiran, and Yanwu Yang.* "Keyword targeting optimization in sponsored search advertising: Combining selection and matching." Electronic Commerce Research and Applications 56 (2022): 101209.

[7]. *Nie, Han, Yanwu Yang, and Daniel Zeng.* "Keyword generation for sponsored search advertising: Balancing coverage and relevance." IEEE intelligent systems 34.5 (2019): 14-24.

[8]. *Valentina Cacchiani, Manuel Iori, Alberto Locatelli, and Silvano Martello*. "Knapsack problems — An overview of recent advances. Part I: Single knapsack problems." Computers & Operations Research (2022) :105692.

[9]. *Amiri, Ali, and Barkhi R.,* "A Lagrangean based solution algorithm for the multiple knapsack problem with setups." Computers & Industrial Engineering (2021):153.

[10]. *Krasňanská, D., Komara, S., and Vojtková, M.* "Keyword categorization using statistical methods". TEM Journal 10.3(2021): 1377-1384.

[11]. *Campos, R., Mangaravite, V., Pasquali, A., Jorge, A., Nunes, C., and Jatowt, A.* "YAKE! Keyword extraction from single documents using multiple local features". Information Sciences 509 (2020): 257-289.

[12]. *Scholz, M., Brenner, C.,* and *Hinz, O.* "AKEGIS: automatic keyword generation for sponsored search advertising in online retailing." Decision Support Systems 119(2019):96-106.

[13]. *Azad, H. K., and* Deepak, *A.* "Query expansion techniques for information retrieval: A survey". Information Processing & Management 56.5(2019): 1698-1735.

[14]. *Bai, X., and Cambazoglu, B. B.* "Impact of response latency on sponsored search". Information Processing & Management 56.1(2019), 110-129.

[15]. *Childers, Courtney Carpenter, Laura L. Lemon, and Mariea G. Hoy.* "# Sponsored# Ad: Agency perspective on influencer marketing campaigns." Journal of Current Issues & Research in Advertising 40.3 (2019): 258-274.

[16]. *Yang, S., Pancras, J., and Song,Y.A.* "Broad or exact? Search Ad matching decisions with keyword specificity and position."*Decision Support Systems*143(2021):113491.

[17]. *Symitsi, E., Markellos, R.N., and Mantrala, M.K.* "Keyword portfolio optimization in paid search advertising." *European Journal of* Operational *Research*,303.2(2022):767-778.