Using Decision Trees to Identify Key Factors of Keyword Advertisements

Long-Sheng Chen*, Tai-Cheng Kuo

Abstract—**In keyword advertising, how to create a high click rate is a critical issue. Briefly speaking, the advertisers buy the preferred keywords. When the search engine is used to conduct a search using any of those keywords, the advertiser's advertisement will appear on the search results page. Advertisers will pay when a user clicks on the triggered advertisement. Because the keyword advertising is different from traditional banner advertisements which force users to read, it has become a popular way for advertising. Because the cost is lower than traditional online advertisements, it's very suitable to small and medium enterprises. Therefore, how to enhance the benefits of keyword advertising by increasing its click rate is one of important problems. Consequently, this study aims to define the factors of keyword advertising and then use decision trees algorithm to extract latent knowledge for adjusting marketing strategies from the viewpoints of customers.**

Index Terms—**Decision trees, Feature selection, Information gain, Keyword advertising.**

I. INTRODUCTION

Keyword advertisements (Ads) are becoming the dominant advertising way as the keywords could be customized and tailored messages relevant to potential consumers [11]. Keyword advertising is a kind of "sit back and wait" models. Keyword advertisements employed search engines to provide the information needs by internet browsers. In this kind advertising model, the advertiser selects its preferred keywords with or without helps of search engines companies [28]. When the search engine is used to conduct a search using any of those selected keywords, the advertiser's advertisement will appear on the search results page [15]. Because the keyword advertising is different from traditional banner advertisements which force users to read, it has become a popular ways for advertising.

In recent years, researchers have paid much attention on the related issues including keyword pricing and auction [11, 16, 28]. For the Internet advertisement market, there are several ways of pricing online advertisements. Among them, the methods based on cost-per-impression and cost-per-click are the two most popular. The cost-per-click fee is proportional to the click-through rate [16]. If one keyword

advertisement has a high rank in Google search, it needs related keywords, higher click cost, and high click rate [11]. Jansen & Resnick [13] found there have no correlation between the amounts of search queries and the hit rate of online advertisements. In their work, they simulated six situations for internet users to search and purchase products. They indicated over 73% internet users will click the advertisements if they used brand related keywords. Sculley et al. [25] used Google database to predict bounce rate by analyzing different features including keyword, creative, and landing page of advertisements. Although main search engines including Google, Yahoo, and Microsoft have different search algorithms, the rank of search ads sorted according to the popularity of selected keywords and the amount that the advertisers want to pay.

To sum up, keyword advertising related studied focus on the effects of customers' memories and consciousness of specific brands, keyword auction mechanism and the factors of affecting the performances of keywords [12, 15, 18]. But, from available published works, few papers aim to discuss how to increase the click rate of keyword Ads.

Moreover, because the advertising cost is calculated by its click amount, it's cheaper than traditional online advertisements. It's very popular and suitable for small and medium enterprises. Therefore, how to enhance the benefits of keyword advertising by increasing its click rate is one of important problems. Consequently, the objective of this work is to identify the key factors of clicking keyword Ads by using decision tree algorithm. The selected important factors can assist advertisers to make their advertising decisions from the viewpoints of customers. Finally, an actual case study will be provided to demonstrate the effectiveness of the proposed methods.

II. LITERATURE REVIEW

A. Keyword Advertising

In keyword advertising, the advertisers select keywords which they wishes its advertising linked to when a search is conducted using that search engine and nominates the maximum amount it is willing to pay when a user clicks on the triggered advertisements [28]. So, the advertisers want to enhance the click rate by understanding customers' viewpoints. Therefore, this study aims to recognize the crucial factors of clicking the keyword advertisements. But, before identifying these crucial factors, we need to define the potential factors.

To define the factors of keyword advertisements, we focus on the factors which can affect internet users to click keyword

advertisements. The title of keyword Ads consists several or couples of terms. The "keywords" in the title is the key of attracting internet users to click Ads. Moreover, from some researches it'll have different effects even they have the same content but in different order [2].

The design of keyword advertisements such as position and content are also important factors [11, 12, 15, 18, 24]. In the design of search Ads, there are four different factors. They are (1) the selection of relevant keywords, (2) the price of every single one keyword, (3) how to design textual content, and (4) how to design the login page [24]. The content of keyword advertisements could be designed depending on the advertising purposes, target customers, and display styles. It main function is to provide the related information regarding the features and functions of products [21]. Kim et al. [15] attempted to define the key factors of influencing the individual keywords' performances. Their study indicated that the conversion rate and the number of reviews significantly affect to the performance of individual keywords, but the click-through rate does not.

Based on the mentioned above factors, lots of Ads provided by companies or websites appeared in the eye-catching places of the search results pages [20]. Other kinds of Ads which embedded the hyperlinks of commercial websites into the search results achieve the purposes of advertising. For examples, lots of local housing, car rental, and gift stores and restaurants put their links within the Google map. People can see the related information of these stores from the searched map [19]. Recently, some search Ads have been integrated into the content of video or movies in YouTube [9, 17, 29].

The interactivity and multi-characteristics (involving multimedia or sound effects) of search Ads is usually the interesting research issues. But, few studies attempted to define them [30]. According to related works, if the keyword Ads can be appeared in famous search websites, it can have better performances [5, 7]. In previous studies, related researches indicated that internet users could have positive attitudes when they have good images toward a specific website [1, 10, 26].

Some researches focus on keyword auctions and keyword advertisements pricing [3, 12, 18]. For examples, Naldi et al. [18] studied the application of the use of generalized second-price auction for sponsored link slot assignment and pricing. They showed the position of keyword Ads are highly related to the click rate. Huang et al. [12] followed the work of Naldi et al. [18], they explored issues such as the distributional assumptions for the click-through rate, sponsored keyword advertising slot valuation over time, participation and click-stream fees and the role of a two-part tariff, alternative advertiser bidding strategies, and the design goals of sponsored keyword advertising slot auctions.

B. Decision Trees

Decision tree [22] has been utilized to extract related knowledge. Decision tree algorithms are one of the most popular knowledge acquisition methods. They have been successfully applied in many areas. For example, Kim et al. [14] employed decision tree to induce knowledge of personalized advertisements on internet storefronts.

Let's briefly introduce decision tree algorithms. Quinlan [22] used information gain to select attributes from a given set to separate the training examples. Briefly speaking, there are three main phases to rule induction. First, create an initial, large rule tree from the training examples based on attribute selection measures (for example, information gain). Second, prune this tree to remove branches with little statistical validity. Finally, process the pruned tree to improve its understandability. The generalization ability of decision trees can enhance through previous two phases even if there may be errors in classifications of the training examples and errors in the attribute values that describe these examples. The ID3 [22] and C4.5 algorithms [23] are the primary foci of research in the field of decision tree methods. This work utilized C5.0 algorithm to discover knowledge from collected examples.

A common understanding is that decision trees have built-in feature selection functions [23]. Feature selection can also be viewed as choosing a subset of features that achieves the highest classification accuracy [4]. Generally speaking, when decision tree algorithms are used for feature selection, a tree is constructed from the given data. Then, all attributes that do appear in the tree are assumed to be important. Recently, decision tree based feature selection methods have lots of successful applications. For instances, Chen et al. [4] presented a support vector machines (SVM) based model which integrates decision tree feature selection method to analyze the failures of turbines in thermal power facilities. Dębska and Świder [8] utilized a decision tree to select key factors of food quality. Cho and Kurup [6] employed decision tree to reduce dimension space to improve the classification performance of electronic nose. Sindhu et al. [27] developed a decision tree based light weight intrusion detection using a wrapper approach. Therefore, decision tree based feature selection methods have been employed to implement feature selection task in this work.

III. IMPLEMENTAL PROCEDURE

The implemental procedure involves 5 major steps. They are to define service factors of keyword advertising, design questionnaire, collect data, build decision tree, and make conclusions.

The details of the procedure have been provided as follows.
Step 1: Define factors of keyword advertising
In this step, we try to define quality factors of keyword advertisements depending upon available literatures. We survey published works and attempt to define possible quality factors from related studies. Next, according to these defined factors, we can go to next step to design questionnaire for building decision trees. Take the factor "title" as example. It could be defined as "the title of a keyword Ad is short, clear, attractive and interesting."
Step 2: Design questionnaire
In this step, an importance level of the defined factors is measured in this questionnaire for constructing decision trees. In this questionnaire, customers express the feeling about the level of importance for a certain factor regarding the probability of clicking a keyword Ad which fulfills this factor. For instance, according to the definition of "title" mentioned above, we can create one question items as follows.

Q: How would you feel the importance of the title of keyword advertisement?
(A) Very important (B) Important (C) Neutral
(D) Unimportant (E) Very unimportant

Step 3: Collect data

Before issuing, the designed questionnaire will be given to a small group of internet users for pilot testing. Based on their responses, we can slightly modify the questionnaire. Then, we attempt to collect some representative samples from the internet. Those internet users who have experiences of using keyword advertisements will be requested to respond our questionnaire.

Step 4: Build decision tree

In this step, we attempt to extract useful knowledge from the collected data. We use the data of the importance items as our input data (attributes) and the label "the probability of clicking a keyword Ad which fulfills the important factors" as our output data. Then, we can construct decision trees and compute information gain for selecting the important factors. The detailed implemental process could be found as following sub-steps.

Step 4.1 Data preparation

Use 5-fold cross validation experiment and construct a decision tree for each fold of data. In other words, the collected data set will be divided into 5 equal sized sets and each set is then in turn used as the test set. Beside test set, we use other 4 sets as our training set to build decision trees. Therefore, we will have 5 trees.

Step 4.2 Determine the input and output variables

Step 4.3 Construct decision trees following C5.0 algorithm for each fold data set

Step 4.3.1 Create an initial rule tree
Step 4.3.2 Prune this tree
Step 4.3.3 Process the pruned tree to improve its understandability

Step 4.4 Pick a tree whose performance is the best among all constructed trees

Step 5: Draw a conclusion

Finally, from the results of step 4, we find the important factors of keyword advertising, understand the internet users' thinks, and then we can draw conclusions based on them.

IV. RESULTS

A. Defined Candidate Factors of Keyword Advertising

After reviewing couples of related works, we summarize some potential quality factors in literatures and then define 10 quality factors for further analysis. Table 1 gives a summary of these factors, their definitions and some supported references. In available works, the potential quality factors might be not directly related to keyword advertising. Some factors are significant for other kinds of online advertisements. But, all potential factors have been included in this work. And, our method can identify which factors are important from customers' respective.

B. The Collected Data

We collected data from January to July, 2013. 150 online questionnaires have been returned. Among these collected

TABLE I
THE DEFINED POTENTIAL FACTORS

No.	Factor	Definitions	Supports
1	Title	The title of a keyword Ad is short, clear, attractive and interesting.	[2]
2	Content	The content of a keyword Ad is designed to clearly illustrate the features and functions of products.	[2, 21]
3	Famous search engines	The keyword Ads are appeared on famous search websites.	[5, 7, 11]
4	Interactivity	The keyword Ads combined with other multimedia such as flash animation to increase their interactivities.	[30]
5	Customer attitude	Customers have good images to a specific website such as Google or Bing. It would help customers to have positive attitudes toward the keyword Ads on this website.	[1, 10, 26]
6	Refund	Customers will be refunded some money if they click the keyword Ads.	[24]
7	Association	The keyword Ads use association style. For example, a keyword Ad use a title "irresistible attractiveness" and a picture of woman wore under wear to sell washing powder of under wears, not to sell "under wears" directly.	[5, 7]
8	Location/Position	The keyword Ads appeared in conspicuous locations.	[12, 20]
9	Combined with Google Maps	The keyword Ads appeared within the searched maps.	[9, 19]
10	Combined with Video websites	The keyword Ads combined with video websites such as YouTube.	[17, 29]

examples, merely 102 respondents of them have the experiences of experiencing keyword advertisements. Therefore, these examples are considered as valid for further analysis. In the collected samples, the proportion of male and female is 6:4. The range of monthly income for top 2 groups are below 10,000 NTD (New Taiwan Dollar) (48.04%) and 10,000 to 30,000 NTD (33.33%). It refers to the main samples are the youth whose income are not very top in Taiwan society. It also could be found as their education background and age.

Regarding education background, most of respondents are under-graduate educated (82.35%). The age of most samples (77.45%) is below 25 years old, and then 26~35 years old (19.61%). About marriage status, 90.20% samples are single. And, 64.71% of these collected samples use the internet every day. Most of samples spend 1~3 hours (50.00%) and 4~6 hours (36.27%) per day for browsing information or social networking in the internet. From the above mentioned, we can find the collected samples can represent the majority of internet users.

About online shopping frequency in recent half year, less than 5 times (77.45%) is top 1, and then 6~10 times (13.73%). We can find that 3.92% samples bought over 20 times through internet shopping. About the frequency of clicking keyword advertisements in a recent month, the distribution is, less than 3 times (58.82%), 4~10 times (32.35%), and more than 10 times (8.82%). From these numbers, we can understand that the keyword advertisements are very popular, but the internet's users seem not click them very often.

In addition, we also provided a multiple choice question to know the reason why respondents click keyword advertisements. The result shows that 28.43% think the true reason is "the advertisement contains some useful information which can satisfy my needs". But, the same proportion (28.43%) samples respond that they incautiously click them. Other reasons include "the keyword advertisement is an attractive or interesting advertising copy (20.59%)" and "introduced by peers (10.78%)". Therefore, we can find the "desired information", its design style, and the influence of peers, are the major motivations of clicking keyword advertisements.

The proportion of samples who think the keyword advertisements is useful (41.18%) is larger than those who think this kind advertisements are a little useful (14.71%) and not useful (0.98%). So, we can confirm that most samples consider the keyword advertisement is one of effective advertising ways. Finally, most samples (64.71%) think their understanding level regarding keyword advertising is "average".

C. Results of Decision trees

We implement five-fold cross validation experiment to the collected data to build decision trees. In this case, we utilize the "the probability of clicking a keyword Ad when fulfills the important factors" as the dependent variable and 10 factors (the importance measure question items) as the independent variables. Table 2 summarized the classification accuracy of 5 decision trees. Among 5 trees, the performance (80.0%) of fold 2 is the best. Consequently, the tree built from fold#2-experiment has been picked for further analysis.

TABLE II
RESULTS OF FIVE-FOLD EXPERIMENTS

Experiment	Accuracy
Fold #1	65%
Fold #2	80%
Fold #3	65%
Fold #4	66.7%
Fold #5	66.7%

Table 3 listed 9 extracted knowledge rules of the tree with the best performance. In decision trees, all attributes that appear in the tree are assumed to be relevant. From this table, rule #1 and #2 are related to the high probability of clicking a keyword Ad. In these two rules, we find that "Famous search engines (Q3)" and "Title (Q1)" have been recognized as important factors. Moreover, rules #3~#6 are the rules related to the medium probability of clicking a keyword Ad. From these 4 rules, except "Famous search engines (Q3)" and "Title (Q1)", we find that " Customer attitude (Q5)" which means customers have good images to a specific website such as Google or Bing. It would help customers to have positive attitudes toward the keyword Ads on this website. Other rules are related to the low probability of clicking a keyword Ad. We don't discuss them.

V. CONCLUSIONS

In this work, we attempt to define 10 potential factors of keyword advertisements from related works. For all collected data, we utilize decision tree to extract useful knowledge and to identify the crucial factors, respectively, for increasing the

TABLE III
THE EXTRACTED RULES

No.	Rules
1	**IF** Q1 = Very important **AND** Q3 = Important **THEN** The probability of clicking a keyword Ad = **High**
2	**IF** Q3 = Very important **THEN** The probability of clicking a keyword Ad = **High**
3	**IF** Q5 = Neutral **THEN** The probability of clicking a keyword Ad = **Medium**
4	**IF** Q1 = Important **THEN** The probability of clicking a keyword Ad = **Medium**
5	**IF** Q3 = Unimportant **THEN** The probability of clicking a keyword Ad = **Medium**
6	**IF** Q3 = Neutral **THEN** The probability of clicking a keyword Ad = **Medium**
7	**IF** Q1 = Neutral AND Q3 = Important **THEN** The probability of clicking a keyword Ad = **Low**
8	**IF** Q3 = Very unimportant **THEN** The probability of clicking a keyword Ad = **Low**
9	**IF** Q3 = Very important **AND** Q5 = Very important **AND** Q7 = Neutral **AND** Q10 = Neutral **THEN** The probability of clicking a keyword Ad = **Low**

probability of clicking a keyword Ad. Finally, we select 3 important factors. They are "Famous search engines (Q3)", "Title (Q1)", and "Customer attitude (Q5)".

"Famous search engines (Q3)" and "Title (Q1)" have been confirmed as important for the situation which customers can have high probability to click keyword Ads. Therefore, we can conclude that these two factors are the most important if advertisers and advertise agencies want to create a high click rate of a specific keyword advertisement. First, they have to make their keyword Ads appear on famous search websites such as Google, Bing, and so on. Second, they should keep the title of a keyword Ad short, clear, attractive and interesting.

Moreover, we have another one recommendations for advertisers and advertise agencies. They have to make their keyword advertisements appear on websites which customers have good images to these websites. It would help customers to have positive attitude toward the keyword Ads.

There are some possible directions for future works. First, the results of survey come from customers' perspective. The difference between customers and advertisers could be studies further. Second, this study define the factor set regarding the keyword advertisements. It's a hard task. In this work, we tried to define them form available works. Among them, some are not highly correlated to the keyword advertising. But, we include them into our candidate set and then select the important one using our methods. The future works can obtain the candidate set by surveying advertisers, advertise agencies, and customers.

REFERENCES

[1] Bartos, R. and T. Dunn. "Advertising and consumers: New perspectives." New York: American Association of Advertising Agencies, 1974.

[2] Belch, G. E. and M. A. Belch. "Advertising and Promotion." New York: McGraw-Hill, Companies, Inc, 2004.

[3] Chen, J., J. Feng, , and A. B. Whinston. (2010) "Keyword Auctions, Unit-Price Contracts, and the Role of Commitment." Production and Operations Management 19, no. 3 (2010): 305-321.

[4] Chen, K. Y., L. S. Chen, M. C. Chen and C. L. Lee. "Using SVM based Method for Equipment Fault Detection in A Thermal Power Plant." Computers in Industry 62, no. 1, (2011): 42-50.

[5] Chevalier, J. and D. Mayzlin. "The Effect of Word of Mouth on Sales: Online Book Reviews." Journal of Marketing Research 43, no. 3 (2006): 345–354.

[6] Cho, J. H. & Kurup, P. U. "Decision tree approach for classification and dimensionality reduction of electronic nose data." Sensors and Actuators B, 160 (2011): 542-548.

[7] Clemons, E., G. Gao and L. Hitt. "When Online Reviews Meet Hyper Differentiation: A Study of the Craft Beer Industry." Journal of Management Information Systems 23, no. 6, (2006): 149–171.

[8] Debska, B. & Guzowska-´Swider, B. "Decision trees in selection of featured determined food quality." Analytica Chimica Acta 705, no. 1-2 (2011): 261-271.

[9] Egeth, H. E. and S. Yanits. "Visual Attention: Control, Representation, and Time Course." Annual Review of PsyChology 48, no. 1 (1997): 269-927.

[10] Gbadamosi, G. and P. Joubert. "Money Ethic, Moral Conduct and Work Related Attitudes: Field Study from the Public Sector in Swaziland." Journal of Management Development 24, no. 8, (2009): 754 – 763.

[11] Gopal, R. "Online Keyword based Advertising: Impact of Ad Impressions on Own-channel and Cross-channel Click-through rates." Decision Support Systems 52, no.1, (2011): 1-8.

[12] Huang, H. "On the Design of Sponsored Keyword Advertising Slot Auctions: An Analysis of A Generalized Second-price Auction Approach." Electronic Commerce Research and Applications 10, no. 5 (2011): 194-202.

[13] Jansen, B. J. and M. Resnick. "An Examination of Searcher's Perceptions of Nonsponsored and Sponsored Links During Ecommerce Web Searching." Journal of the American Society for Information Science and Technology 57, no. 14 (2006): 1949-1961.

[14] Kim, J., B. Lee, M. Shaw, H. Chang, M. Nelson. "Application of Decision-tree Induction Techniques to Personalized Advertisements on Internet Storefronts." International Journal of Electronic Commerce 5, no. 3 (2001): 45–62.

[15] Kim, C., S. Park, K. Kwon, W. Chang. "How to Select Search Keywords for Online Advertising Depending on Consumer Involvement: An Empirical Investigation." Expert Systems with Applications 39, no. 1 (2012): 594-610.

[16] Moon, Y. and C. Kwon. "Online Advertisement Service Pricing and An Option Contract." Electronic Commerce Research and Applications 10 (2011): 38–48

[17] Mrazek, P., "Creating Brands Online: Third Party Opinions and Their Effect on Consumers' Trust in Brands and Purchase Intentions." Master thesis, Agricultural and Mechanical College, Louisiana State University, 2010.

[18] Naldi, M., G. D'Acquisto and G. F. Italiano. "The Value of Location in Keyword Auctions." Electronic Commerce Research and Applications 9, no. 2 (2010): 160–170.

[19] Park, D., J. Lee and I. Han. "The Effect of On-line Consumer Reviews on Consumer Purchasing Intention: The Moderating Role of Involvement." International Journal of Electronic Commerce 11, no. 2 (2007): 125–148.

[20] Park, D. and J. Lee. "eWOM Overload and Its Effect on Consumer Behavioral Intention Depending on Consumer Involvement." Electronic Commerce Research and Applications 7, no. 4 (2008): 386–398.

[21] Polyorat, K., D. L. Alden and E. S. Kim. "Impact of Narrative versus Factual Print Ad Copy on Product Evaluation: The Mediating Role of Ad Message Involvement." Psychology and Marketing 24, no. 6 (2007): 539-554.

[22] Quinlan, J.R. "Induction of decision trees" Machine Learning, 1 (1986): 81-106.

[23] Quinlan, J. R. "Book Review: C4.5: Programs for Machine Learning," Morgan Kaufmann Publishers 16, no. 1 (1993): 235-240.

[24] Rutz, O. J. and R. E. Bucklin. "A Model of Individual Keyword Performance in Paid Search Advertising." Available at SSRN: http://ssrn.com/abstract=1024765 or http://dx.doi.org/10.2139/ssrn.1024765 (2007).

[25] Sculley, D., R. G. Malkin, S. Basu and R. J. Bayardo. "Predicting Bounce Rates in Sponsored Search Advertisements." In Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, Paris, June 28 - July 01 (2009): 1325-1334.

[26] Schlosser, A. E., S. Shavitt and A. Kanfer. "Survey of Internet Users Attitudes Toward Internet Advertising." Journal of Interactive Marketing 13, no 3 (1999): 34-54.

[27] Sindhu, S., Geetha, S. & Kannan, A. "Decision tree based light weight intrusion detection using a wrapper approach." Expert Systems with Applications, 39, no. 1 (2012): 129-141.

[28] Tyacke, N., R.Higgins, C. Utz. "Searching for Trouble – Keyword Advertising and Trade Mark Infringement." Computer Law & Security Report 20, no. 6 (2004): 453-465.

[29] UWang, J. C. and R. F. Day. "The Effects of Attention Inertia on Advertisements on the WWW." Computers in Human Behavior 36, no. 6 (2007): 1390-1407.

[30] Zaichkowsky, J. L. "Conceptualizing Involvement." Journal of Advertising 15, no. 2 (1986): 4-14.