Constructing Social Intentional Corpora to Predict Click-Through Rate for Search Advertising

Yi-Ting Chen, Hung-Yu Kao
Department of Computer Science and Information Engineering
National Cheng Kung University
P76001221@mail.ncku.edu.tw, hykao@mail.ncku.edu.tw

Abstract

In the beginning, search engines provide placements next to the original search results for advertisers on specific keywords. Since users often search for their interests or purchasing decision, timely presenting proper advertisements to users will encourage them to click on search ads. With the rapid growth of advertising, there is a bidding mechanism that advertisers need to bid keywords on their ads. They should carefully compose keywords in order to enhance the opportunity for their ads to be clicked. Until now, how to efficiently improve the ad performance to earn more clicks remains a main task.

In this paper, we focus on the scope of smart phone and produce a social intentional model with advertising based features to forecast future trend on ads' click-through rate (CTR). In terms of social intentional model, we analyze Chinese text content of technology forum to derive social intentional factors which are Hotness, Sentiment, Promotion, and Event. Our results indicate that with knowing public opinions or occurring events beforehand can efficiently enhance click prediction. This will be very helpful for advertisers on adjusting bidding keywords to improve ad performance via social intention.

Keywords: Advertising, Sponsored Search, Click-Through Rate, Social Intention.

1. Introduction

For online search advertising, the well-known search engines such as Bing, Google, and Yahoo! enable ads to be shown on the top banner or alongside the search results. This generates most of the revenue for search engines. The most common mechanism is cost-per-click (CPC), which means the advertiser bid on keywords but only be charged for each user click on the ad. Both search engines and advertisers look forward to enhancing the ad's click-through rate (CTR), which indicates the probability of the number of ad clicks divided by the number of ad impression. The ad position is on the basis of the ranking score which is computed by the multiplication of CPC and ad quality score. The ad quality depends on plenty of factors that cause an ad to be clicked like ad's keywords, historical CTR, title, description, display URL, landing page, etc. Moreover, CTR is an important and direct metric

for measuring advertised performance.

This paper will focus on forecasting ad keyword's CTR trend, since different bidding keywords in the same ad have various CTR values. Target on the popular 3C products: smart phones, we use the public information from technology forum to predict ups and downs of the next day CTR.

As [1] statistics, the top three most important factors influencing consumer choice of mobile phones are: innovative features, recommendation and price. We extend these criteria as following factors: **Hotness**, **Sentiment**, **Promotion**, and **Event**. All these factors may affect ad's future CTR as Figure 1 depicted. For example the releasing news, a kind of events, may trigger users search on search engine or forum to look for product comments in detail. Users may click more on ads while the ads containing promotion terms or the promotion news is releasing.

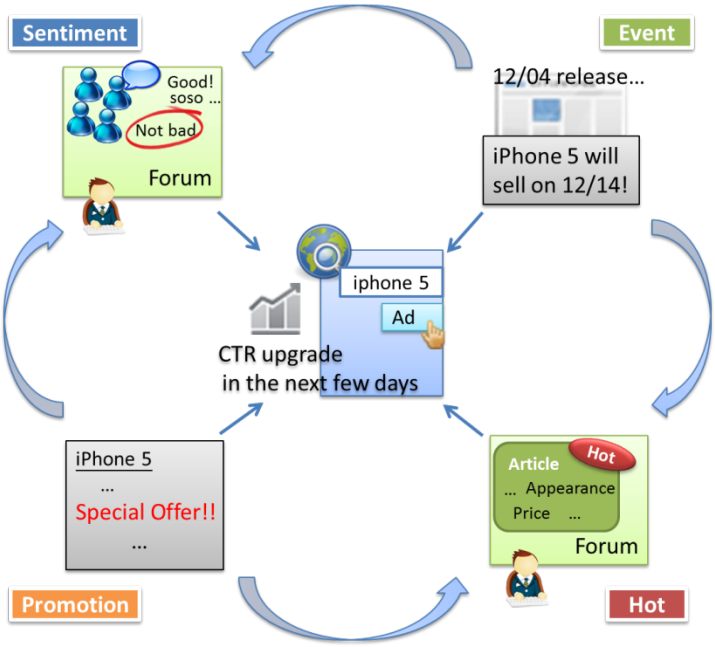

Figure 1. The impact on CTR from releasing news to people reaction

The purpose of this study is to predict and analyze which factors that affect ad keywords' CTR in the next day. This work could previously inform advertisers of user intention on product keywords and assist them to judge whether to change ad strategy or not. It appears that research has not yet been available concerning the effect on search advertising from forum opinions. We expect that this work could significantly aid advertisers in advertising production.

2. Related Work

Even now, there are still lots of research on improving advertising performance in order to verify which features could probably affect ad clicks. We will introduce some related researches that predicted clicks on search advertising; moreover, for the same predicting task but in different research domain, there exist some studies that use public mood to predict the stock trend in the stock market.

2.1 Traditional click prediction problem

Regelson and Fain [2] claimed that historical click information provides tangible examples of user behavior. To predict future click-through rate by term level, for those terms with low frequent or completely novel terms, they use hierarchical clusters of related terms to compute. Apart from terms, Richardson et al. [3] suggested that adding features of ads, and advertisers can accurately predicts the click-through rate for new ads. The collected information of ads contained landing page, bid terms, title, body, display URL, clicks, and impressions (views).

User intentions may significantly vary in the same query. Guo et al. [4] develop a fine-grained user interaction model for inferring searcher receptiveness to advertising. They modified the Firefox version of the OpenSource LibX toolbar to instrument mouse movements and other user action events on search result pages. Cheng and Cantú-Paz [5] develop demographic-based and user-specific features that reflect the click behavior of groups and individuals.

To strengthen the relation between query and ad, Dave and Varma [6] proposed a similarity method to give prediction. Especially for those rare/new ads, they used cosine similarity between two queries or two ads. Xiong et al. [7] designed a continuous conditional random fields (CRF) based model, which considered both features of an ad and its similarity to the surrounding ads.

2.2 Using social media for prediction

The prediction problem on trend is analogous to click prediction. Bollen et al. [8] first used six dimensions of mood(tension, depression, anger, vigor, fatigue, confusion) from Profile of Mood States(POMS), a well-established psychometric instrument to observe the relation between moods and socio-economic phenomena. After that, Bollen et al. [9] expanded terms of POMS from Google webpages, named it GPOMS. GPOMS contained six different mood dimensions: *Calm*, *Alert*, *Sure*, *Vital*, *Kind*, and *Happy*. They used Granger causality analysis to investigate the hypothesis of public mood states and a Self-Organizing Fuzzy Neural Network to predict the daily up and down changes of Dow Jones Industrial Average (DJIA) in the stock market by the OpinionFinder and GPOMS mood time series.

3. Method

In this section, we present our proposed framework as shown in Figure 2 to address the problem of predicting ad keyword CTR via adding social phenomena. In brief, given an ad keyword as an input, our system returns the direction of movement in the next day based on previous advertising data and social intention effects. First, in advertising-based part, we do the CTR filtering to be basic information on an ad keyword. Next, before running the main process, the social intentional factors have been built from historical public behaviors on technology forum. After that, we crawl the related articles on technology forum in recent time duration to calculate social intentional scores. Thus, with these two-part values, we can run the prediction model in the last process. The results are produced from Linear Regression model and SVM classification model.

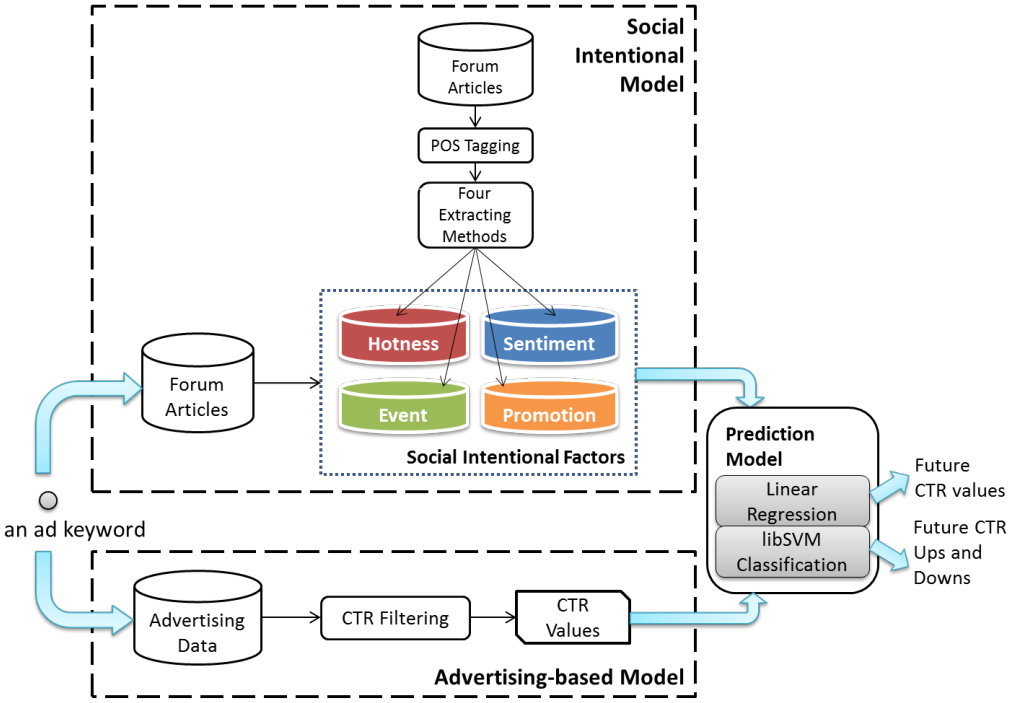

Figure 2. Proposed framework

According to user preference on purchasing, we propose four extracting methods to produce *Hotness*, *Sentiment*, *Promotion*, and *Event* that may be sufficient to affect user click on ads. The data used for methods in this part is Mobile01 articles from November 1, 2012 to January 31, 2013 which contains 21,674 articles. In the following parts, we will introduce these methods with Mobile01 articles in detail.

- Hotness -

The "*Hot*" means feverish, to become lively or exciting[1] that can informal arouse

[1] http://en.wiktionary.org/wiki/hot

intense interest, excitement, or controversy[2]. What we need to do is find out those proper themes that stimulate public to discuss on technology products. Focusing on smart phone in our work, we consider the phrases are broadly and frequently mentioned between articles, such as the phone's appearance, functionality, price, etc. Inverse Document Frequency (IDF) is a measure of whether the term is common or rare across all articles as shown in Eq.(1), where $|D|$ is the number of all articles, and $|\{d_i|d_i \in D\}|$ is the number of articles containing the phrase t_i. We choose the IDF range from 0 to 4 which contains 379 terms to be hot candidates.

$$IDF_i = log \frac{|D|}{|\{d_i|d_i \in D\}|} \quad (1)$$

We randomly pick some terms in IDF of all articles less than 4 and greater than 8 to check what the terms look like and display it in Table 1. The range of IDF less than 4 closely meet our expectation.

Table 1. Terms look like when IDF less than 4 and IDF greater than 8

Terms in IDF < 4	**Terms in IDF > 8**
功能, 蝴蝶, 三星, 智慧型, 蘋果, 品質, 規格, 價錢, 價位, 耗電, 解析度, 畫素, 優勢, 瑕疵, 配件, 廠牌, 費率, …	抗刮性, 輕量版, 超薄超順超, 機王戰, 獨家版, 獨特感, 磨砂款, 機防撥水, 高精度, 超簿, 優質感, 質量感, …

When a hot article comes up, there must be widely discussed and viewed by a crowd of people. Thus we gather the articles having top 1 percent high prestige[3] in each category and obtain 222 of them in all articles. Because hot terms are feverish and most talked-about subjects, we sort these 379 candidate terms by Term Frequency (TF) value in a descending order from all articles. The TF value is calculated by Eq.(2), where $n_{i,j}$ is the number of term t_i appears in article d_j, and $\sum_k n_{k,j}$ is total number of terms in article d_j. It means each candidate terms has 21,674 TF ranking value from all articles. Next, we set a threshold on 222. That is, if top-222 TF article values contain one of hot articles (222 articles), this candidate term will be chosen as hot term.

$$TF(t_{ij}) = \frac{n_{i,j}}{\sum_k n_{k,j}} \quad (2)$$

With hotness lexicon $lexicon_H$, input an ad keyword and a date, we could calculate the

[2] http://www.thefreedictionary.com/hot

[3] "Prestige" here is said the number of views to the article.

hot score from daily articles by Eq.(3), where a_i is one of keyword-related articles from the set $Articles_{k,d}$, $|Articles_{k,d}|$ is the number of keyword-related articles that are crawled in the date time, and $Count(h_j, a_i)$ is the count of hot term h_j appears in article a_i.

$$Score_{Hot,k}(d) = \frac{\sum_{a_i \in Articles_{k,d}} \sum_{h_j \in lexicon_H} Count(h_j, a_i)}{|Articles_{k,d}|} \quad (3)$$

- Sentiment -

In this part, we want to analyze public moods and opinions for a product from articles. The first step is to build a sentiment lexicon. We utilize NTUSD[10] and HOWNet[4] to obtain 4140 positive terms and 6608 negative terms with no repeats as our sentiment lexicon $lexicon_s$. Although the number of negative terms is more than positive terms used, it does not affect the orientation of public opinions.

The sentiment score for an ad keyword with a date is calculated by Eq.(4), where $Score(s_j) = +1$ if s_j is a positive term, otherwise is -1, and $Count(s_j, a_i)$ is the count of sentiment term s_j appears in article a_i.

$$Score_{Senti,k}(d) = \frac{\sum_{a_i \in Articles_{k,d}} \sum_{s_j \in lexicon_s} Score(s_j) * Count(s_j, a_i)}{|Articles_{k,d}|} \quad (4)$$

- Promotion -

Everyone knows that selling products with discount phrases is noteworthy to public. At first we pick 15 terms that contain promotional meaning to be seed words. They are 特價 (Special offer), 降價(Price reduction), 優惠(Preferential), 特賣(Clearance), 特惠(Specials), 福袋(Lucky bag), 抽獎(Lottery), 折扣(Discount), 獨享(Exclusive), 好康(Good things), 下殺(an auxiliary verb for discount in Chinese), 免費(Free), 放送(Gift), 便宜(Cheap), and 划算(Saving). To build a lexicon on promotion, we expand these terms by analyzing word co-occurrences in front and rear 5-term collections by Yahoo! top-200 query results in a past year.

For calculating promotion score, we produce a formula in Eq.(5), where $Count(p_j, a_i)$ is the count of promotion term p_j appears in article a_i.

$$Score_{Promote,k}(d) = \frac{\sum_{a_i \in Articles_{k,d}} \sum_{p_j \in lexicon_P} Count(p_j, a_i)}{|Articles_{k,d}|} \quad (5)$$

[4] http://www.keenage.com/html/c_index.html

- Event -

We have observed that news or events may affect ad keyword's CTR in the next few days. Thus we propose the number of bursty replies on forum articles to model an event effect in a numerical manner. By using Eq.(6), where t_{a_i} is the post time of the article a_i, t_d is the time duration we set to a half-day, and $RC(a_i, t_{a_i}, t_d)$ is the reply counts based on two former parameters, our event score is produced.

$$Score_{Event,k}(d) = \frac{\sum_{a_i \in Articles_{k,d}} RC(a_i, t_{a_i}, t_d)}{|Articles_{k,d}|} \quad (6)$$

3.3 Advertising-based model

Usually, advertisers would combine the product name with some terms like: 價格 (price), 便宜(cheap) to be a bidding keyword. Hence if we wonder to look for the specific keyword's data on certain day, all kinds of keyword combination should be taken into for consideration. Table 2 displays a part of bidding keywords on "iPhone 5".

Table 2. Bidding Keywords on "iPhone 5"

apple iphone 5 16G 評價, apple iphone 5 功能, Apple iphone 5 哪裡買,
Apple iphone 5 售價, apple iphone 5 發表, Apple iphone 5 開箱,
iphone 5 價格, iphone 5 規格, …

Thus, for those keyword-related ads that are crawled in the date time, we define them as $\mathcal{AD}_{k,d} = \{ad_1, ad_2, \ldots, ad_n\}$. For those bidding keywords from the keyword-related ad on the certain day are presented as $\mathcal{B}_{\mathrm{ad_j}} = \{k_1, k_2, \ldots, k_m\}$. $CTR(k_i)$ is the click-through rate of the bidding keyword k_i in the ad ad_j. With these advertisements and bidding keywords, we could compute CTR value for the objective keyword on certain day as follows:

$$CTR_k(d) = \frac{\sum_{ad_j \in \mathcal{AD}_{k,d}} \frac{\sum_{k_i \in \mathcal{B}_{ad_j}} CTR(k_i)}{|\mathcal{B}_{ad_j}|}}{|\mathcal{AD}_{k,d}|} \quad (7)$$

4. Experiments

4.1 Dataset and preprocessing

Our dataset of technology forum is from Mobile01.com[5] which is an Internet forum being devoted to discussing a variety of mobile phones, mobile devices, 3C products, etc. We crawl 4 months data from November 1, 2012 to February 28, 2013 with twelve categories. The information we extract from forum articles includes 15 available attributes. The ultimate decision on attributes using are Category, Prestige, Title, Replies, Post Date, and Post Content.

WIS Internet Inc.[6] is currently a Yahoo! Taiwan Search Marketing Ambassador. It is thanks to WIS assist in providing advertising data to us that our research is getting more credibility. The duration of advertising data is 3 months from December 1, 2012 to February 28, 2013. Since our study is focused on smart phone, the dataset consists of 10 related advertisers, 2,283 ads and 14,537 ad keywords. The information we use to experiment are Advertiser ID, Ad ID, Date, Keyword, Ad Group, Ad Campaign, Impressions, Click-Through Rate, Clicks, and Keyword average Ranking.

Before we do our experiments, we preprocess our dataset in advance. We use CKIP to split Chinese phrases from content of articles and obtain POS tags. The distribution of the number of articles and replies in training and testing data are shown in Table 3.

Table 3. Data statistics in training and testing

Item	In training	In testing
Date	Dec.1, 2012~Feb.14, 2013	Feb.15~Feb.28, 2013
# of categories	12	12
# of articles	18,125	2,984
# of replies	187,821	35,353

4.2 Results and discussion

In order to evaluate the performance of our system and to compare with the baseline, forecasting CTR value and CTR up or down prediction is measured in terms of the Average Mean Absolute Error (MAPE) and the direction accuracy. Based on the CTR values produced from advertising model, we add keyword's daily average position as our baseline to strengthen the predicting capability.

In Figure 3, we observe that for using previous 4 days data, some of factors predict well

[5] http://www.mobile01.com/

[6] http://www.wis.com.tw/eng.html

than baseline but not all of them do. Since each factor has its characteristic which are demonstrated from daily forum articles. With using more previous data, the increasing reference data can aid the prediction.

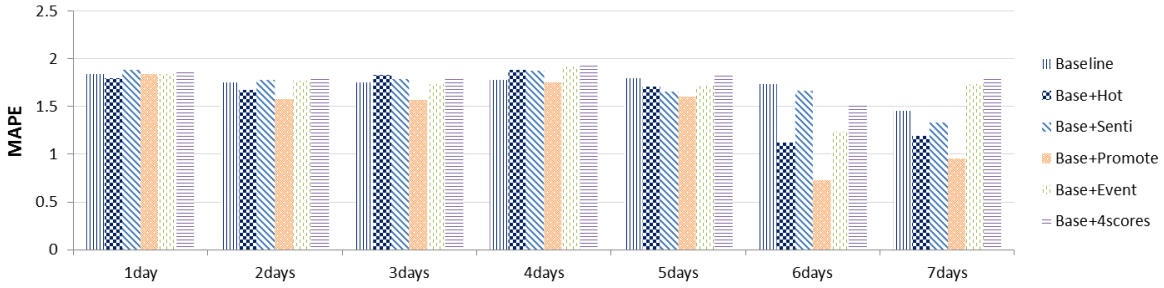

Figure 3. CTR daily prediction with different social intentional factors

For ups and downs prediction, Figure 4 illustrates that for using previous 2 days data, adding *Sentiment* information has an outstanding performance. Besides, the overall conditions for using previous 6 or 7 days data have better prediction. We observe that only using advertising data may not enough to predict future CTR trend. However, with our social intentional methods, the prediction will more accurate and each social intentional factors are more significant in different previous days usage.

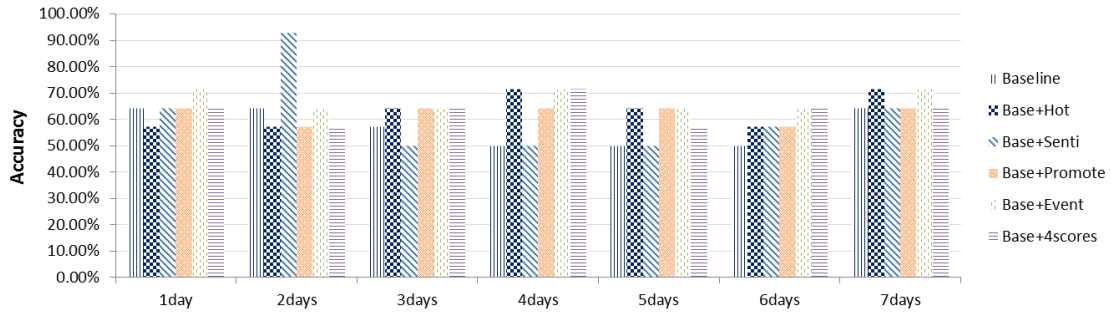

Figure 4. CTR ups and downs prediction with different social intentional factors

5. Conclusion

In this paper, we propose the social intentional methods derived from a popular technology forum to forecast CTR trend on Chinese search advertising. Differs from traditional advertisement click or not prediction, our model focuses on the specific ad keyword and predicts its next day CTR value and direction of movement. In particular, we construct three corpora and one factor from forum to represent public perspectives on mobile phones. Based on these corpora, we can find which terms are most discussed by people in Hotness, or which terms are probably attractive to people in Promotion, etc. Our results present that social intention will affect an ad keyword's future CTR soon or delayed a few days. The reason is that people may discuss or read experiential articles on forum before searching or purchasing on search engine. With public disposition and market tendency, we

can precisely indicate which factors influence the specific ad keyword the most in recent days. This approach is very helpful to advertisers who want to publish a new ad or adjust the keywords of the ad. Furthermore, our proposed method can not only use in the scope of mobile phones but also expand to other marketing fields like brand analysis, beauty makeup, or clothes.

References

[1] S. Y. Mokhlis, Azizul Yadi, "Consumer Choice Criteria in Mobile Phone Selection: An Investigation of Malaysian University Students," *International Review of Social Sciences & Humanities,* vol. 2, pp. 203-212, 2012.

[2] M. Regelson and D. C. Fain, "Predicting Click-Through Rate Using Keyword Clusters," presented at the 06th ACM Conference on Electronic Commerce, 2006.

[3] M. Richardson, E. Dominowska, and R. Ragno, "Predicting clicks: Estimating the click-through rate for new ads," in *In Proceedings of the 16th International World Wide Web Conference (WWW-07)*, ed: ACM Press, 2007, pp. 521-530.

[4] Q. Guo, E. Agichtein, C. L. A. Clarke, and A. Ashkan, "In the Mood to Click? Towards Inferring Receptiveness to Search Advertising," presented at the Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology - Volume 01, 2009.

[5] H. Cheng and E. Cantú-Paz, "Personalized click prediction in sponsored search," presented at the Proceedings of the third ACM international conference on Web search and data mining, New York, New York, USA, 2010.

[6] K. S. Dave and V. Varma, "Learning the click-through rate for rare/new ads from similar ads," presented at the Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval, Geneva, Switzerland, 2010.

[7] C. Xiong, T. Wang, W. Ding, Y. Shen, and T.-Y. Liu, "Relational click prediction for sponsored search," presented at the Proceedings of the fifth ACM international conference on Web search and data mining, Seattle, Washington, USA, 2012.

[8] J. Bollen, A. Pepe, and H. Mao, "Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena," *CoRR,* vol. abs/0911.1583, 2009.

[9] J. Bollen, H. Mao, and X.-J. Zeng, "Twitter mood predicts the stock market," *Journal of Computational Science,* vol. Volume 2, pp. Pages 1-8, March 2011 2011.

[10] L.-W. Ku and H.-H. Chen, "Mining opinions from the Web: Beyond relevance retrieval," *J. Am. Soc. Inf. Sci. Technol.,* vol. 58, pp. 1838-1850, 2007.