Modeling Click and Relevance Relationship for Sponsored Search

Wei Vivian Zhang, Ye Chen, Mitali Gupta, Swaraj Sett, Tak W. Yan
Microsoft Corporation
1020 Enterprise Way
Sunnyvale, CA 94089
{wzha, yec, mitalig, swasett, takyan}@microsoft.com

ABSTRACT

Click-through rate (CTR) prediction and relevance ranking are two fundamental problems in web advertising. In this study, we address the problem of modeling the relationship between CTR and relevance for sponsored search. We used normalized relevance scores comparable across all queries to represent relevance when modeling with CTR, instead of directly using human judgment labels or relevance scores valid only within same query. We classified clicks by identifying their relevance quality using dwell time and session information, and compared all clicks versus selective clicks effects when modeling relevance.

Our results showed that the cleaned click signal outperforms raw click signal and others we explored, in terms of relevance score fitting. The cleaned clicks include clicks with dwell time greater than 5 seconds and last clicks in session. Besides traditional thoughts that there is no linear relation between click and relevance, we showed that the cleaned click based CTR can be fitted well with the normalized relevance scores using a quadratic regression model. This relevance-click model could help to train ranking models using processed click feedback to complement expensive human editorial relevance labels, or better leverage relevance signals in CTR prediction.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Relevance feedback, Retrieval models, Search process

General Terms

Algorithms, Experimentation.

Keywords

Ad relevance; ad click; click-through rate; sponsored search

1. INTRODUCTION

There are two important modeling tasks in web search and advertising: the ranking model used to order results based on relevance, and the click model used to optimize user engagement or revenue based on user click-through behavior. The relevance ranking model is typically learned from human editorial judgment labels, while user click optimization is mined from click logs.

One key challenge is that human relevance labels are not always consistent with click signals [5]. Much work has attempted

to bridge this gap. The cascade, DBN and PCC click models [6] aimed to improve document relevance. R. Agrawal et al. [1] proposed to generate ranking model relevance training labels from clicks. In this paper, we study the relationship between clicks and relevance in the sponsored search domain. Our results show that clicks with dwell time less than 5 seconds barely contribute to relevance prediction. We presented the well-fitted quadratic regression model. It uses the normalized relevance score trained with human editorial relevance labels as the independent variable. It uses the position bias removed CTR as the dependent variable. The model using cleaned click signal (clicks with dwell time greater than 5 seconds and last clicks in user search sessions) outperforms other five proceeded click alternatives we explored.

2. RELEVANCE SCORE AND CLICK SIGNAL GENERATION

2.1 Normalized Relevance Score Generation

We trained a McRank model [3] to generate relevance scores for given (query, ads listing) pairs. McRank is a tree boosting model that casts ranking as a multiple classification task. Other than its state-of-the-art learning-to-rank performance, McRank naturally produces $[0, 1]$ normalized relevance scores that are comparable across different queries, unlike other popular learning-to-rank models. This is necessary for the purpose of studying the relationship between relevance and click signals across all queries. Training data were sampled from search logs. Classical learning-to-rank features were used as the predictors [4]. A four-point scale of human editorial judgments on (query, ad listing) pairs was used as training labels (excellent, good, fair and bad).

2.2 Click Collection and Processing

We ran an A/B testing flight to collect user queries, ad listings and clicks. For the collected (query, ad listing) pairs, we used the McRank model trained in Section 2.1 to generate their normalized relevance scores. We then fetched their clicks from the click logs, as well as user post-click behaviors for click signal processing. All clicks have been normalized to remove positional bias as follows. Positional priors $p(click|position)$ were first generated through a randomized flight. In this flight, the relevance-wise pre-qualified ad listings were randomized for their display positions in search result page to remove positional bias, while not hurting user experience much. We calculated the positional prior as the expected CTR for each position. An in-log click is then normalized by the inverse of its corresponding positional prior [2]. Intuitively, a click from a less prominent position shall cast more feedback than the one from a better position.

We experimented with six click signal alternatives as summa-

Table 1: Click signals, their volume ratio and R squared measure of goodness of fit on relevance score

| Click Type	Definition	Percentage of all clicks	R-square
AllClicks	All position-normalized clicks	100%	0.9584
CleanedClicks	Position-normalized clicks with dwell time $\geq$ 5 seconds or last clicks	96.2%	0.9595
FilteredClicks	Position-normalized clicks with dwell time $\geq$ 10 seconds or last clicks	89.9%	0.9584
GoodClicks	Position-normalized clicks with dwell time $\geq$ 20 seconds or last clicks	77.3%	0.9576
ExcellentClicks	Position-normalized clicks with dwell time $\geq$ 30 seconds or last clicks	68.1%	0.9542
LastClicksOnly	Position-normalized last clicks in user search session	17.0%	0.9093
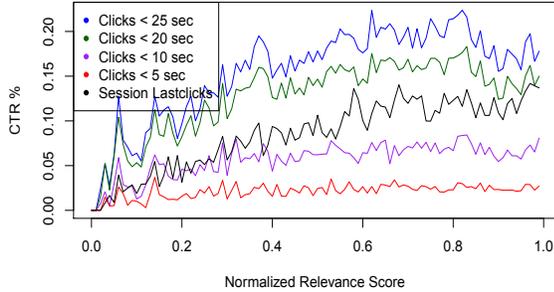
Figure 1: CTR using different click signals vs. relevance score
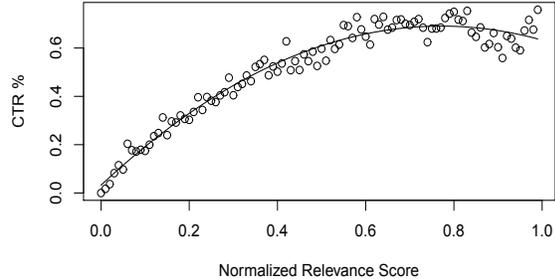

Figure 2: The quadratic model fitted using cleaned clicks

rized in Table 1. The goal is to select the most informative click set with appropriate post-processing to best reflect the relevance for (query, ad listing) pairs.

3. EXPERIMENTAL RESULTS

3.1 Identify Low Quality Clicks

Many clicks occurred on irrelevant documents. They are noises in relevance modeling. Figure 1 shows that clicks with dwell time less than 5 seconds barely have any prediction power for relevance. Their CTR remains flat when relevance score increases, while a good predictor should have significant increasing pattern. Clicks with dwell time $\geq$ 5 seconds and last clicks in session capture relevance signals well. They are important clicks to predict relevance. We suggest removing clicks with dwell time < 5 seconds for click noise reduction. Table 1 shows this removal results in only 3.8% click loss, which is affordable considering nearly unlimited click log sources. Notice that the CTR dropping with dwell time maximum threshold decreasing in Figure 1 is expected because we were experimenting with removing noise clicks.

3.2 Model CTR and Relevance Relationship

We aim to use the simplest model to capture the relationship between CTR and relevance when possible. The kth order polynomial regression model can be expressed as

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_k x^k + \epsilon \tag{1}$$

Where x is the normalized relevance score, y is the CTR, and β_k is the kth order coefficient to be estimated. We collected 16 million (query, ad listing) pairs with their relevance scores generated, and 85 thousand clicks using the approach described in Section 2. This dataset was fitted into the kth order polynomial regression models with $k = 1, 2, 3 \ldots$. Their R-squared measures using the cleaned-clicks (clicks with dwell time < 5 removed) are listed in Table 2.

As shown in Table 2, R-square is significantly improved from the linear regression model to quadratic model, while stays the same from quadratic to the 3rd order polynomial. Therefore, we will use the quadratic regression model as follows, with its fitting curve illustrated in Figure 2.

$$y = 0.032 + 1.708x - 1.108x^2 \tag{2}$$

Table 2: The measure of goodness of fits

	$k=1$	$k=2$	$k=3$
R-square	0.7873	0.9595	0.9595

Interestingly, the model trained using cleaned clicks outperformed other click signals listed in Table 1. This is consistent with our observation in Section 3.1 and shows noise reduction does help.

4. CONCLUSIONS AND CONTRIBUTIONS

In this study, we investigated the relationship between CTR and (query, ad listing) relevance. We used normalized relevance score as the predictor, to allow relevance comparison across different queries. We experimented with different click post-processing approaches using user post-click behavior and session information. We fit a quadratic model to formally capture the relationship between CTR and the relevance score.

Our major contributions include: (1) we have established the relationship between relevance and clicks in sponsored search via a non-linear model; (2) the formal relevance-click model provides foundation for leveraging click feedback in relevance ranking, and vice versa; and (3) the procedures of normalizing relevance scores and qualifying clicks will be instrumental in practice.

5. REFERENCES

[1] R. Agrawal, A. Halverson, K. Kenthapadi, N. Mishra, and P. Tsaparas. Generating labels from clicks. In *WSDM*, 2009.

[2] Y. Chen and T. W. Yan. Position-normalized click prediction in search advertising. In *KDD*, pages 795–803, 2012.

[3] P. Li, C. Burges, , and Q. Wu. Mcrank: Learning to rank using multiple classification and gradient boosting. In *NIPS*, 2007.

[4] T. Qin, T.-Y. Liu, J. Xu, and H. Li. Letor: A benchmark collection for research on learning to rank for information retrieval. *Journal of Information Retrieval*, 2010.

[5] W. V. Zhang and R. Jones. Comparing click logs and editorial labels for training query rewriting. In *WWW 2007 Workshop on Query Log Analysis*, 2007.

[6] F. Zhong, D. Wang, G. Wang, W. Chen, Y. Zhang, Z. Chen, and H. Wang. Incorporating post-click behaviors into a click model. In *SIGIR*, 2010.