New Features for Query Dependent Sponsored Search Click Prediction

Ilya Trofimov
Yandex
16 Leo Tolstoy St.
Moscow, Russia
trofim@yandex-team.ru

ABSTRACT

Click prediction for sponsored search is an important problem for commercial search engines. Good click prediction algorithm greatly affects on the revenue of the search engine, user experience and brings more clicks to landing pages of advertisers. This paper presents new query-dependent features for the click prediction algorithm based on treating query and advertisement as bags of words. New features can improve prediction accuracy both for ads having many and few views.

Categories and Subject Descriptors

H.3.3 [**Informational Storage and Retrieval**]: Information Search and Retrieval; I.2.6 [**Artificial Intelligence**]: Learning

General Terms

Algorithms, Measurement, Performance

Keywords

Click prediction, sponsored search, web advertising

1. INTRODUCTION

Most major search engines today present two types of results: *organic search results*, the short snippets of text with links to relevant web pages, and *sponsored search results*, the small textual advertisements, displayed close by the organic results. Search engine Yandex displays most of ads on the right hand side on the page (*right ad placement*), and the most profitable and relevant ads are placed straight above the organic results (*top ad placement*).

Search engine returns the most relevant search results trying to give an answer to a user's query. That is why organic results are ranked solely based on relevance. In the same time, sponsored search results are the major source of the search engine revenue. In the most common pay-per-click model an advertiser is charged each time as the ad being clicked by a user. Ad's CTR multiplied by the bid is recognized as the revenue estimate. That is why click-through rate prediction is of special importance. We recommend lectures on computational advertisement [1] for more detailed survey of on-line ads problems.

WWW 2013 Companion, May 13–17, 2013, Rio de Janeiro, Brazil.
ACM 978-1-4503-2038-2/13/05.

Sponsored search is based on the keyword auction. The keyword auction is described in [2]: advertiser bids on the selected set of keywords, describing the product or the service. When a user types a query, search engine matches it with all keywords and selects appropriate ads to display. There are various types of matches, namely, similarities between a query and a keyword: exact match, phrase match, broad match. In the case of the exact match, the ad is eligible to appear only when the user's query is identical to the keyword. In case of the phrase match, the keyword must be a subset of the query. Finally, the broad match allows the ad to appear when the query coincides with a keyword automatically chosen by the advertising system.

Because sponsored search ads are only textual, text driven features are an important source of information for click prediction. Search engines often show each ad for many different queries and a click-through rate can vary significantly across them. But for most of ad-query pairs historical click data is insufficient and click-though rate can't be estimated directly. In this case statistical model can solve this problem. This short paper makes following contributions:

- we present a new set of binary features based on conjunctions between query and ad's creative;
- we study previous users's query in a search session as a source of useful information;
- we adopt hashing for handling large feature space.

2. PROPOSED MODEL

Suppose we have already some formula for click probability $CTR_{base}(z)$, which uses historical click statistics. Click statistics include $CTR = clicks/views$ for several types of objects with many views, like ad and advertiser's domain. The base formula is described in [6]. Let's build a logistic model for click prediction which *adjusts* the base formula

$$P(click|x, z) = f(af^{-1}(CTR_{base}(z)) + b + \sum_{i=1}^{n} w_i x_i) \quad (1)$$

$$f(t) = \exp(t)/(1 + \exp(t))$$

Click prediction formula 1 is optimized for maximum likelihood estimation with respect to the variables a, b, w_i. In this formula x_i are binary text-based features. Each feature is non-zero if some word is present in some bag of words, derived from the ad's creative and the query. It is important to distinguish query-dependent and query-independent

Table 1: Relative quality improvement for different features set

Features	ΔNLL	ΔauPRC
Baseline	0.0%	0.0%
Shaparenko et al. [5]	-0.6%	+2.6%
The proposed model	-1.2%	+6.7%

features. A feature x_i is query independent if it is a function only of the advertisment's creative and the keyword. Non-zero corresponding weight w_i will adjust the base formula CTR_{base} both for ads having many views and few views. This is inefficient, because for ads having many views CTR_{base} is already a good estimate of average click-through rate. This idea leads to the explicit dividing a query into two parts: a keyword (which is always a subset of a query) and a rest part.

2.1 Text sources of an ad

The following table presents different sources of a text in the ad.

Table 2: Bags of words from ad's text sources

Bag of words	Definition	Query-depend.
K	$v \in keyword$	No
T	$v \in title$	No
D	$v \in description$	No
QK	$v \in query$ & $v \notin keyword$	Yes

We consider presence of a given word v in a bag as a binary feature, e.g. $x_i = 1 \Leftrightarrow (v_i \in query \ \& \ v_i \notin keyword)$. The same word present in different bags yields different features. First three bags of words are query-independent, others are query-dependent. The proposed model also contains all conjunctions between query-dependent and query-independent binary features from different bags, e.g.
$x_k = 1 \Leftrightarrow (v_n \in title \ \& \ (w_m \in query \ \& \ w_m \notin keyword))$.

2.2 Previous query

Similarity between two successive queries in a search session clarifies user's information need and can help to build more accurate prediction. We model this similarity by means of the binary features, which are shown in table 3. Q denotes query words set, R - previous query words set.

Table 3: Similarity features

Feature	Condition
IsMain	$R = http://www.yandex.ru$
IsEmpty	R *is empty*
InRefresh	$Q = R, page\ number = prev.\ page\ number$
InNext	$Q = R, page\ number = prev.\ page\ number + 1$
IsSpec	$R \subset Q$ (*specification*)
IsGen	$R \supset Q$ (*generalization*)
IsDisjoint	$Q \cap R = \varnothing$ (*disjoint queries*)
IsOther	*All previous conditions are false*

For several similarity types between two successive queries we build the following bags of words which are shown in table 4. Again we consider a presence of some word in a bag as a binary feature.

3. RESULTS

The training and testing data sets are the top ad placement logs with users' clicks from the Yandex search engine. Training set contains 215×10^6 examples randomly sampled

Table 4: Bags of words yielded by successive queries

Bag of words	Definition
Q1	IsMain & $v \in query$
PQ1	IsSpec & $v \in prev.\ query$
PQ2	IsSpec & $v \in query$ & $v \notin prev.\ query$
PQ3	IsOther & $v \in prev.\ query$

from the one week period. Testing set contains 28×10^6 examples randomly sampled from one day after training set. Basic text normalization like lemmatization and punctuation removal is performed for the both sets. For fitting logistic regression we used free Vowpal Wabbit software [4]. It uses hash representation [3] of all features with hash table having size 2^{24}. Thus our model handles approximately 17 millions of features. Table 1 presents quality measures at the testing set. Baseline quality is a quality of the best logistic transformation of $CTR_{base}(z)$:

$$P(click|z) = f(af^{-1}(CTR_{base}(z)) + b)$$

Our set of features has two main differences from the features described in [5]. First, we build different conjunctions between the query and the ad's text sources. Second, we use information yielded by the previous query.

4. CONCLUSIONS

Click-through rate estimation plays a crucial role for ads selection. It greatly affect the search engine revenue, traffic received by advertisers' landing pages and user experience. In this short paper we described a new set of query-dependent text features. We showed that these features outperform state-of-art ones. Further research can include developing a more accurate way of mixing binary text-based features with baseline click prediction formula and studying bigrams. It is important to add these features to the production click prediction system. Live traffic experiment will show how new features improve observed advertising system metrics like average CTR, ads relevance and revenue.

5. REFERENCES

[1] A. Broder and V. Josifovski. *Introduction to Computational Advertising.* http://www.stanford.edu/class/msande239/.

[2] B. Edelman, M. Ostrovsky, and M. Schwarz. Internet advertising and the generalized second price auction: Selling billions of dollars worth of keywords. *American Economic Review*, 97(1):242–259, 2007.

[3] A. S. Kilian Weinberger, Anirban Dasgupta, Josh Attenberg, John Langford. Feature Hashing for Large Scale Multitask Learning. In *ICML*, 2009.

[4] J. Langford. Vowpal wabbit. Technical report, http://hunch.net/ vw, 2007-2012., 2007.

[5] B. Shaparenko, O. Çetin, and R. Iyer. Data-driven text features for sponsored search click prediction. In *KDD, ADKDD Workshop*, pages 46–54, New York, New York, USA, 2009. ACM Press.

[6] I. Trofimov, A. Kornetova, and V. Topinskiy. Using boosted trees for click-through rate prediction for sponsored search. In *KDD, ADKDD Workshop*, Beijing, China, 2012. ACM Press.