Learning the Click-Through Rate for Rare/New Ads from Similar Ads

by

Kushal Dave, Vasudeva Varma

in

33rd Annual ACM SIGIR Conference

Geneva, Switzerland

Report No: IIIT/TR/2010/21

Centre for Search and Information Extraction Lab
International Institute of Information Technology
Hyderabad - 500 032, INDIA
July 2010

Learning the Click-Through Rate for Rare/New Ads from Similar Ads

Kushal Dave
Language Technologies Research Centre
International Institute of Information Technology
Hyderabad, India
kushal.dave@research.iiit.ac.in

Vasudeva Varma
Language Technologies Research Centre
International Institute of Information Technology
Hyderabad, India
vv@iiit.ac.in

ABSTRACT

Ads on the search engine (SE) are generally ranked based on their Click-through rates (CTR). Hence, accurately predicting the CTR of an ad is of paramount importance for maximizing the SE's revenue. We present a model that inherits the click information of rare/new ads from other semantically related ads. The semantic features are derived from the query ad click-through graphs and advertisers account information. We show that the model learned using these features give a very good prediction for the CTR values.

Categories and Subject Descriptors

I.2.6 [**Computing Methodologies**]: Artificial Intelligence—*Learning*; I.6.5 [**Computing Methodologies**]: Simulation and Modeling—*model development*; H.3.3 [**Information Systems**]: Information Storage and Retrieval

General Terms

Algorithms, Economics, Experimentation.

Keywords

Sponsored Search, Click-Through Rate Prediction, Ranking

1. INTRODUCTION

Sponsored search can be seen as an interaction between three parties - SE, User and the Advertiser. The user issues a query to a SE related to the topic on which he/she seeks information. Advertisers and SEs try to exploit the immediate interest of user in the topic by displaying ads relevant to the query topic. Advertisers bid on certain keywords known as bid terms and their ads may get displayed based on the match between bid term and the user query. SEs try to rank the ads in a way that maximizes its revenue.

Search engines typically rank ads based on the expected revenue ($\epsilon_{ad}(Rev)$). Expected revenue from an ad is a function of both bid and relevance: $\epsilon_{ad}(Rev) = Bid{*}Relevance_{ad}$. The relevace of an ad is measured using its CTR. The CTR of an ad for a query is the no. of clicks normalized by no. of impressions for that query. CTR of an ad is a function of both ad and the query, i.e. an ad can have a different CTR for different queries. The CTR value for an ad-query pair is calculated form past click logs. For new/rare ads, we do not have any/sufficient past click data. Hence CTR for such ads need to be predicted so that they can be ranked along with other frequent ads. Richardson et. al. [5] predict the CTR based on ad text, ad quality etc. Fain et. al. [4] predict the CTR based on term clusters. We propose similarity features derived from click logs and advertisers hierarchy to accurately predict the CTR for new ads.

SIGIR'10, July 19–23, 2010, Geneva, Switzerland.
ACM 978-1-60558-896-4/10/07.

2. DATASET

The dataset used in our experiments comprised 12 days search log from Yahoo! search engine's US market. After removing redundant fields, each record in the dataset contained following fields: 1.Query 2.Term Id 3. Creative ID 4.Adgroup ID 5.Campaign ID 6.Account ID 8.CTR. Fields 2-6 point to a unique ad. Creative id points to the ad text. An ad text comprises bid term, title, abstract & display URL. The CTR values are normalized by removing the position & presentation bias. After some preprocessing, we got 1,447,543 unique query-ad pairs from the click through logs. It contained 1,97,080 unique queries and 9,43,431 unique ads. We randomly divide this dataset into 65-25-10 ratio for training, testing and validation respectively. We use Gradient boosted decision trees (GBDT) as a regression model [3]. Using validation set, the number of trees and no. of nodes parameters of GBDT were set to 600 and 150 respectively.

3. FEATURES

Features from Query-ad click graph: These features are based on the semantic relations of the queries and ads with other similar queries and ads. Regelson [4] have shown that similar ads (bid terms in their case) follow similar CTR distribution. The idea here is to learn the CTR values of query-ad pair from semantically similar queries and ads. We derive the semantic similarity from the query ad click-through graph. The click graph is built from 12 days query log (same period from which we generated our dataset). Queries are represented as vectors and these query vectors are compared to find similarity amongst the queries. A query q is represented as a vector of transition probability from q to all the ads in the graph. Edges are weighted using click frequency-inverse query frequency (CF-IQF) model: $cfiqf(q_i, a_j) = c_{ij} * iqf(a_j)$.

The transition probability from a query to an ad, $P(a_j|q_i) = cfiqf(q_i, a_j)/cfiqf(q_i)$. Each query is represented as $q = (P(a_1|q_i), P(a_2|q_i), \cdots, P(a_n|q_i))$. The similarity between two queries $\mathbf{q_i}$ and $\mathbf{q_j}$ is the cosine similarity between the two query vectors. $Sim(q_i, q_j) = Cosine\frac{\vec{q_i} * \vec{q_j}}{\| q_i \| \cdot \| q_j \|}$. This

Table 1: Improvement for various features (p-value $\leq$ 0.01)

Feature	RMSE (1e-3)	KL Diver-gence (1e-1)	% Improvement	Feature	RMSE (1e-3)	KL Diver-gence (1e-1)	% Improvement
Baseline	7.20	1.72	-	**Campaign**	**5.67**	**1.32**	**21.25%**
Sim-Q	5.86	1.42	18.61%	Account	5.94	1.39	17.50%
Sim-A	6.31	1.53	12.36%	AdH	6.20	1.46	13.9%
Sim-QA	**5.68**	**1.38**	**21.11%**	**SimQA+Camp**	**5.28**	**1.24**	**26.67%**
Term	6.24	1.45	13.34%	QADL	6.50	1.56	9.72%
Creative	6.51	1.50	9.6%	**SimQA+Camp**			
Adgroup	5.87	1.35	18.48%	**+QADL**	**5.14**	**1.21**	**28.61%**

similarity is used to predict the CTR for new query-ad pair by retrieving top k queries similar to q' and calculating the weighted average of the CTR values for all the ads over query q' as in [1]. Using query similarity,the CTR is estimated as:

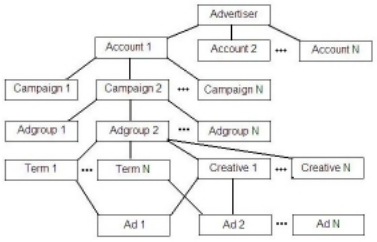
$$QCTR(a_i) = \frac{\sum_k CTR(q_k) * Sim(q_i, q_k)}{\sum_k Sim(q_i, q_k)}$$

The similarity between ads is also calculated in a similar fashion, with each ad being represented by the transition probability from ad to query $P(q_j|a_i)$ and similarity between two ads is reffered as $Sim(a_i, a_k)$. Using ad similarity, The CTR of is estimated as follows:

$$ACTR(a_i) = \frac{\sum_k CTR(a_k) * Sim(a_i, a_k)}{\sum_k Sim(a_i, a_k)}$$

Along with QCTR/ACTR We also consider the number of similar queries/ads retrieved (N_q/N_a). The Query and ad similarity featuers are called *Sim-Q* & *Sim-A*.

Figure 1: A typical Ad hierarchy



Features from Ad Hierarchy: Advertisements on an ad engine are typically maintained in some kind of a hierarchy. One such hierarchy is shown in Fig. 1. There are numerous reasons for maintaining ads in a hierarchy: (1)Advertiser's business may span various business units (BU). Ads from the same advertiser but from different BUs are maintained in different accounts. (2)For each BU, the advertisers can have ads on a range of products. Advertisements from the same account on similar products fall under the same Campaign. (3)Adgroups do further granular classification of ads. (4) Finally, an ad comprises a bid term and ad text (creative). Combination of these two makes an ad. We aggregate ads at each level viz. Term, Creative, Adgroup, Campaign and Account, compute the average within each group and use them as features in our model. In addition, number of featuers in each group are also taken as features. We call these features as *AdH* features. Detailed explanation of all the features is available in [2].

Features from Query-ad lexical match: In an attempt to capture how relevant an ad is to the query, we compute the lexical overlap between the query and these ad units. We compute various text matching features such as cosine similarity, word overlap, character overlap, and string edit distance for each combination of unigrams and bi-grams. We refer to this category of features as *QADL*. For all the set of features we also consider log of each feature as a feature. In all we have 50 features.

As shown in Table 1, Sim-Q & Sim-A give good improvements and when combined (Sim-QA) give an improvement of 21.11%. In the AdH category, Campaign (Camp) gave the best result and when *Sim-QA* was clubbed with *Camp* the improvement over baseline reached 26.67%. Finally, lexical feature did not yeild much improvement alone, but (Sim-QA+Camp+QADL) give the best performance with a good 28.61% improvement over the baseline. All these improvements are staistically significant at 99% significance level.

When all the features were ranked according to the feature importance [3]. Features like Campaign, ACTR, log(ACTR), No. of ads in campaign were amongst the top few.

4. CONCLUSIONS

We have proposed an approach to predict the CTR for new ads based on the similarity with other ads/queries. The similarity of ads is derived from sources like query ad click-through graph and advertisement hierarchies maintained by the ad engine. The model gives good prediction on the CTR values of new ads. Analysis of the feature's contribution shows that the features derived from the ad hierarchy and from the click-through graphs contribute the most to the model followed by some of the word overlap features.

5. ACKNOWLEDGMENTS

We are grateful to Yahoo! labs Bangalore for granting access to the ad click-through logs.

6. REFERENCES

[1] T. Anastasakos, D. Hillard, S. Kshetramade, and H. Raghavan. A collaborative filtering approach to ad recommendation using the query-ad click graph. In *CIKM '09*, pages 1927–1930, 2009.

[2] K. Dave and V. Varma. Predicting the click-through rate for rare/new ads. *Technical report IIIT/TR/2010/15*, IIIT-H, 2010.

[3] J. H. Friedman. Stochastic gradient boosting. *Comput. Stat. Data Anal.*, 38(4):367–378, 2002.

[4] M. Regelson and D. C. Fain. Predicting click-through rate using keyword clusters. In *Electronic Commerce (EC)*. ACM, 2006.

[5] M. Richardson, E. Dominowska, and R. Ragno. Predicting clicks: estimating the click-through rate for new ads. In *WWW '07*, pages 521–530, 2007.