Predicting Click-Through Rate

Your Name

Date

Predicting Click-Through Rate for New Users/Ads

Properly targeting ads to users so that the ads have a good click-through rate is a fundamental problem for the $50 billion US Internet ad market.

Short text messages usually marked as “sponsored ads” are the most common form of Internet ads. There are many types of text ads: *sponsored search advertising* places ads on the result pages from a web search engine based on the search query.

Contextual advertising places ads within the content of a generic, third party web page based on target words and the words on the page.

The central challenge is to find the "best match" between a given user in a given context and available ads. One common metric that can be used to measure this goodness of match is the click-through rate (CTR), that is, how often users click on an ad divided by how often they see the ad. The click thru rate is also affected by the position and prominance of the text ad on the page. Many online advertisers have a cost-per-click (CPC) model where advertisers bid on search terms with the highest bids receiving the best positions.

In this project, we want to investigate the problem of pricing the search terms. That is, estimating the probability that an ad will be clicked on as a function of the bid price.

Dealing with new ads or new users is called cold start scenarios, in which no prior events, like ratings or clicks, are known for certain users or advertisements.

Background

The authors in [2] proposed a new model for predicting click-through-rate (CTR) for new ads based on a logistic regression model given its features. The performance measurement is the average KL-divergence between the model’s predicted and true CTRs, where a perfect match would result in 0. They use Term CTR and Related Term CTR to boost the prediction result. The Term CTR feature is the CTR of other ads that contain the same bid term while Related Term CTR considers ads with subsets or supersets of the bi term. The authors also looked at ad quality, order specificity, and external search data. Combining these other features the method is able to achieve 29% improvement. The method achieved significant improvements in the area of reducing the error rate. One thing to note is that for an ad to be included in the data set it had to have at least 100 views. The question of are these

still considered new ads comes into play. The methods of this paper only consider ads' features and not on users. So the display of ads for users will be the same. This can be solved by personalized recommendation and could be an approach we will consider.

In [3], the authors combine the two different lines of work and propose a relevance-based approach that is augmented to use the click data to produce a CTR estimate. First, they used classical information retrieval approach to represent ads as vectors of features (unigrams, phrases, classes etc.). Representing page in the same vector space model, searching for best ads matching content can be measured by vector space similarity between two vectors. To incorporate the feedback, logistic regression is used to combine click feedback and semantic information extracted from pages and ads. This is accomplished in three main steps: feature extraction, feature selection, and coefficient estimation for features. The solution presented is scalable and can process large amounts of data in parallel using distributed frameworks such as Hadoop. However the experimental data used, as stated in the paper, might not be representative of the general online advertising community. Another drawback of this method is that a part of the method is based on ad impressions (which is the display of an ad on a page). This does not guarantee that the user will see the ad.

In [4] the authors imbibe features from three complementary approaches, though specifics of the proposal have been excluded. Their proposal predicts the CTR for new/rare ads for which other conventional CTR prediction models may not apply due to lack of any sufficient past click data. CTR prediction based on term clusters and ad text, as quality etc. are considered. The first feature derives from query-ad click graphs. A semantic relation is drawn between the queries-ads in context and other similar queriesads, which have known CTR using a query-ad click graph. Secondly all ads are aggregated into levels in an ad hierarchy based on the advertiser, accounts, campaign and ad groups. The third feature deals with a simple query-ad lexical match, in which the lexical overlap between the query and ad units is computed. Though this model is simple and boasted of considerable CTR improvements, it barely acknowledges other factors, which would affect an ad's CTR. The geographical location of the user coupled with current trends may greatly affect the user's interest in the ads presented.

[5]'s model proposes a technique to determine the relevance of an ad for a query search using CTR. It builds on a collaborative filtering scheme to discover new ads related to the query using a bipartite click graph of queries and ads by first determining similarities between queries. The authors construct bipartite graph G (Q, A, E) whose edges E represents a user click between queries Q and ads A. The weight of the edges is an indication of the strength of the association between the two and measures as the number or click or other metrics. In the first step of the algorithm the similarity between queries is estimated with the Pearson correlation coefficient. Next a set of neighbors' queries to the query in context is computed. A predicted response to ads associated to each of those neighbors is then estimated as

the weighted average of the neighboring responses. This provides a CTR prediction for the new query-ad combination.

Experimental results suggested a definitive edge over other commercial search engine baselines in terms of percentage of relevant listings an improved average CTR score. A drawback of this approach is that the semantics of queries that are captured in the ads tend to be highly diffused as the number of queries associated with an ad increases. Another concern would be the position bias affecting the response strength to the ads.

The authors of [6] also present an algorithm to predict the CTR for new ads following the estimation of similar values for existing ads from historical logs. They follow a three pronged approach. First the sample dataset acquired from Microsoft and the inherent anomalies found in it are described. This acquires significance as such anomalies in test data can manifest as skew results. Secondly, the challenge in computing CTR for existing ads is addresses using maximum likelihood estimation. Assuming that a particular position and page do not affect the comparison of two ads at the same position and page, a mathematical probability that an ad is clicked is derived, which servers as an interpretation of CTR in the paper. The drawback of this model is that a hidden variable is used to interpret the probability such that the ad is always seen at the first position of the page. The authors refute this claim since the actual position may not be relevant in comparing two ads. Lastly an algorithm for learning a set of decision rules for new ads CTR predictions is described. Here again a mathematical probability of increase or decrease of an ad click is arrived at using a simple set of if-then style rules. This in turn can be used to give recommendations to the advertisers suggesting changes to their ads. The authors claim that the initial results of this approaching are promising and recommend further work on the same.

Data sources

Data mining and website analytics are a popular and effective way to gather data about CTR for website advertisements. In addition, you can use Google's Webmaster tools (https://www.google.com/webmasters/tools).

- UCS Toolkit in R

Algorithms

Recommender System algorithms such as collaborative filtering or other neighborhood-based algorithms can be used to predict CTR.

- Recommender Systems: https://en.wikipedia.org/wiki/Recommender_system

- Collaborative Filtering Algorithms: https://en.wikipedia.org/wiki/Collaborative_filtering

Resources

-Kagggle's CTR prediction training files serve as a useful exercise for CTR prediction

- Here is their example code: https://www.kaggle.com/c/avazu-ctr-prediction/forums/t/10927/beat-the-benchmark-with-less-than-1mb-of-memory

-SciKit has tutorials and other lessons on subjects that are relevant to CTR prediction, such as regression modeling and clustering.

-Regression Model Tutorials in R

References

[1]http://www.businessinsider.com/the-iab-just-released-its-annual-presentation-and-itsa- must-see-for-anyone-in-the-industry-2012-6

[2] Richardson, Matthew, Ewa Dominowska, and Robert Ragno. "Predicting clicks: estimating the click-through rate for new ads." Proceedings of the 16th international conference on World Wide Web. ACM, 2007.

[3] Chakrabarti, Deepayan, Deepak Agarwal, and Vanja Josifovski. "Contextual advertising by combining relevance with click feedback." Proceedings of the 17th international conference on World Wide Web. ACM, 2008.

[4] Dave, Kushal S., and Vasudeva Varma. "Learning the click-through rate for rare/new ads from similar ads." Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval. ACM, 2010.

[5] Anastasakos, Tasos, et al. "A collaborative filtering approach to ad recommendation using the query-ad click graph." Proceeding of the 18th ACM conference on Information and knowledge management. ACM, 2009.

[6] Dembczynski, Krzysztof, W. Kotlowski, and Dawid Weiss. "Predicting ads' clickthrough rate with decision rules." Workshop on Targeting and Ranking in Online Advertising. Vol. 2008. 2008.

[7] Yanzhi Niu, Yi Wang, Gordon Sun, Aden Yue, Brian Dalessandro, Claudia Perlich, Ben Hamner "The Tencent Dataset and KDD-Cup'12". KDD-Cup Workshop, 2012.