A Relevance Model Based Filter for Improving Ad Quality

Hema Raghavan
Yahoo! Inc
Great America Parkway
Santa Clara, CA, 95054
raghavan@yahoo-inc.com

Dustin Hillard
Yahoo! Inc
Great America Parkway
Santa Clara, CA, 95054
dhillard@yahoo-inc.com

ABSTRACT

Recently there has been a surge in research that predicts retrieval relevance using historical click-through data[5]. While a larger number of clicks between a query and a document provides a stronger "confidence" of relevance, most models in the literature that learn from clicks are error-prone as they do not take into account any confidence estimates. Sponsored Search models are especially prone to this error as they are typically trained on search engine logs in order to predict click-through-rate (CTR). The estimated CTR ultimately determines the rank at which an ad is shown and also impacts the price (cost-per-click) for the advertiser. In this paper, we improve a model that applies collaborative filtering on click data by training a filter that has been trained to predict pure relevance. Applying the filter to ads that have seen few clicks on live traffic results in improved CTR and click-yield (CY). Additionally, in offline experiments we find that using features based on the *organic* results improves the relevance based filter's performance.

Categories and Subject Descriptors

H.3.3 [**Information Retrieval**]: Information filtering

General Terms

Algorithms, Experimentation

1. INTRODUCTION

Sponsored Search is the problem of finding candidate ads and ranking them for a search engine query. Advertisers submit creatives and bid on keywords or search queries. Ads are ranked by a combination of their estimated click-through-rate (CTR) and bid. When an ad is clicked on by a user the advertiser is charged by the search engine for the click. In the generalized second price auction used by most search engines, the cost of the click is a function of the bid and relevance of the ad below the clicked ad [4]. Models to estimate CTR are typically learned from click logs (eg., [5]). Clicks can be interpreted as a weak indicator of relevance, and models that learn from clicks are noisy due to issues of click fraud, accidental or exploratory clicks, and position-bias. Removing bad or irrelevant ads is particularly important in sponsored search problem because poor ads not only lead to a bad user experience, but also create a bad advertiser experience because a bad advertisement with a high bid can affect the cost paid by an advertiser whose ad is shown above this bad ad in the second price auction.

2. RELEVANCE MODELING

Our relevance filter is a binary classifier trained to detect *relevant* and *non relevant* advertisements, given a particular search. We experimented with Support Vector Machines, Maximum Entropy and adaBoost Decision Trees. While all algorithms had similar performance, boosting was slightly better and hence we report those results. The baseline model had 22 features: query length and 7 features that separately compared the query to the three *zones* of an ad (the title, description and display url). These seven features included word overlap (unigram and bigram), character overlap (unigram and bigram), string edit distance, cosine similarity, and a feature that counted the number of bigrams in the query that had the order of the words preserved in the ad zone. The first five features selected in the boosting iterations were (1) the cosine similarity between query and title, (2) character overlap between query and abstract (3) character overlap between query and display url (4) query length and (5) a feature that counted the number of bigrams in the query that had the order of the words preserved in the title. This 22 feature model (M1) forms our baseline model.

Web-queries and ad-creatives are both very short, so we hypothesized that query-expansion would be useful. In the past query-expansion on web-results has been shown to be useful for ad retrieval[2]. We expand the query using the summaries of the top 10 results from the Yahoo! search engine by computing a language model for each query as: $P(w|Q) = \sum_D P(w|D)P(D|Q)$, where $P(w|D)$ is a smoothed maximum likelihood probability and $P(D|Q)$ is a triangle function of rank. We compute the similarity between the expanded query and each of the title and abstract to get an additional two features for our second model (M2). We also experimented with a model (M3) that had features that determined whether each of the top 100 words from an ad-corpus occurred in the title, description or display url. Model M4 has the web-expansion features in addition to the word features.

3. EXPERIMENTAL SET UP AND RESULTS

We report performance on 10-fold cross validation and on a held out test set. The dataset for cross validation has about 117,000 query-ad pairs sampled from a major search engine. The held out-test set is sampled from a method that predicts query-ad relevance scores based on a collaborative

SIGIR'09, July 19–23, 2009, Boston, Massachusetts, USA.
ACM 978-1-60558-483-6/09/07.

	Cross Validation				Held-out			
	M1	M2	M3	M4	M1	M2	M3	M4
Precision	0.647(0.013)	0.653 (0.011)	0.652 (0.012)	**0.659 (0.012)**	0.566	0.557	**0.587**	0.558
Recall	0.856(0.015)	**0.866 (0.021)**	0.855 (0.016)	0.861 (0.017)	0.800	**0.866**	0.756	0.862
F1	0.737 (0.007)	0.744 (0.005)	0.739 (0.007)	**0.746 (0.004)**	0.663	**0.678**	0.661	**0.678**

Table 1: Results: Values in parentheses indicate cross validation error. Scores are reported at the threshold that achieves maximum F1.

filtering algorithm and has about 8000 query-ad pairs. The collaborative filtering algorithm is based on a common family of approaches called "neighborhood methods" (eg., [1]). The method can also be considered analogous to one that does a random-walk of 3 on the query-ad click graph[3]. Relevance assessors labeled the training and held out data using a graded scale. The set of possible labels is shown in column 1 of Table 2. For training and evaluating our classifiers all instances with a rating of *somewhat attractive* or better were marked *relevant*, and we measure precision and recall of relevant ads. In Table 1 we present results at the max F1 score for each model to provide a measure of the ranking quality of the models.

We see that model M2 improves all metrics over model M1 on cross validation, while on the held out test set recall is improved, but precision is slightly hurt. Model M3 has better precision over model M1. In model M3, many features that are triggered based on the occurrence of words like *cheap*, *insurance* etc in the query and words like *shipping* in the creative are very important. The addition of such words seems to increase precision at the expense of recall. Overall, in terms of F1, models M2 and M4 are significantly better than model M1 and M3, showing that web-expansions provide some benefit.

For the held-out test set we also report filtration rate (percentage of ads filtered) on the various editorial grades for ads in table 2. We find that using web-features (M2) and word features (M3) improves the performance of the model on the top 2 grades but causes more aggressive filtering on grades 3 and 4. Model M4 is excellent at not filtering *Perfect* ads and doing a good job at filtering non-relevant ads.

Grade	M1	M2	M3	M4
1) Perfect (15)	0.133	0.067	0.067	**0.000**
2) Certainly Attractive (87)	**0.056**	0.079	0.079	0.090
3) Probly. Attractive (218)	**0.172**	0.239	0.176	0.235
4) Somewhat Attractive (349)	**0.277**	0.399	0.322	0.406
5) Prob.ly not attr. (293)	0.445	0.563	0.519	**0.597**
6) Certainly not attr. (164)	0.697	0.786	0.721	**0.819**
7) Offensive/Risky (83)	0.760	0.760	0.760	0.760

Table 2: Filtration rates of the different models at operating points s.t. the Offensive/Risky filtration rate is the same for all methods. Numbers in parentheses indicate the number of query-ad pairs.

4. EXPERIMENTS ON LIVE TRAFFIC

Methods that infer relationships based on clicks can find relationships between phrases like *running shoes* and *sneakers* because they do not rely on strict word overlap. If we are "confident" about the clicks observed, these models can draw many meaningful associations. Our proposed relevance models rely heavily on word overlap, so (for live traffic) we wanted to apply it only on ads where the ad had not seen sufficient impressions to collect a robust observed click rate. The simplest way to implement this would be to apply a strict threshold on the number of impressions. However, the position bias in search engine logs makes comparing absolute metrics problematic. For example, if one ad is observed 5 times in the first position (where user visual attention is high) and a second ad has been observed the same number of time on the right hand side of the page (where user visual attention is very low) then we expect the second ad to have actually been seen by a user much fewer times. We compute an "expected clicks" metric as: ec(q,a) $= \sum_r \mathrm{imp}(q,a,r)P(\mathrm{click}|r)$. The quantity ec$(q,a)$ is the expected number of clicks for a query (q) and ad (a) computed over all possible rank positions (r). The quantity $P(\mathrm{click}|r)$ is estimated by observing the per-position click-through-rate on a size-able portion of the traffic for several days. For our experiment, a fraction of the search engine's users were bucketed into two bins, a control (or baseline bin) and an experimental bin that applied our filtering model to only those ads for which the observed ec(q,a) as well as the observed clicks for the given query ad pair (q,a) were less than a threshold. We measured *nCTR*, a version of the CTR metric that computes *clicks over the expected clicks.* This metric removes the position bias from the raw CTR metric (clicks/views). We then compared *nCTR* for ads suggested by our collaborative filtering approach and found an 8% improvement in *nCTR* for those ads in our experimental bucket, where 1.5% of all ads were filtered. This improved click rate can be attributed to filtering low relevance ads.

5. CONCLUSIONS

In this work we have proposed a relevance model based filter to remove poor ads. We have shown in both offline (human assessment) and online (live-traffic CTR) scenarios that the model improves a collaborative filtering based system that was learned on noisy click log data.

6. REFERENCES

[1] J. S. Breese, D. Heckerman, and C. Kadie. Empirical analysis of predictive algorithms for collaborative filtering. In *Proceedings of UAI '98 Uncertainty in Artificial Intelligence*, July 1998.

[2] A. Z. Broder, P. Ciccolo, M. Fontoura, E. Gabrilovich, V. Josifovski, and L. Riedel. Search advertising using web relevance feedback. In *CIKM*, 2008.

[3] N. Craswell and M. Szummer. Random walks on the click graph. In *SIGIR '07*, 2007.

[4] 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), March 2007.

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