Click-Through Rate Estimation for Rare Events in Online Advertising

Xuerui Wang, Wei Li, Ying Cui, Ruofei (Bruce) Zhang, Jianchang Mao
Yahoo! Labs, Silicon Valley
United States

ABSTRACT

In online advertising campaigns, to measure purchase propensity, click-through rate (CTR), defined as a ratio of number of clicks to number of impressions, is one of the most informative metrics used in business activities such as performance evaluation and budget planning. No matter what channel an ad goes through (display ads, sponsored search or contextual advertising), CTR estimation for rare events is essential but challenging, often incurring with huge variance, due to the sparsity in data. In this chapter, to alleviate this sparsity, we develop models and methods to smoothen CTR estimation by taking advantage of the data hierarchy in nature or by clustering and data continuity in time to leverage information from data close to the events of interest. In a contextual advertising system running at Yahoo!, we demonstrate that our methods lead to significantly more accurate estimation of CTRs.

INTRODUCTION

The Internet revolution has driven a transformation of how people experience information, media and advertising. Web advertising did not even exist twenty years ago, but nowadays it has become a vital component of the modern Internet, where advertisements are delivered from advertisers to users through different channels such as display ads, sponsored search and contextual advertising. Based on a recent study by Interactive Advertising Bureau and PricewaterhouseCoopers International, online advertising revenue in the first half of 2009 reached over 10.9 billion US dollars (IAB & PwC, 2009). Recent trends have also witnessed that larger and larger share of advertisers' budgets are devoted to the online world, and online advertising spending has greatly outpaced some of the traditional advertising media, such as radio and magazine.

There are many pricing models for online advertising. For example, in a pay-per-impression campaign, the advertisers pay for the number of ad exposures. This is a very popular model in display ads. Another example is pay-per-action, where the advertisers pay only if the ad leads to a sale or similar transaction. Both models have their own limitations. The pay-per-impression model does not consider ad performance while the pay-per-action model often has difficulty in tracking user transactions. Therefore, the success and effectiveness of online advertising campaigns is prevalently measured by whether users click on ads. This pricing model is called pay-per-click, where the advertisers pay a certain amount for every click. As a result, the click-through rate (CTR), defined as a ratio of number of clicks to number of impressions of an ad for a query or on a page, is one of the most useful and informative metrics to measure user response in online advertising. The use of CTR in online advertising is everywhere. For many advertising systems, CTR is one important factor to determine what advertisements to be displayed and in what order. In addition, advertisers often plan their budget based on historical CTRs or predicted CTRs.

In online advertising, a serving event refers to the showing of an ad in response to a user query, which can be a search query in sponsored search or a Web page in display and contextual advertising. The serving frequency varies for different queries and ads. Some popular ones account for a large fraction of Internet traffic while the overwhelming counterparts are extremely rare events, whose behaviors follow a typical

power-law pattern. Even being infrequent, the revenue from the rare events generates a several-billion-dollar business, which makes estimating CTR for such rare events an extremely important and practical problem. Unfortunately, due to the sparseness, the CTR estimation for rare events is notoriously challenging and extremely unreliable, often with huge variance. There are two scenarios that are especially difficult to deal with. The first one is when we observe a click out of only few serving events. A direct estimation of the CTR will be much higher than the true CTR. Another case is when we observe no clicks at all out of thousands of serving events. This is very common in contextual advertising because many page/ad CTRs are on the magnitude of 0.1% or even lower. Again, a direct estimation is not desirable here. In order to address these issues, we need to take into account that rare events are not strictly independent to each other. Instead, the existing correlations or associations among events can be leveraged into the CTR estimation for rare events.

First, in many scenarios, attributes in event data are naturally organized in hierarchies, and/or can be clustered via data-driven methods. An illustrative example of a page to publisher hierarchy is shown in Figure 1. As we can see, the publisher "yahoo_food_specialday" (special day site for Yahoo! Food) contains many pages that are similar to each other, such as foods for birthday and anniversary. These pages differ in content but share some common themes in the big picture, and we would imagine CTRs on these pages should be closer than the one on a random page. In general, events close in data hierarchy or in the same cluster share certain commonality compared to the events far in hierarchy or in different clusters. Information from "close" events can be utilized to enrich what we can learn from rare events themselves.
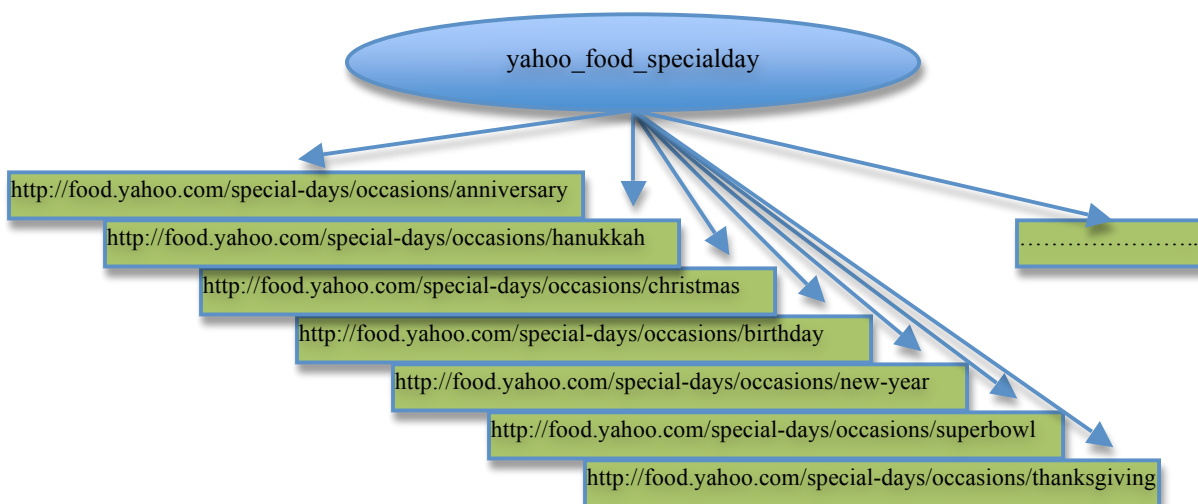

Figure 1. An illustrative example of a page to publisher hierarchy.

Second, most event data are often collected over time and advertising campaigns run over a certain period. Even though such dynamics make the patterns lying in the data earlier generally disappear later, the continuation in short-term still provides us with more reliable information about the events we are interested in. For example, an exponential smoothing technique can be employed to gather information from events closer in time to the events of interest, and at the same time, events far away play less significant role due to the dynamic changes in patterns.

It is worth pointing out that the two ideas described above can be employed either together or separately to estimate CTRs for rare events more robustly in most cases, since they exploit data commonality from two different angles. And we find the benefits are additive instead of being cancelled out by each other.

In this chapter, we develop models and methods to smoothen CTR estimation by taking advantage of the data hierarchy and data continuity in time to leverage information from data close to the events of interest. We use contextual advertising as an example to illustrate our approaches, but the models and methods discussed in this chapter could be well and easily extended to other online advertising channels such as display ads and sponsored search. Further more, the concept of smoothing comes into play handily as well when other ratios (such as conversion-rates) are under investigation, for which, the sparseness concern is even more severe, and therefore we expect more effectiveness.

BACKGROUND

The problem of CTR estimation in contextual advertising could be formalized as follows: for a given page/ad pair (*p, a*), with the number of impressions (*I*) and clicks (*C*), we want to estimate the CTR associated with this pair. For rare pairs, which drive the vast majority of impressions and clicks online, the standard estimation of CTR as *C/I* is inevitably unreliable. For example, if we only get one impression of a page/ad pair, and a click is recorded from that impression, the naïve calculation (*C/I*) would give a CTR = 1, which is a totally unrealistic estimate for most advertising campaigns. However, if we can leverage information from the data hierarchy to which the page/ad pair belongs to and/or the time series data of the page/ad pair, more reliable CTR estimation could be obtained.

Data Hierarchies

Pages and ads can be either associated with hierarchies in nature or grouped by applying clustering techniques in different metric spaces. On the page side, there is a natural hierarchy indicated by the URL path. For example, many pages belong to the same domain and further to the same publisher. On the ad side, an advertiser can have multiple accounts that roughly partition the ads. Furthermore, each account can have multiple ad groups where the ads in one group are closely related to each other. One could take advantage of the hierarch on the ad side, or on the page side, or both. To illustrate the idea, for simplicity, we describe the hierarchy (from ad to account) on the ad side in details, and the similar hierarchy on the page side (e.g., from page to publisher) also exists.

Assume that we have N ads in the same account $(a_1, a_2, \cdots, a_N)$ and a page p, and we are interested in the CTR of the page/ad pair (p, a_i). A two-stage empirical Bayes method is employed to estimate this CTR by combining information from empirical measurements on ad a_i and all other ads in the same account on page p. In our setting, the observed clicks $(C_1, C_2, \cdots, C_N)$ are assumed to be generated from an underlying set of CTR $(r_1, r_2, \cdots, r_N)$, according to a probability distribution, for instance, Binomial (I_i, r_i), where I_i is the number of impressions for the page/ad pair (p, a_i). The parameters r_i in turn can be considered as samples drawn from a population of the account characterized by further hyperparameters, according to a probability distribution, for instance, Beta (α, β). Information about r_i therefore comes not only from (I_i, C_i) which directly depends on it, but also from the CTRs of other ads in the same account inferred from the data as a whole, summarized by the hyperparameters α and β. The conjugacy between Binomial and Beta allows us to compute the marginal likelihood of all the data for the account and to estimate the hyperparameters α and β using maximum likelihood estimate (MLE). Once α and β are estimated (denoted by $\hat{\alpha}$ and $\hat{\beta}$, respectively), the posterior estimate of r_i could be calculated as

$$\hat{r}_i = (C_i + \hat{\alpha})/(I_i + \hat{\alpha} + \hat{\beta}),$$

which is considered as a smoothed CTR with smaller variance and more stability than C_i/I_i. As we discussed earlier, other hierarchies (such as page to domain) and multiple hierarchical levels (such as page to domain to publisher) can be included as well easily, and other compound distributions (such as Gamma-Poisson) can be possibly chosen based on the modeling need.

Data Continuity

In many cases, we are more interested in the trend of CTRs more than just a snapshot of the absolute values. For rare page/ad pairs, at any time point, the CTR estimation could be very noisy. By acknowledging some kind of data continuity, we treat impressions and clicks as discrete sets of repeat measurements, and an exponential smoothing technique is adopted to smoothen these measurements, and at the end we calculate a smoothed CTR.

Assume that, for a page/ad pair, we have M consecutive days of measurements of impressions $(I_1, I_2, \cdots, I_M)$ and clicks $(C_1, C_2, \cdots, C_M)$, and we want to estimate the CTR on the Mth day. Denote the smoothed impressions and clicks as $\hat{I}$ and $\hat{C}$ respectively, and we can calculate the smoothed CTR as $\hat{C}/\hat{I}$ in contrast to C/I. $\hat{C}$ is defined as follows ($\hat{I}$ is defined similarly):

$$\hat{C}_j = C_j \qquad\qquad\qquad j = 1$$
$$\hat{C}_j = \gamma C_j + (1 - \gamma)\hat{C}_{j-1} \qquad j = 2, \cdots, M$$

where γ is the smoothing factor, and $0 < \gamma < 1$ (the value of γ controls how much history we want to include in the smoothing: when γ is close to 0, we get very heavy smoothing from historical values; when γ is close to 1, little information in history is used for smoothing). In other words, as time passes the smoothed $\hat{C}_j$ becomes the exponentially decreasing weighted average of its past observations. One could apply exponential smoothing directly on CTR instead of impressions and clicks respectively, but we find that is often less effective.

As we pointed out earlier, the above two smoothing methods can be combined together with ease, for example, we can first do exponential smoothing on impressions and clicks before leveraging the data hierarchy.

SMOOTHING CTR ESTIMATION

In this section, we present the details of the smoothing methods discussed in the previous section, and their mathematical foundations.

Leveraging Data Hierarchies with Empirical Bayes Methods

In statistics, with empirical Bayes methods, the integrals over conditional probability distributions are substituted by the empirical statistics in the observed data, which allows us to estimate the posterior probabilities (such as CTRs) about one page/ad pair by leveraging the information from all page/ad pairs that are close to the page/ad pair of interest in the hierarchy.

There are multiple hierarchies that we could take advantage of in online advertising, and for illustration, we choose the hierarchy from page/ad pairs to publisher/account pairs. In another word, every page belongs to a publisher, and every ad is associated with an account, and in combination, every page/ad pair is mapped to a publisher/account pair.

Considering the hierarchical nature of the empirical Bayes methods, two assumptions are made in our experimental setting: 1) for each publisher/account pair, there is an underlying/unknown probability distribution of CTR, and the CTR of every page/ad pair in the publisher/account pair could be regarded a stochastic outcome of the distribution; and 2) for every page/ad pair, we observe the impressions and clicks generated from a distribution of the CTR of this pair. To parameterize the generative process, we use Beta and Binomial distributions, respectively in the above two assumptions. Essentially, we have a

Beta-Binomial compound distribution for the observed data. The generative process of this hierarchical Bayesian model is described as follows:

- Given a publisher/account pair that has N page/ad pairs, we have a distribution Beta(α,β).
- For every page/ad pair belonging to the publish/account pair:
    - Sample CTR $r \sim$ Beta(α,β), $p(r \mid \alpha,\beta) = \dfrac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} r^{\alpha-1}(1-r)^{\beta-1}$
    - Sample clicks C $\sim$ Binomial(I, r), $P(C \mid I, r) \propto r^{C}(1-r)^{I-C}$

The graphical model representation is shown in Figure 2. Shaded variables are observed, and others are hidden. Plate denotes replicates.

Figure 2. The graphical model representation of the CTR estimation with the empirical Bayes methods.
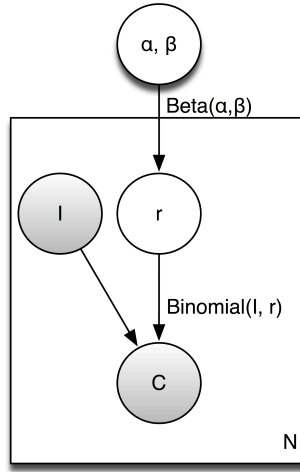
The resulting data likelihood over all the clicks of all page/ad pairs is

$$P(C_1,C_2,\cdots,C_N \mid I_1,I_2,\cdots,I_N,\alpha,\beta) = \prod_{i=1}^{N} P(C_i \mid I_i,\alpha,\beta) = \prod_{i=1}^{N} \int_{r_i} P(C_i,r_i \mid I_i,\alpha,\beta)dr_i$$

$$= \prod_{i=1}^{N} \int_{r_i} P(C_i \mid I_i,r_i)p(r_i \mid \alpha,\beta)dr_i \propto \prod_{i=1}^{N} \int_{r_i} r_i^{C_i}(1-r_i)^{I_i-C_i} r_i^{\alpha-1}(1-r_i)^{\beta-1}dr_i$$

$$= \prod_{i=1}^{N} \frac{\Gamma(\alpha+\beta)}{\Gamma(I_i+\alpha+\beta)} \frac{\Gamma(C_i+\alpha)}{\Gamma(\alpha)} \frac{\Gamma(I_i-C_i+\beta)}{\Gamma(\beta)}$$

The gradient of the log-likelihood is

$$\frac{d\log P(C_1,C_2,\cdots,C_N \mid I_1,I_2,\cdots,I_N,\alpha,\beta)}{d\alpha} = \sum_{i=1}^{N}\left[\Psi(\alpha+\beta) - \Psi(I_i+\alpha+\beta) + \Psi(C_i+\alpha) - \Psi(\alpha)\right]$$

$$\frac{d\log P(C_1,C_2,\cdots,C_N \mid I_1,I_2,\cdots,I_N,\alpha,\beta)}{d\beta} = \sum_{i=1}^{N}\left[\Psi(\alpha+\beta) - \Psi(I_i+\alpha+\beta) + \Psi(I_i-C_i+\beta) - \Psi(\beta)\right]$$

where $\Psi(x) = \dfrac{d}{dx}\ln\Gamma(x)$ is the digamma function, and can be calculated quickly using the Bernardo's algorithm (Bernardo, 1976).

The maximum can be calculated through the fixed-point iteration (Minka, 2003)

$$\alpha^{new} = \alpha \frac{\sum_{i=1}^{N} \left[\Psi(C_i + \alpha) - \Psi(\alpha) \right]}{\sum_{i=1}^{N} \left[\Psi(I_i + \alpha + \beta) - \Psi(\alpha + \beta) \right]}$$

$$\beta^{new} = \beta \frac{\sum_{i=1}^{N} \left[\Psi(I_i - C_i + \beta) - \Psi(\beta) \right]}{\sum_{i=1}^{N} \left[\Psi(I_i + \alpha + \beta) - \Psi(\alpha + \beta) \right]}$$

The stopping criterion we use is either 1000 iterations or the changes in both α and β are smaller than 1E-10. Once α and β are estimated (denoted by $\hat{\alpha}$ and $\hat{\beta}$, respectively), the posterior estimate of r_i could be calculated as $\hat{r}_i = (C_i + \hat{\alpha})/(I_i + \hat{\alpha} + \hat{\beta})$.

It is a common practice to smoothen an estimate with some background estimate, however, a combining parameter needs to be carefully tuned to get desirable performance. Our method avoids such hassle, and requires little to no human interaction.

Utilizing Event History with Exponential Smoothing

In statistics, with exponential smoothing, the time series data, as a sequence of noisy observations of an underlying random process, can be used to produce smoothed data for processing.

Directly substituting the recursively defining equation in the previous section for exponential smoothing back into itself, we can find that

$$\hat{C}_M = \gamma C_M + (1 - \gamma)\hat{C}_{M-1} = \gamma C_M + (1 - \gamma)(\gamma C_{M-1} + (1 - \gamma)\hat{C}_{M-2})$$
$$= \gamma C_M + \gamma(1 - \gamma)C_{M-1} + \gamma(1 - \gamma)^2 C_{M-2} + \cdots + \gamma(1 - \gamma)^j C_{M-j} + \cdots + \gamma(1 - \gamma)^{M-1} C_1$$

In another word, as time passes, the smoother estimate $\hat{C}_j$ becomes a weighted average of more and more past observations in the history, and the weights assigned to the historical observations are exponentially discounted as shown above, i.e., the weighting factors give more weight to the most recent measurements in the time series and less weight to older data.

Technically speaking, exponential smoothing can be regarded as an ARIMA model with one nonseasonal difference and an MA(1) coefficient, and no constant term, that is, an ARIMA(0,1,1) model with no constant term in statistics. The MA(1) term come from the quantity $1 - \gamma$. Therefore, one could take advantage of this to calculate confidence intervals for the exponential smoothing model.

Combining Data Hierarchies and Temporal Smoothing

We discussed that the two smoothing techniques discussed in this chapter can be combined together to achieve better performance than each individual model alone. For example, after exponential smoothing of the event data, it is straightforward to exploit the data hierarchies to further alleviate the sparsity concern.

One would think about conducting empirical Bayes with data hierarchy first, and then do temporal smoothing with CTR instead of impressions and clicks. Preliminary experiments show that directly smoothing CTR usually leads to large fluctuation and ends up with inferior results than smoothing

impressions and clicks separately. For this reason, we do not pursue this direction further, and only show results with empirical Bayes first, and then exponential smoothing. From now on, we refer to this as the combined smoothing strategy.

EXPERIMENTAL RESULTS

In this section, we discuss the evaluation metrics we adopted, and the corresponding experimental results for the methods that were laid out in the previous section: empirical Bayes, exponential smoothing and the combined smoothing.

Evaluation Metrics

We use the mean squared error (MSE) between the estimated CTR and the true CTR to measure the performance of our methods. The MSE is defined as

$$MSE(\hat{r}) = E[(\hat{r} - r)^2] = \sum_{i=1}^{N} (\hat{r}_i - r_i)^2 / N$$

MSE is always non-negative, and a zero MSE means that the estimator $\hat{r}$ predicts observations of the parameter r with a perfect accuracy. Note that, in this particular measurement, the constraint that $0 \le \hat{r} \le 1$ is not utilized.

On the other hand, the CTR we estimate is essentially parameterized with a Bernoulli probability distribution, thus, we can use the Kullback-Leibler (KL) divergence (Kullback&Leibler, 1951) to measure how the estimated CTR differs from the "true" CTR in the data. For probability distributions $P(r)$ and $Q(\hat{r})$, where r and $\hat{r}$ are the true CTR and the estimated CTR respectively, the KL divergence of $Q(\hat{r})$ from $P(r)$ is defined as

$$D_{KL}(P(r) \| Q(\hat{r})) = r \log \frac{r}{\hat{r}} + (1 - r) \log \frac{1 - r}{1 - \hat{r}}$$

$D_{KL}(P(r) \| Q(\hat{r}))$ is also always non-negative, and 0 when $\hat{r} = r$. Unfortunately, as we can see when r and $\hat{r}$ are in extreme values (such as 0 and 1), $D_{KL}(P(r) \| Q(\hat{r}))$ is not well defined, we ignore such cases in our KL divergence evaluation. However, a big trunk of the data does have extreme CTRs (e.g., $\hat{r} = 0$ or $r = 0$), and, not like the MSE, the KL divergence evaluation only provides results on a portion of data under investigation. Nevertheless, we have found that MSE and the average KL divergence in general agree with each other based on our experiments.

Experiments

We collect event data from a contextual advertising system at Yahoo! from Aug. 1 to Nov. 20, 2009 to provide the ground truth of the CTR of the page/ad pairs we are interested in. To make sure that we indeed have a relatively reliable ground truth, all the page/ad pairs we consider in this chapter have at least 10,000 impressions. We also eliminate non-Yahoo! publishers. This is done primarily for the following reason: the publisher outside of Yahoo! can be arbitrarily defined, there is no guarantee that the pages under one publisher are generally more similar than the pages from other publishers. To measure the effectiveness of the hierarchical smoothing with empirical Bayes, we also exclude the publisher/account pairs that do not have at least 10 page/ad pairs.

In the ground truth data, we have 30 billion impressions, 30K pages, 80K ads, 20K accounts and a little more than 100 publishers. Note that not every page/ad is a valid combination due to the sparsity nature of the data or our previous filtering, and we have 260K page/ad pairs that are mapped to 49K publisher/account pairs.

To test our methods against different sparsity settings, we uniformly sample 0.01%, 0.1% and 1% of the event data described above to get our test data. The statistics of the sampled test data are summarized in Table 1. As we can see, even with the lowest sampling ratio (0.01%), we still have a large data set with more than 1M impressions.

Table 1: Statistics of sampled data set with different sampling ratios

Sampling ratio	1%	0.1%	0.01%
Impressions	136.37M	13.58M	1.34M
Publishers	73	73	71
Accounts	1,245	1,245	1,102
Pages	27,739	27,739	25,316
Ads	30,922	30,922	26,400
Publisher/account pairs	2,588	2,588	2,249
Page/ad pairs	176,389	176,389	149,121

Results on Smoothing with Data Hierarchies

In Table 2, we report the MSE and the average KL divergence for the test data we collected with different sampling ratios using the empirical Bayes method, where the baseline method is not involved with any smoothing.

Table 2: Smoothing with data hierarchies on different levels of sparsity

Sampling ratio	Methods	MSE (%improvement)	Avg KL divergence (%improvement)
1%	Baseline	9.46041e-7	0.00141036
	Empirical Bayes	8.43402e-7 (10.85%)	0.00138567 (1.75%)
0.1%	Baseline	1.09668e-5	0.0157694
	Empirical Bayes	9.05408e-6 (17.44%)	0.0146967(6.80%)
0.01%	Baseline	6.45237e-5	0.103176
	Empirical Bayes	3.18008e-5 (50.71%)	0.070434 (31.73%)

From Table 2, we can find that both the MSE and the KL divergence become worse when the data are sparser (smaller sampling ratios) in both baseline and empirical Bayes, which indicates the CTR estimation indeed becomes more challenging when data are sparser as expected. On the other hand, the improvement with smoothing is more and more dramatic when the data are sparser. With the 0.01% sampling ratio, we achieve 50% deduction in MSE and 30% deduction in average KL divergence.

Results on Temporal Exponential Smoothing

As to exponential smoothing, it puts more emphasis on observations in recent history than in earlier history. For example, with $\gamma = 0.7$, the contribution weight from ten days ago only accounts for 2E-6 of the overall observations. In our experiments, with the sampled events (sampling ratio = 1%) on Sep. 10-30, we apply the exponential smoothing technique, and compare the estimated daily CTR accuracy on Sep 26-30, in comparison to the one of the baseline method that does not have smoothing at all. Theoretically, the effect of smoothing can be controlled by the value of γ, and we experimented 3 different values of γ.

For temporal smoothing, the MSE and the average KL divergence with different values of γ are shown in Table 3. At first glance, no matter what value of γ we used, the exponential smoothing significantly

outperform the baseline in both MSE and the average KL divergence. Also as expected, different values of γ have significant influence on the smoothing performance. As γ decreases, the more historical information that gets incorporated into the estimation, the more improvement we obtain.

Table 3: (Temporal) exponential smoothing with different smoothing factors

Methods	Smoothing factor (γ)	MSE (%improvement)	Avg KL divergence (%improvement)
Baseline	N.A.	1.97448e-5	0.0623606
Exponential smoothing	0.9	1.62459e-5 (17.72%)	0.0519508 (16.69%)
	0.7	1.34620e-5 (31.82%)	0.0459063 (26.39%)
	0.5	1.14469e-5 (42.03%)	0.0408431 (34.51%)

Results on the Combined Smoothing

As we discussed before we can combine the two different smoothing techniques to further improve the estimates because they utilize different information to deal with sparsity. In our experiment, we first employ the exponential smoothing on the test data we experimented for Table 3, and then utilize the empirical Bayes method on the temporally smoothed data to get the final CTR estimates. The experimental results are listed in Table 4, with some results copied from Table 3 for comparison purpose. To study the advantage of combing two different smoothing techniques, the results from each technique alone are also listed in Table 4.

First of all, both empirical Bayes alone and exponential smoothing individually have significantly outperformed the baseline that does not conduct smoothing. Second, the advantages we get from two different smoothing techniques are additive as expected, since they utilize different perspective of the data. Third, when the effect from temporal smoothing gets improved (while decreasing the value of γ), the performance from the combined model is increased as well. In another word, even if we improve one of the models alone, a significant boost in the combined model is still demonstrated in Table 4. This is a strong indicator that we should combine the two different and orthogonal techniques whenever it is possible to do so.

Table 4: Comparison of different smoothing techniques

Smoothing factor (γ)	Methods	MSE (%improvement)	Avg KL divergence (%improvement)
N.A.	Baseline	1.97448e-5	0.0623606
	Empirical Bayes	1.07309e-5 (45.65%)	0.0459972 (26.24%)
0.9	Exponential smoothing	1.62459e-5 (17.72%)	0.0519508 (16.69%)
	Combined smoothing	8.30497e-6 (57.94%)	0.0371832 (40.37%)
0.7	Exponential smoothing	1.34620e-5 (31.82%)	0.0459063 (26.39%)
	Combined smoothing	6.15689e-6 (68.82%)	0.0304852 (51.11%)
0.5	Exponential smoothing	1.14469e-5 (42.03%)	0.0408431 (34.51%)
	Combined smoothing	4.63981e-6 (76.50%),	0.0241584 (61.26%)

RELATED WORK

Click-through rate is one of the most studied topics in online advertising. Historically, the majority of methods deals with events with high frequency, and ignores rare events or aggregates them to higher frequency. More recently, researchers have paid more attention to CTR estimation for rare events. Richardson et al. (2007) discussed a method of learning CTRs of new ads in sponsored search using a logistic regression model. They carefully designed features such as top unigrams, how many words in the

title, the fraction of images in the landing page, etc. The CTRs reported in their work is higher than ours because the user intention in sponsored search is much clearer than in other channels where CTRs are generally lower, and event data are sparser.

Agarwal et al. (2007) mapped pages and ads into a hierarchy that is derived from contextual information. However, the coverage in the hierarchy is very unbalanced, and for most regions of the hierarchy, the sparsity is still a concern. They designed a sampling mechanism and a two-stage model to overcome the sparsity problem: the first stage adjusts the sampling bias and the second stage incorporates correlation among sibling nodes through a tree-structured Markov model. Agarwal et al. (2009) designed a spatio-temporal model to estimate CTRs in the context of content recommendation. They employed a dynamic Gamma-Poisson model to track CTR over time at specific locations and combined information from correlated locations through dynamic linear regressions. Large-scale experiments showed that their model provided very encouraging results.

CONCLUSION

Estimating CTRs for rare events are extremely important but very challenging. We designed two different smoothing techniques to deal with the sparsity problem lying in the rare events. The first technique utilizes the natural data hierarchy with an empirical Bayes method, and the second takes advantage of temporal continuity in certain period. Both techniques leverage information from events that are closely related to the one of interest, but from orthogonal perspectives. Thus we also experiment a combined model that achieves more than 70% reduction in mean squared error and more than 60% deduction in average KL divergence against the baseline without smoothing.

In this chapter, we have demonstrated the effectiveness of estimating CTRs in a contextual advertising system, but the same techniques could be easily extended to other advertising channels or other ratios of interest. Also, the ad ranks are not considered in this chapter, while rank has a decisive role in CTR in many online advertising circumstances due to user attention capture. However, breaking-down the events that are already rare by rank will increase the severity of sparseness, and we believe that the smoothing techniques discussed here could be more effective to deal with that problem.

The contextual information of pages and ads provides very useful information in CTR estimation as demonstrated in Richardson et al. (2007) and Agarwal et al. (2007), and we believe that this is another orthogonal dimension we can potentially enhance our model additively.

REFERENCES

Agarwal, D., Broder, A. Z., Chakrabarti, D., Diklic, D., Josifovski, V., and Sayyadian, M. (2007). Estimating rates of rare events at multiple resolutions. *In Proceedings of the 13th ACM SIGKDD international Conference on Knowledge Discovery and Data Mining* (pp. 16-25).

Agarwal, D., Chen, B., & Elango, P. (2009). Spatio-temporal models for estimating click-through rate. *In Proceedings of the 18th international Conference on World Wide Web* (pp. 21-30).

Bernardo, J. M. (1976). The Psi (Digamma) function: An algorithm. *Applied Statistics 25*, 315-317.

Interactive Advertising Bureau (IAB) & PricewaterhouseCoopers International (PwC). (2009). Internet ad revenues at $10.9 billion for first half of '09. Retrieved Dec. 15, 2009, from http://www.iab.net/about_the_iab/recent_press_releases/press_release_archive/press_release/pr-100509

Kullback, S., & Leibler, R. A. (1951). On Information and Sufficiency. *Annals of Mathematical Statistics 22*(1): 79–86.

Minka, T. P. (2003). Estimating a Dirichlet distribution. http://research.microsoft.com/en-us/um/people/minka/papers/dirichlet/

Richardson, M., Dominowska, E., & Ragno, R. (2007). Predicting clicks: estimating the click-through rate for new ads. *In Proceedings of the 16th international Conference on World Wide Web* (pp. 521-530).