ISSN 0005-1055, Automatic Documentation and Mathematical Linguistics, 2013, Vol. 47, No. 2, pp. 52–58. © Allerton Press, Inc., 2013.
Original Russian Text © K.E. Bauman, A.N. Kornetova, V.A. Topinskii, D.A. Khakimova, 2013, published in Nauchno-Tekhnicheskaya Informatsiya, Seriya 2, 2013, No. 4, pp. 1–8.

Optimization of Click-Through Rate Prediction in the Yandex Search Engine

K. E. Bauman, A. N. Kornetova, V. A. Topinskii, and D. A. Khakimova

Higher School of Economics, Myasnitskaya ul. 20, Moscow, Russia

e-mail: kbauman@yandex-team.ru, kornetovaanna@yandex.ru, topinsky@gmail.ru, d.a.khakimova@gmail.com

Received September 19, 2012

Abstract—The problem of the estimation of the click-through rate on advertisements that are placed on a search-engine results page is discussed. The proposed methods improved the prediction quality (both in terms of likelihood metrics and the principle parameters of the engine). The cases of advertisement displays are considered when the history of an ad is rather short (i.e., advertisements that are considered to be new). The proposed prediction formula takes the dispersion and high risk of displaying a new advertisement into account.

Keywords: click-through rate estimation, prediction quality metrics, off-line prediction, experiment planning

DOI: 10.3103/S0005105513020040

INTRODUCTION

Yandex is the largest Russian Internet search engine. In response to a user's query Yandex shows two different types of information on the results page, namely, the search results in response to the query and ads that are chosen in a certain way for a given query.

The search results are presented as a list of results. The search engine forms a list of results in order to respond to a particular query with maximal accuracy. They are ranked based on the relevance to the query prior to being displayed.

Ads, the second type of information, are displayed on the right side of the search engine results page (so-called ad displays at the right) and above the search results (premium placement). In all large Internet search engines, the revenue that is obtained via search advertising occupies an important and often a decisive part in the company's total revenue. Based on this criterion, the advertising results should be formed with due regard for several factors, such as the correspondence of an ad to a user's query and the expected revenue from this ad.

At present, the problem of ad selection for a particular query is solved using several selection mechanisms. An advertiser assigns a set of keywords to each ad. Depending on the selection mechanism, the ad can be selected for showing in the following cases: at least one key phrase and the query match totally, key phrases and the query match partially (one of the key phrases is part of the query), or key phrases match the query semantically (broad matching by sense or theme).

This paper aims to solve the second subproblem of ad selection, namely, the consideration of the expected revenue from an ad that is displayed.

Most search engines, including Yandex, use the mechanism of advertising sales during searching where an advertiser pays for a user click on their ad. In other words, the advertiser pays only for real user transitions from search pages to advertising sites. In this case, the rates are determined by a generalized second-price auction [2].

The quantity of ads that can be shown to a user on the search results page is limited. As a rule, it can be a maximum of three ads above the results list and four or five ads on the right-hand side of the page. The engine selects the most effective ads among the selected ones.

In practice, the selection is performed by choosing ads with the highest cost-per-mille (*CPM*). This name is attributed to the history of Internet advertising, when ad displays were sold. Today, advertisers deal in clicks; therefore, *CPM* is now calculated as

$$CPM = BID\,CTR,$$

where *COST* is the advertiser's rate for a click on their ad and *CTR* is the click-through rate.

Advertising technologies were described in more detail in [3].

Therefore, the effective selection of ads is reduced to the problem of the estimation of the click-through rate. This problem is one of the most important ones in modern search engines and the quality of its solution has a direct effect on the degree of user satisfaction, the effectiveness of context advertising for advertisers, and the revenue of the search engine itself.

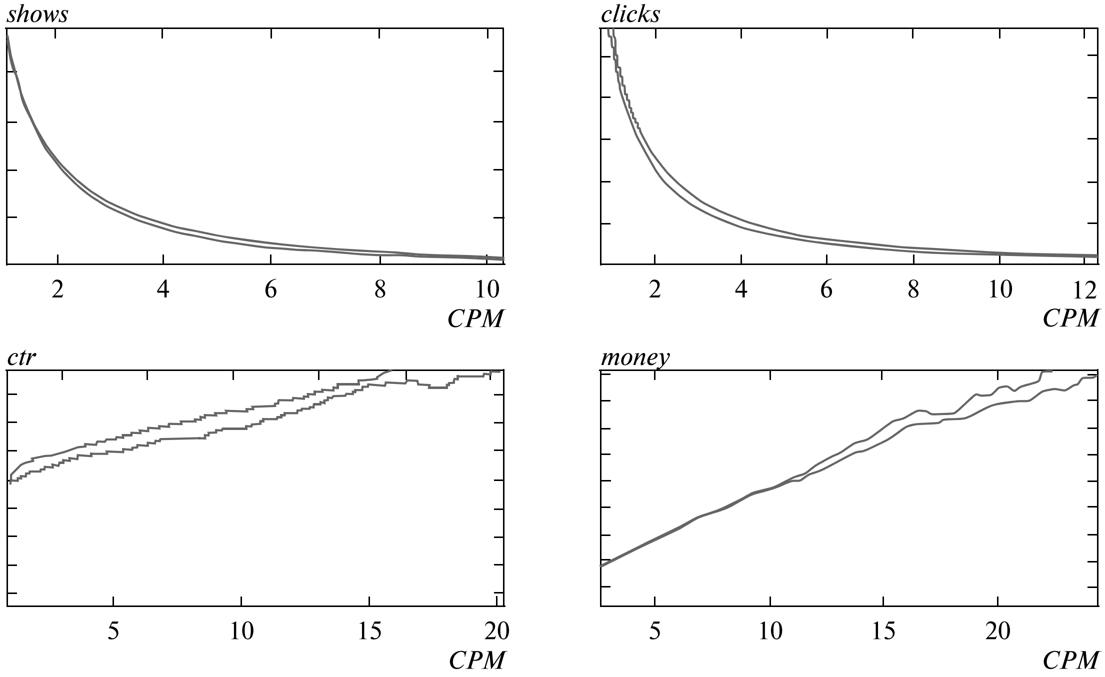

Fig. 1. The method RTC-similar curves: *shows*—the number of impressions, *clicks*—the number of clicks, *ctr*—the average *CTR* of the system, *money*—the amount of money.

The authors of this paper discuss the problems that occur when estimating *CTR*.

1. OFF-LINE OPTIMIZATION

Let us consider metrics that we can use to compare different predictions for click-through rates. In the literature on applied problems of the theory of probability and mathematical statistics, widely used classical metrics include the likelihood algorithm, mean square error and mean absolute error, correlation measures, the area under *ROC*-curve [5], and so forth.

In practice, the *CTR* prediction or estimation is only part of a large and complex system of displaying ads on the Internet. It is extremely interesting to determine how a new *CTR* prediction influences such general system indicators as the average characteristic of the entire engine or of individual traffic components, total revenue per day (per month), traffic level which is expressed in clicks, etc.

The best method, which is rather complicated and expensive, is conducting an on-line experiment on a certain part of the search queries; however, these experiments can take time and cause a company possible losses on experimental traffic (in both revenues and clicks). Therefore, we have developed an off-line technique for comparing two different predictions of click-through rates.

1.1. Method of ROC-Similar Curves

We will describe the essence of building plots to compare predictions.

We have records from the database about previous displays of ads as input data. Using *CTR* prediction, we assign a number of derivative characteristics including *CTR*, *CPM*, and *cost* per click, to each record.

Further, by ordering all the events by *CPM* we build the following curve, e.g., for the analysis of click quantity, namely, we consider the current *X*-coordinate value as the *CPM* threshold. The *Y*-coordinate value is then the quantity of clicks in those events whose corresponding *CPM* values exceeded this threshold. In the same way, we build plots for revenue, *CTR* (Fig. 1).

These curves allow us to compare different predictions. If the curve for a new prediction dominates the curve for the current prediction then the new prediction is obviously preferable. This can be interpreted as if we rejected or prohibited the ads with the lowest *CPM* (which occurs in practice). Then, according to the first prediction, the ads with the highest estimates of expected clicks or revenue would be shown (according to the curve with the most evident domination).

This approach has one complication. Two curves can often intersect inside the range of the *CPM*-thresholds. However, it is obvious that at the ends of the region, these curves are always identical. Additional limitations can help to resolve the situation with internal intersections and, consequently, with alterations in domination. For example, the search engine can have a limitation on the share of generated search pages that contain ads. In this case, a corresponding *CPM* threshold can easily be calculated and the predictions can be compared at corresponding points.

1.2. Simulation of a System for Displaying Ads

The second method for comparing two predictions in terms of expected clicks and money is building an off-line model for the entire system of advertising display events.

With this aim in mind, we copy all the current characteristics of the system, namely, information on the ads that are available, their rates, budgets, etc. After this, we combine these copies with the current version of the implementation of the program complex for the system of ad display events and thus, we obtain a local copy of the system of ad display events with actual data.

In order to obtain the basic characteristics of the system, we take a random set of queries (possibly, with repetitions to keep the proportion of the natural traffic) and input it in the local copy of the system. The output is a set of all ads that went through a selection procedure. For each ad, we can count *CTR* prediction versions that are of interest for a given query and complete the formation of the ad list for display. At the output, we obtain as many sets of ads for each user's query as there are predictions that we want to study.

After this, we can count the expected revenue, clicks, etc. using a particular version of the search engine results page with advertisements. To calculate mathematical expectations as a true distribution of probabilities, we use a version of the prediction with the best parameters of the quality metrics (likelihood, linear correlation, etc.)

We will describe these calculations in detail.

Let *H(Q, CTR)* designate the set of sets of ads that went through the selection for the queries from the set Q using the prediction $CTR(\cdot)$. Each $h \in H(Q, CTR)$ is the set of advertisements that were selected for showing in response to a particular query. The lower index i will be used for the reference to a particular ad from all that were chosen to display. Let $CTR_{new}(\cdot)$ be a new prediction of the click-through rate (under study) and $CTR_{base}(\cdot)$ be the current prediction.

Then, we can calculate the expected changes in the basic characteristics of the system according to the following formulas:

$$\Delta\text{Coverage} = \frac{|H(Q, CTR_{new})|}{|H(Q, CTR_{base})|} - 1,$$

$$\Delta\text{Traffic} = \frac{\sum_{h \in H(Q, CTR_{new})} \sum_{i \in h} CTR_{new}(i)}{\sum_{h \in H(Q, CTR_{base})} \sum_{i \in h} CTR_{new}(i)} - 1,$$

$$\Delta\text{Revenue} = \frac{\sum_{h \in H(Q, CTR_{new})} \sum_{i \in h} \cos t_i\, CTR_{new}(i)}{\sum_{h \in H(Q, CTR_{base})} \sum_{i \in h} \cos t_i\, CTR_{new}(i)} - 1.$$

Here, Coverage and Traffic are taken to mean the share of the queries from the general set that contain at least one ad and the quantity of clicks on these ads, $CTR_{new}(i)$, $CTR_{base}(i)$, $\cos t_i$, respectively, are predicted click-through rates according to the new and the current predictions and possible payment per click calculated for a particular ad i.

We compare the derivative characteristics of the algorithm operation with different prediction versions and, possibly, with different specially selected model hyperparameters. For the calculated characteristics to be comparable, we should use the same probability distribution that recognize it as the estimation of a true unknown distribution. In the case of Coverage we do not use multiplication by probabilities, since in this case we do not depend on clicks as the only source of randomness in the given model.

Which particular distribution is better? We propose to choose the distribution that meets the best characteristics of the prediction quality as the best approximation of an unknown true distribution (in our case, it is worthwhile to study a new prediction only if it is superior in such characteristics as likelihood, mean error, etc.).

It should be noted that this method allows us to perform optimization using the free hyperparameters of the system by means of full reconstruction of the selection rules for advertising display events.

2. CONSTRUCTION OF ESTIMATIONS AND REGULARIZATION

2.1. Data in Use and Current Estimation of CTR

The current system for ad that are displayed on Yandex is arranged in such a way that we have statistics on each ad that was selected for a given display in a given query, such as the number of times it was displayed and the quantity of clicks that correspond to these display events on different parts of the search traffic.

Here and below, let $clicks_b$ be the quantity of clicks on a given ad b and $displays_b$ be the corresponding quantity of displays of the given ad during which these clicks were collected. Let us assume that we fixed a certain section, e.g., we collect the statistics for the ad–purchased phrase pair by which a display occurred. We designate through λ the set of ads for which we keep statistical data.

An event during which the ad is clicked on we simulate through the Bernoulli random value taking values 0 (non-click) and 1(click), where the click-through rate is $p = CTR_b$.

At the time of the present study, Yandex has already implemented an estimation of the click-through rate based on the set of click statistics. This prediction possessed certain features, namely, on sufficiently large volumes of display events, the prediction was reduced to the clicks/display ratio and in the absence of sufficient statistics, a statistical a priori prediction was used (a rather simple piecewise constant function).

Our aim is to build a new prediction based on the same data which possesses the same properties and whose usage could yield the best results according to the basic metrics of the system operation quality.

2.2. Estimation of CTR and the Minimization of MSE

Let us consider how one can build a new prediction for *CTR* based on the statistics on a certain section.

Assume that we have additional information on the set $C = \{CTR_b | b \in \Lambda\}$. For example, this information is reduced to $\overline{CTR_b}$, which is the estimation of the mean value of the CTR_b. Then we will try to build such a prediction that, depending on the quantity of display events in the statistics, would be reduced from this additional information to an actual ratio of clicks to displays. More formally, we are searching for the *CTR* prediction in the form of the linear combination:

$$C\hat{T}R(displays)$$

$$= \alpha(displays)\overline{CTR} + \beta(displays)\frac{clicks}{displays},$$

$$\text{where} \quad \overline{CTR} \approx mean_{b \in \Lambda}(CTR_b).$$

In order to find unknown functions α(*displays*) and β(*displays*), we can choose the mathematical expectation of error squares *Q*(*displays*) as a target functional for minimization.

$$Q(displays) = E\sum_{b \in \Lambda}(CTR_b - CTR_b(displays))^2.$$

Then, it is easily to find that optimal values for these functions are:

$$\alpha(displays) + \beta(displays) = 1,$$

$$\alpha(displays) = \frac{S_0}{S_0 + displays},$$

where $S_0 \approx \dfrac{mean_{b \in \Lambda}(CTR_b(1 - CTR_b))}{\text{var}_{b \in \Lambda}(CTR_b)}$.

Therefore, the final prediction can be written as

$$C\hat{T}R_b = \alpha(displays)\overline{CTR} + \beta(displays)\frac{clicks}{displays}$$

$$= \frac{clicks + C_0}{displays + S_0}, \quad \text{where } C_0 = \overline{CTR} \ S_0. \tag{1}$$

Below, Fig. 2 presents the comparison of the given prediction with the current one.

2.3. Estimation of CTR and Likelihood Maximization

The Bayesian approach with assumed a priori distribution is another widely known method for obtaining a prediction of the same type. Let $P_{prior}(CTR)$ be an a priori distribution of the true *CTR* in the set Λ. Then, the a posteriori probability for CTR_b can easily be calculated as

$$P_{posterior}(CTR_b)$$

$$= P_{prior}(CTR_b)CTR_b^{clicks_b}(1 - CTR_b)^{displays_b - clicks_b}.$$

If we assume that the a priori distribution $P_{prior}(CTR)$ belonged to the family of Beta- distributions (i.e., $P_{prior}(CTR) = \dfrac{CTR^a(1 - CTR)^d}{B(a + 1, d + 1)}$), then the corresponding a posteriori distribution, as is known, will also be the Beta-distribution with parameters (a + $clicks_b$, d + *display events*$_b$ – $clicks_b$). Therefore, the a posteriori mathematical expectation of click-through rate *b* is

$$C\hat{T}R_b = \frac{a + clicks_b}{a + clicks_b + d + displays_b - clicks_b}$$

$$= \frac{clicks_b + a}{displays_b + a + d} = \frac{clicks + C_1}{displays + S_1}.$$

Therefore, using different initial assumptions we have arrived twice at the same estimate. In the given case, we have two unknown constants that can be obtained, if necessary, via building an estimation of the a priori distribution. Instead, we will propose another heuristics, which resulted from the considerations in section 2.2.

2.4. Practical Use of the Proposed Estimate

As was shown, the regularized ratios of clicks to displays with regularization is a fairly widely spread form for estimating click-through rates. We write this ratio as it was derived in section 2.2.

$$C\hat{T}R_b = \frac{clicks + D_0C\tilde{T}R}{shows +}D_0,$$

where $C\tilde{T}R$ is additional information.

Now, let us assume that we use a statistical prediction (i.e., a prediction that is calculated without the direct use of statistics of clicks and display events, e.g., it can be a logistic regression built on the set of test attributes, information about a particular user, etc.) as additional information.

Assume that this statistical prediction has a positive correlation *R* with the corresponding random value

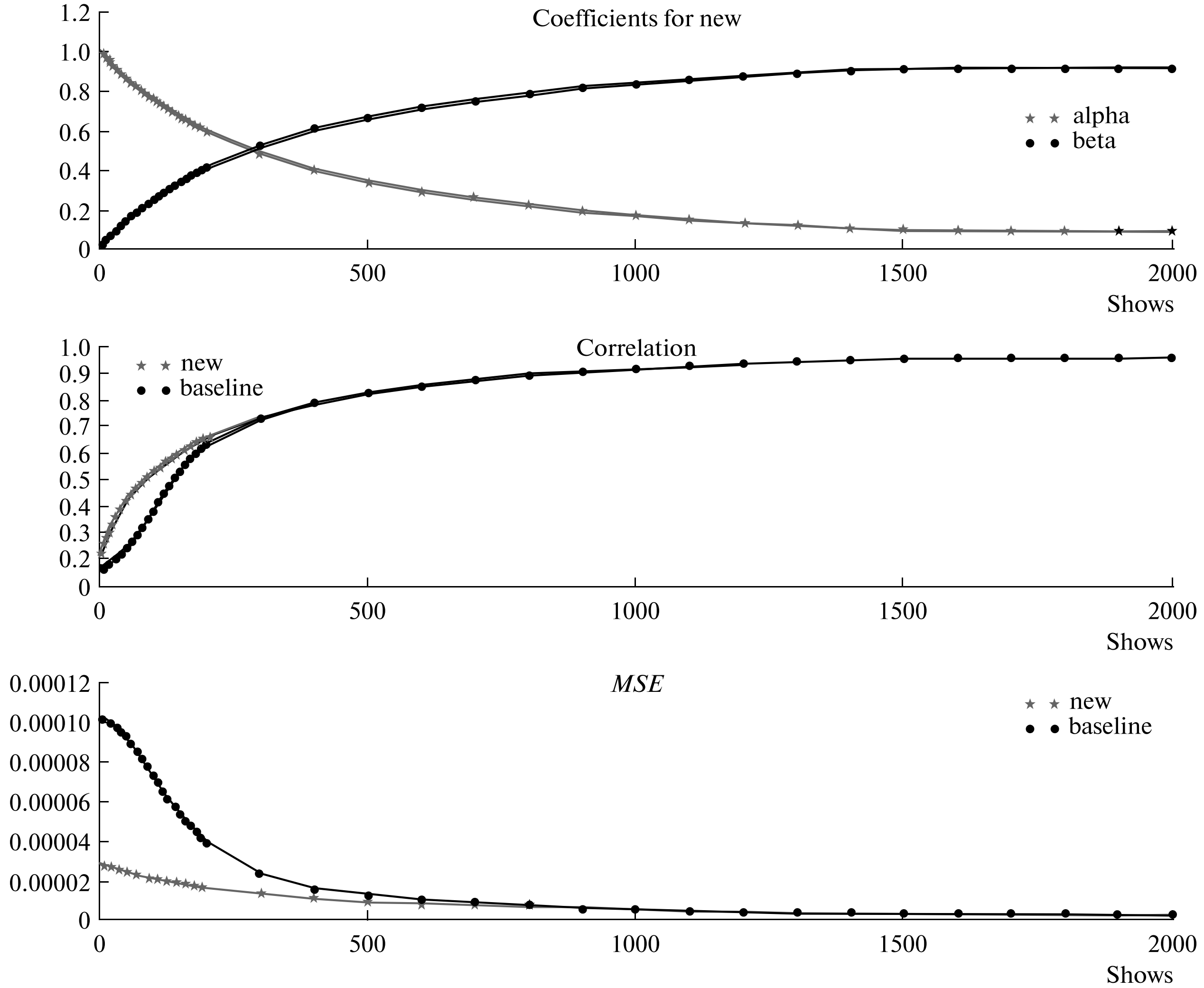

Fig. 2. The upper plot demonstrates the behavior of functions α and β; the middle plot represents the correlation of two predictions with clicks; the lower one is the mean error square.

$click = \sum_{b \in \Lambda} click_b \cdot displays_b$, where $display_b = 1_{(b,\text{ was shown})}$, $click_b = 1_{(b,\text{ was shown})}$. Then, if we assume that the given prediction has a low dispersion: $\Sigma < R^2\text{var}[clicks]$ it is easy to demonstrate that $D_0 > S_0$ under the condition $\alpha(displays) + \beta(displays) = 1$.

This fact can be interpreted in a natural way. If we build a statistical prediction for a better *CTR* quality than the mean *CTR* estimation of true rates then the final prediction will fit our initial approximation better. In other words, the convergence to the click–display ratio will be slower.

3. MINIMIZATION OF RISKS

When using any click-through rate prediction in practice in the current system of ad displays we may encounter undesirable effects and risks.

A prediction clearly has a lower dispersion at a large volume of statistics. Therefore, for new ads we, encounter e.g., highly scattered *CTR* predictions but owing to the ad selection algorithm for display events we obtain high risks that the minimum cost for placement in a desirable part of the search engine results page might be unreasonably high.

For this purpose, we undertook a number of simulations to calculate the probability that the minimum cost for entering the corresponding blocks increased k times compared to the costs at the current prediction for a specified true *CTR* value.

The simulation confirmed the fact that the new prediction had a high risk for low-statistic ads; therefore, an undesirable obstacle can occur, namely, a high initial cost, for the participation of these ads.

The proposed prediction type was modified in such a way that its dispersion at low values of display events agreed with the dispersion that is set at a large quantity of display events.

Let n_0 be the minimum quantity of display events for which we consider the prediction dispersion appli-

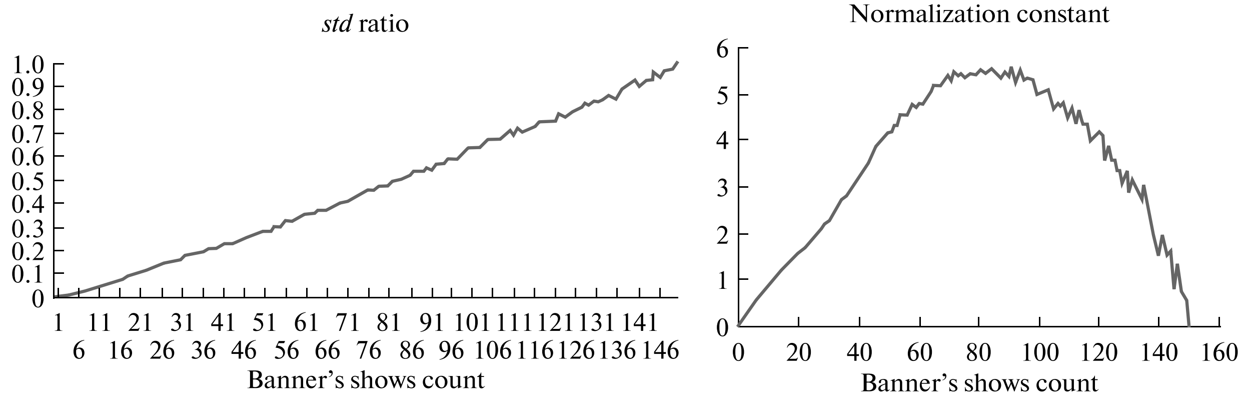

Fig. 3. To the left, ratio $\frac{std_{n_0}}{std_n}$ plot; to the right, shear parameter $\left(1 - \frac{std_{n_0}}{std_n}\right) M_n$.

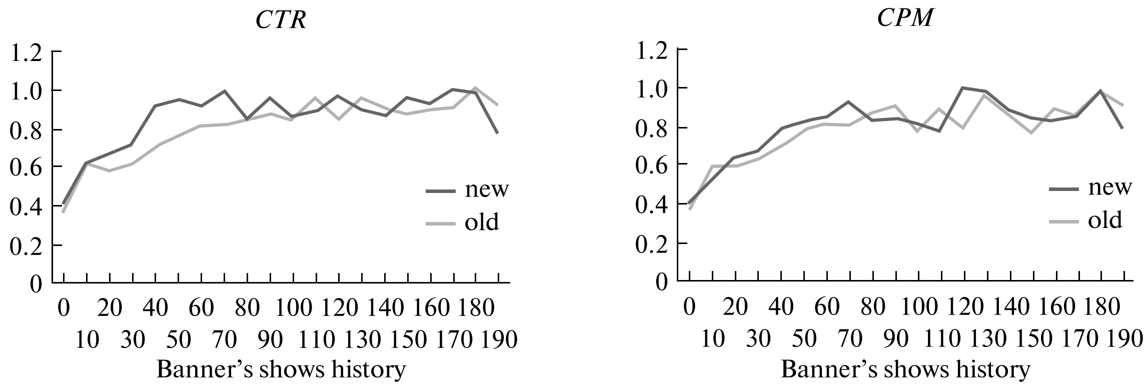

Fig. 4. To the left, *CTR*; to the right, *CPM* (depending on display events in ad statistics).

cable. For each quantity of display events n in statistics, we count an expected quantity of clicks M_n and standard deviation std_n for this quantity of clicks. Then, the wish to avoid high dispersion of the prediction while keeping it unbiased can be expressed using the following constraints:

$$(std_n)_{new} = std_{n_0} \text{ for each } n < n_0.$$

$$(M_n)_{new} = M_{n_0} \text{ for each } n < n_0.$$

In order to satisfy these constraints, the following transformation is sufficient:

$$CTR^*_{new} = \frac{std_{n_0}}{std_n} CTR_{new} + \left(1 - \frac{std_{n_0}}{std_n}\right) M_n, \quad \forall n < n_0.$$

After this transformation, CTR^*_{new} has the specified dispersion level (Fig. 3).

Thus, we reduced the advertiser risks of overpayment per click, which are caused by the high volatility of the initial prediction.

3.1. The Experiment

New formula (1) was tested on previously selected test material and a *CTR* increase by 1.9% was obtained. We believe that such growth is sufficient to carry out an on-line experiment.

According to the on-line experiment, the average *CTR* increase in the system was +1.53%. Figure 4 (at the left) demonstrates the expected improvement of *CTR* to occur over the application region of the new prediction. It is seen in Fig. 4 (at the right) that the new prediction generated no less revenue compared to the current prediction.

Let us consider another important characteristic of the system, namely, *cost-per-click* (*CPC*).

As is seen in Fig. 5, for ads that have a short history the product becomes cheaper and, consequently more attractive. In such a way we are increasing the inflow of clients and providing them with more profitable conditions for ad placement. Similar effects can give positive results in the long-term perspective, i.e., advertising becomes more effective for advertisers, thus stimulating an increase in existing budgets and attracting new resources.

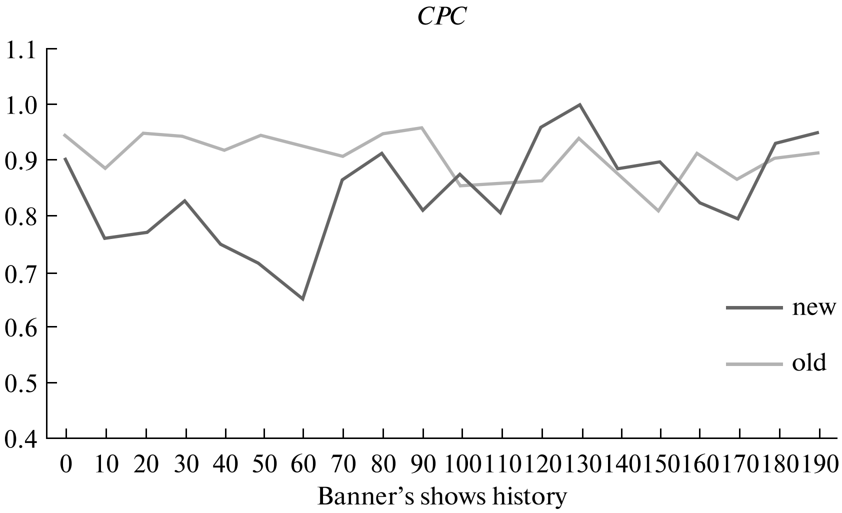

Fig. 5. Dynamics of variation in the average cost per click *CPC*.

CONCLUSIONS

We discussed the problems that occur while optimizing the current system for displaying ads on Yandex.

Primarily, it was noted that for any optimization, the ability to calculate characteristics with optimization at new points should exist. However, launching a new on-line experiment cannot be provided each time because of the natural expensiveness of this approach. An alternative "static" prognosis was used because of the uncertainty that statistics for new ads will predict the click-through rate. As a consequence, the problem of the optimal transition from this initial prediction to the basic one arises.

Prediction properties often vary depending on the volume of accessible statistics. In this connection, for ads with a short history within the frameworks of their advertising companies, undesirable negative effects may occur. As an example, we refer to the problem of impermissibly high prediction volatility.

Methods for solving arising problems are proposed and the result of the operation of an algorithm built for the prediction on a real plot is presented.

Two methods for comparing the quality of algorithm operation with different prediction versions were described in terms of the final characteristics of the system. In this connection, it was proposed to optimize all parameters off-line and only after this to carry out an on-line experiment as the verification of the given approximation.

We described two different approaches to solving the problem of optimal transition from an initial prediction to a basic one. The proposed variant seems to be sufficiently convincing by virtue of the fact that we used a form for its construction that appears to be common for both error square minimization and the Bayesian approach with a wide class of a priori rates for the given task.

The problem of declining prediction volatility was discussed as the problem of reducing undesirable effects. We described the technique for calculating the required prediction characteristics and proposed a linear transformation to solve this problem.

The results of the on-line experiment are presented as the summarization of the work that was performed. It was conducted for comparing the characteristics of an ad-display system that were obtained using the proposed modification of the prediction and the current version. The accuracy of the off-line prediction of improvement of natural characteristics of the system proved to be acceptable, which allows us to consider that the approach was successful.

REFERENCES

1. Ashkan, A., Clarke, C.L.A., Agichtein, E., and Guo, Q., Estimating ad clickthrough rate through query intent analysis, *Proc. 2009 IEEE/WIC/ACM Int. Joint Conf. on Web Intelligence and Intelligent Agent Technol. (wi-iat '09),* 2009, pp. 222–229.
2. Edelman, B., Ostrovsky, M., and Schwarz, M., Internet advertising and the generalized second price auction: selling billions of dollars worth of keywords, *Am. Econ. Rev.* 2007, vol. 97, pp. 242–259.
3. Broder, A. and Josifovski, V., Introduction to Computational Advertising, 2011. http://www.stanford.edu/class/msande239/
4. Dembczynski, K., Kotlowski, W., and Weiss, D., Predicting ads' click-through rate with decision rules, *Proc. Workshop on Targeting and Ranking in Online Advertising*, 2008.
5. Fawcett, T., ROC graphs: notes and practical considerations for researchers, in *HP Labs. Tech. Report*, no. HPL–2003–4.
6. Graepel, T., Candela, J.Q., Borchert, T., and Herbrich, R., Web-scale bayesian click-through rate prediction for sponsored search advertising in microsoft's bing search engine, in *ICMLOmnipress*, 2010, pp. 13–20.

Translated by I. Kekukh