The Role of Relevance in Sponsored Search*

Luca Aiello
Bell Labs
Cambridge, UK
luca.aiello@nokia.com

Ioannis Arapakis
Eurecat
Barcelona, Spain
arapakis.ioannis@gmail.com

Ricardo Baeza-Yates
NTENT
Carlsbad CA, USA
rbaeza@acm.org

Xiao Bai
Yahoo! Research
Sunnyvale CA, USA
xbai@yahoo-inc.com

Nicola Barbieri
Tumblr
New York NY, USA
barbieri@yahoo-inc.com

Amin Mantrach
Yahoo! Research
Sunnyvale CA, USA
amantrac@yahoo-inc.com

Fabrizio Silvestri
Facebook
London, UK
fabrizio.silvestri@gmail.com

ABSTRACT

Sponsored search aims at retrieving the advertisements that in the one hand meet users' intent reflected in their search queries, and in the other hand attract user clicks to generate revenue. Advertisements are typically ranked based on their expected revenue that is computed as the product between their predicted probability of being clicked (i.e., namely clickability) and their advertiser provided bid. The relevance of an advertisement to a user query is implicitly captured by the predicted clickability of the advertisement, assuming that relevant advertisements are more likely to attract user clicks. However, this approach easily biases the ranking toward advertisements having rich click history. This may incorrectly lead to showing irrelevant advertisements whose clickability is not accurately predicted due to lack of click history. Another side effect consists of never giving a chance to new advertisements that may be highly relevant to be printed due to their lack of click history.

To address this problem, we explicitly measure the relevance between an advertisement and a query without relying on the advertisement's click history, and present different ways of leveraging this relevance to improve user search experience without reducing search engine revenue. Specifically, we propose a machine learning approach that solely relies on text-based features to measure the relevance between an advertisement and a query. We discuss how the introduced relevance can be used in four important use cases: pre-filtering of irrelevant advertisements, recovering advertisements with little history, improving clickability prediction, and re-ranking of the advertisements on the final search result page. Offline experiments using large-scale query logs and online A/B tests demonstrate the superiority of the proposed click-oblivious relevance model and the important roles that relevance plays in sponsored search.

CCS Concepts

•Information systems → Sponsored search advertising; Information retrieval; •Applied computing → *Electronic commerce;*

Keywords

Relevance Model, Relevance in Sponsored Search

1. INTRODUCTION

Online advertising, also known as Internet advertising or online marketing, consists in delivering commercial messages (*adverts*) through the Internet to consumers [3]. Online advertising has been used as a marketing mechanism since the 90's and its popularity has grown exponentially in these last years. Online advertising comes under different formats: sponsored search, display advertising, native advertising, etc. The common underlying idea, though, consists in showing advertisements to users while they are visiting pages of an online service, and differently from traditional advertising, the goal is to differentiate and personalize the advertisements shown to each user [1]. Let us consider Facebook as an example, the goal of Facebook advertising is to leverage the huge amount of information about their users in order to get businesses closer to people that might potentially be interested in their activities.

One of the most popular and successful advertisement models is the one implemented by most Web search companies: "*sponsored search*". In sponsored search the objective is to match demand (advertisements) with supply (queries). When a user submits a query, the sponsored search system has the challenge of showing to that user the "best" advertisements possible. As we shall see in the reminder of the paper, in fact, there are many, valid, definitions of "best" advertisements and the majority of them are usually oriented toward returning an advertisement that is both in line with

*This work was done at Yahoo Labs.

CIKM'16 , October 24-28, 2016, Indianapolis, IN, USA

DOI: http://dx.doi.org/10.1145/2983323.2983840

what the user was searching for and profitable for the publisher (in this case the search engine company).

In order to explain how sponsored search works we need to explain how is an advertisement represented in the system and what are the information that an advertiser has to provide to the publisher in order to be considered by the sponsored search system.

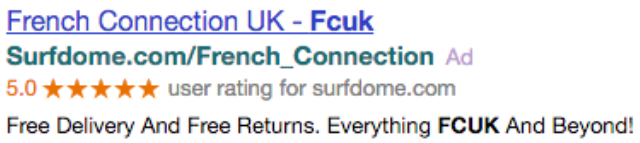


Figure 1: An example of a creative as visualized by a major search engine.

When an advertiser decides to market a service or a product with an online system it has to provide the following:

- *Title and Abstract.* These two pieces of information represent the main message the advertiser wants to convey through the advertisement. In the example in Figure 1 Title and Abstract are "French Connection UK - Fcuk", and "Free Delivery And Free Return. Everything FCUK and Beyond!" respectively

- *A Display and Target URL.* The Target URL is the hyperlink to the landing page that will be shown to the user after she/he will click on the link that has as anchor text the Display URL. In the example in Figure 1, "Surfdome.com/French_Connection" is the Display URL.

- *Bidded Terms or Keywords.* Along with the creative elements the advertiser has also to produce a set of keywords that identify the advertised product. As we shall see bidded terms play a central role in advertising matching, as they are the main signal with which queries, and therefore user intents, are targeted by advertisers.

Along with each bidded terms advertisers have to provide the *matching type* they want to have for that particular bidded term. Matching can be *Exact*, *Phrase*, or *Broad*.

Exact Matching is the most straightforward: given an advertisement and an associated bidded term the advertisement matches the submitted query when it is exactly equal to the bidded term. *Phrase Matching* consists in exactly matching a bidded term with a sub phrase of the query; for instance, within phrase matching the bidded term "tennis shoes" will match the query "how to buy tennis shoes". Finally, *Broad Matching* is the loosest matching type as a query could match all the bidded terms that are in some sense related to that query. As an example, the query "holiday in Europe" could broad match the bidded term "vacation in Italy" as the two concepts are related. Broad matching, of course, offers lots of freedom to advertisers as they only need to specify a concept in order to capture a whole set of queries altogether.

Broad matching is very similar, in principle, to the problem of information retrieval (IR): given a user intent expressed under the form of a text query, retrieve all the related bidded terms and the associated advertisements. In the more traditional IR setting advertisements are then sorted according to the probability of being "relevant" to the query, whereas in sponsored search the sorting has to be done according to the "revenue" for the search engine [11].[1]

If we would only consider revenue as the main factor concerning the ranking of broadly matching advertisements for a given query we would end up always showing the same set of advertisements disregarding the queries and the bidded terms. In essence, "relevance" has to play a role also in the ranking of advertisements.

In fact, the real ranking factor for advertisements in response to a query is typically the "*expected*" revenue that for an advertisement A is given by the following expected revenue per mille impressions (eRPM) formula:

$$\text{eRPM}_A = \mathbb{P}(C = 1|A) \times \text{bid}_A \tag{1}$$

In the above Equation, $\mathbb{P}(C = 1|A)$ represents the probability of clicking on the advertisement A once it has been displayed (i.e., clickability), and bid_A represents the bid associated with the advertisement A.

Clearly, the term $\mathbb{P}(C = 1|A)$ has to be estimated possibly using machine learning methods that also consider factors such as the quality of the matching between the query and the advertisement. However, such methods easily bias the prediction toward advertisements having rich click history. This may incorrectly lead to showing irrelevant advertisements whose clickability is not accurately predicted due to lack of click history. This also risks to never giving a chance to new advertisements that may be highly relevant to be printed due to their lack of click history.

To address this problem, in this paper we consider a particular perspective on the optimization of the quality of the matching between a query and an advertisement. We consider quality from the point of view of relevance. That is, an advertisement should match all those queries for which the user might consider the advertisement itself relevant. While this is the typical approach considered when selecting algorithmic results in web search (i.e., the ten blue links), in the sponsored search scientific literature this has surprisingly not been studied with the necessary depth. For this reason, the present paper tries to fill this gap by defining, experimenting, and discussing how relevance-based scores can be used in a real-world sponsored search system to improve the search user experience. We consider several uses of a relevance score ranging from filtering irrelevant advertisements to using the score to recover advertisements going through different ways of using the relevance score in the click model. Specifically, the contributions of this paper are as follows:

- We propose a supervised machine learning approach that solely relies on text features to predict the relevance between a query and an advertisement, as well as a 6-point editorial guideline that enables fine-grain assessments and thus ensures higher prediction accuracy;

- We show through offline experiments that the proposed relevance model improves the AUC of the state-of-the-art relevance model that uses basic text features and click history by 14.5%;

[1]Indeed, the final ranking is the result of an auction that takes place as the very last phase in the process. For the sake of simplicity we omit details on this phase.

- We propose 4 different use cases where the proposed relevance model can be applied in sponsored search;
- We run several A/B tests in a popular commercial search engine and demonstrate that the proposed relevance model helps improving user experience and search revenue in all the proposed use cases.

The rest of the paper is organized as follows. Section 2 discusses the related work. Section 3 discusses the motivation of our relevance model design and introduces the potential use cases of the model in sponsored search. Section 4 presents our relevance model along with the 6-point editorial guideline. Section 5 discusses the 4 use cases of the proposed relevance model through online A/B tests and offline search log analysis. Section 6 concludes the work.

2. RELATED WORK

Computational advertisement, and more particularly sponsored search, has been a subject of study particularly active since the beginning of the century [13].

A large body of work discussing computational advertising is devoted to finding models and techniques enabling the most precise prediction of clickability of an ad when returned to a user [12, 7, 4, 18, 10]. In this context one of the most recent works is that of McMahan *et al.* [12] where authors show some of the insights gained from developing a large-scale ad serving system in use at Google. In the paper the major focus is on learning how to predict the clickability of an advertisement and, in particular, on developing scalable and effective methods for Google-like systems. On the same line, but for a different type of application, the work of He *et al.* [7] shows how advertisements are selected on a streaming system like that in use at Facebook, while the paper of Graepel *et al.* [4] shows how Bing selects its advertisements to show to users. Due to the pervasive success of deep learning, recent works have detailed how deep systems can be beneficial to sponsored search systems [18, 10, 5, 6]. The paper of Zhai *et al.* [18] exploits a Recurrent Neural Network model to assess the importance of words in advertisements in order to better weight terms in click models for advertisements. Jiang *et al.* [10] propose a deep neural network model integrated with a classical logistic regression model for CTR prediction in contextual advertising. In their model, the deep model is used for automatically extracting abstract and sophisticated features from advertisements content, users' profiles, and clicks. Those features are then used to train a logistic regression model.

Only very few papers have dealt with the study of the impact of relevance on advertisement ranking. The research work that is closer to ours is the one from Hillard *et al.* [8]. As in this paper, Hillard *et al.* develop a machine learning based model to score the relevance between a query and an advertisement. The model uses features including text overlap between query and ads, and past user clicks and exploits a translation model to learn the propensity of users to click on that advertisement when returned for a given query. The results in their paper show that relevance plays a very important role. In particular, when mixed with users feedback signals, (e.g. clicks) relevance quality can be improved considerably. Another paper considering relevance between queries and advertisements is the one from Raghavan and Iyer [15]. In their paper Raghavan and Iyer show a complete implementation of an ad retrieval system using a Language Model (LM) as a first pass to retrieve potentially useful ads. From a modeling perspective they propose an approach that aims at incorporating query segmentation and phrases in the LM framework, discuss impact of score normalization for relevance filtering, and present preliminary results of incorporating query expansions using query-rewriting techniques. Their LM formulation is considerably better in terms of accuracy metrics such as nDCG (about 8% improvement in nDCG@5) on editorial data and also demonstrates significant improvements in clicks in live user tests. A nice feature of that study is that they also show the feasibility of implementing such a system in practice therefore enabling the system to serve millions of users everyday.

3. MOTIVATING A RELEVANCE MODEL IN SPONSORED SEARCH

The role of relevance has been well recognized in web search. However, the role of relevance in sponsored search has been underestimated. As a consequence, modern search engines typically rank the advertisements to a query based on their predicted revenue, which is computed as the product of the predicted clickability of an advertisement and the bid provided by its advertiser. This ranking approach is based on the assumption that user click is highly correlated with ad relevance. That is, advertisement with higher relevance to query is more likely to be clicked than advertisement with lower relevance. Relevance of an advertisement to a query is thus not explicitly captured in ad serving systems.

Although sponsored search is the major source of income for commercial search engines, which somehow justifies the motivation behind the eRPM-based ranking in sponsored search, relevance is still a key factor of user satisfaction. It is thus important for search engines to select advertisements that are relevant to user queries. However, we observe that clickability itself is not enough to represent relevance. As reported in Figure 2, more than 70% of the query-ad pairs are getting less than 10 impressions, which means great majority have no click history exploitable. More interestingly, as developed further (see Section 5.3), according to Hoeffding Inequality, less than 1% of the impressions do have enough click history so that it can be directly exploited as a reliable feature. This is why, unlike the previous work of [8] that model relevance using click history, we focus at leveraging a relevance model that can target the full set of impressions.

Furthermore, when click history is available, many advertisements that have high click through rate (i.e., CTR) are not necessarily relevant to user queries, while other advertisements that have low CTR may be very relevant. Figure 3 shows the correlation between CTR and relevance for query-advertisement pairs estimated on 28K pairs with click history (extracted from the data sets used in this study see Section 4.6). Editors label the relevance of an ad to a giving query by using the 6-scale editorial guideline defined in Section 4.3.

Although perfectly relevant advertisements get the highest CTR values and irrelevant advertisements among the lowest CTR values, nevertheless, the CTR values associated to the other relevance labels scatter from 0 to 1. There is a significant amount of advertisements (i.e., 60%) that are labeled by editors as "Highly Relevant" that does not receive any click (i.e., CTR=0). Similar observation exists for the advertisements labeled as "Relevant". This implies

that if the ad serving system relies on eRPM to rank advertisements, even a model that would perfectly predict the actual CTR of an advertisement, 60% of "Highly Relevant" advertisements may still not get a chance to be presented to users as their eRPM is very low due to their low CTR. Similarly, we observe that 17% of advertisements having relatively high CTR (e.g., between 0.01 and 0.1) are judged as "Irrelevant". This means if the ranking system decides to serve these advertisements given their high eRPM due to their high CTR, they may not lead to any user click, as they are not relevant. Thus, both cases can negatively impact user satisfaction with search engines. Given this observation, we believe that relevance of an advertisement to a query should be modeled separately from clickability, allowing relevance to play a more important role in sponsored search.

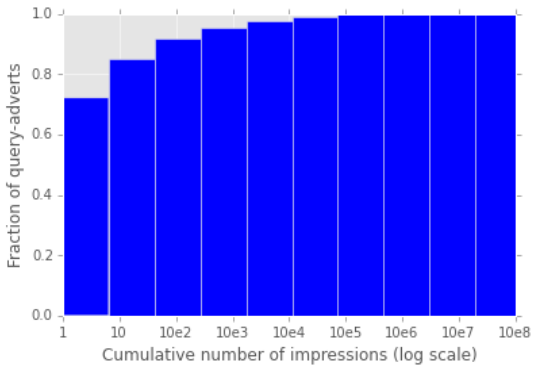


Figure 2: Distribution of impressions and query-advertisements pairs estimated on a random sample of around 1 billion impressions extracted during a period of three months.

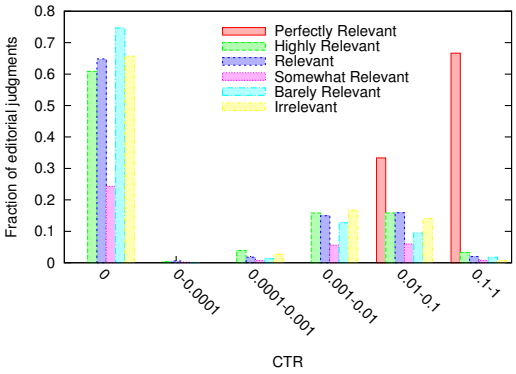


Figure 3: Correlation between CTR and editorially judged relevance.

Motivated by the two observations above, we aim at designing a new model that only depends on text features extracted from query and advertisement to predict query-ad relevance. With this model, we demonstrate the important roles a relevance model can play in sponsored search. We describe these roles of relevance through 4 applications:

- Filtering irrelevant advertisements: This prevents advertisements having low relevance scores to be passed to the ranking phase of the ad serving system. This will on the one hand reduce the operational cost of the system, and on the other hand improve user search experience by not showing irrelevant advertisements.

- Improving click model: Relevance is a strong indicator of click. If relevance score is leveraged as a feature in the click model that predicts the clickability of a query-advertisement pair, the accuracy of clickability prediction should increase, which will in turn improve user satisfaction during search and lead to more clicks on advertisements.

- Recovering filtered cold advertisements: Advertisements with low predicted clickability may not be ranked high enough to be presented to users. Cold advertisementisements that have little click history may not get an accurate predicted clickability, and thus may be wrongly filtered in the ranking phase. Relevance can be used in this scenario to recover the cold ads that get filtered by clickability but are very relevant to user query to improve the overall quality of the served advertisements.

- Reranking advertisements using relevance: Relevance can also be used in the ranking function to adjust the pure eRPM-based ranking. Again, as relevant advertisements are more likely to attract clicks, explicitly leveraging relevance in ranking may help improving user experience by showing relevant advertisements at better positions of search result pages.

4. AD RELEVANCE MODELING

Unlike clickability, which to some extent, indicates whether an ad is interesting or attractive, relevance can capture a different set of ad qualities such as the underlying intent (as relation, context, inference, and interaction) or its topical and situational connection to the associated search query [16]. These qualities may be equally or more important factors to consider when addressing the problem of sponsored search. However, the editorial control of advertisements is a difficult task that requires manual effort and it is hardly scalable. This warrants the research of appropriate benchmarks and proxies of ad relevance (e.g., global and local textual features) using machine learning methods that are cost-effective and can scale. In what follows, we discuss an approach to predicting ad relevance by modeling the quality assessment and selection process applied by professional editors.

4.1 Dataset

Our study is conducted on a search log collection consisting of query-advertisement pairs taken from a popular commercial search engine. More specifically, we randomly sample an equivalent number of search queries from each decile bucket from the query frequency distribution. Then, for each sampled query we retrieve the top advertisement candidates obtained by various matching algorithms that produce query-advertisement pairs on the serving platform (i.e. Exact and Broad matching as presented in Section 1). This results in a sample of about 1.2K unique queries and 81K unique text advertisements that forms a final collection of 170K query-ad pairs. Figure 4 shows the distribution of number advertisements per query.

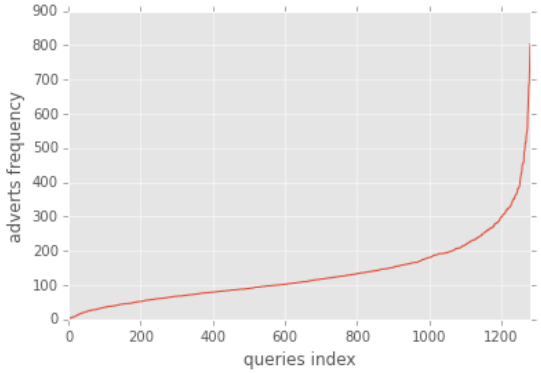


Figure 4: Advertisements per query distribution.

4.2 Participants

To characterize ad relevance, we rely on the domain knowledge and human intuition of expert judges whom we employ in a rigorous, crowdsourcing-based evaluation for generating a labeled dataset. For our editorial study, we employed 45 expert judges (male = 9, female = 36) whose age ranged from 20 to 50 and who, in their majority, had a background in linguistics, journalism, psychology, or computer science. The expert judges are either native English speakers (42.2%), are proficient with the English language (46.7%), or have an advanced level (11.1%).

4.3 Editorial Guidelines

For assessing ad relevance, we define a set of guidelines used to annotate query-advertisement pairs with one of the following categories:

Perfect Match: the advertisement captures exactly the intent of the query; the query itself must have a very specific intent that is answerable by an official vendor or official web page.

Highly Relevant: there is no constraint on the query (e.g. to be navigational) but you anticipate that clicking on the ad will take the issuer of the query to a very useful page.

Relevant: the advertisement has good topical match with the original query; although it does not align perfectly with the query intent, it is judged as relevant.

Somewhat Relevant: the advertisement has a small degree of relevance to the query; clicking on the ad will most likely take the issuer of the query to a somewhat useful page.

Barely Relevant: the advertisement does not make any useful promise for the query in question; however, the vendor may possibly have products related to the query.

Irrelevant: the query and the advertisement are clearly mismatched and have nothing to do with one another, or the vendor cannot possibly offer what they claim they can.

Our relevance assessment report consists of a 6-point relevance scale, where low and high scores suggest weak or strong relevance respectively. The advantages of this approach is that it gives the editors more scope to express how they feel about a query-advertisement example and are easily understood. While binary assessments may result in higher inter-rater agreement and are easier to complete, the Likert-type scale, when clearly-defined, can show good agreement and better construct validity, which can be useful when learning relevance models. In addition, our Likert-type scale reduces the risk of observing biased responses due to the limited options. If there are not enough response options (e.g., as in binary judgments or scales with very few items) editors will be forced to choose the next best alternative and this introduces a systematic measurement error. Most importantly, our Likert-type scale allows for a better performance assessment of our models. From a business perspective, filtering out "Barely Relevant" matches is less costly and more acceptable than filtering out "Somewhat Relevant" matches. Similar, retrieving a "Perfect Match" is more desirable than retrieving a "Highly Relevant". With the proposed approach, our models can account for this difference and weight differently the training samples.

Finally, when assessing the ad relevance, the editors were instructed to take into account all three components of an advertisement, namely: (1) title, (2) abstract, and (3) display URL (without clicking on the link to check the relevance of the landing page). Even if some of the ad components did not provide an indication that the advertisement is relevant, the editors were asked to consider the remaining components and whether they provide some relevant context (Figure 1).

4.4 Procedure

Prior to the study, there was a pilot session where each expert judge was asked to become familiar with the relevance criteria and annotate several trial query-advertisement pairs. Next, a meeting (physical or online) was arranged and the authors discussed with the expert judge the rationale behind assigning the scores, and appropriate corrections and recommendations were made. This step ensured that we had disambiguated any question prior to the editorial study and also assured that the expert judges followed the same scoring procedure. The annotation took place remotely, and each expert judge could annotate between 150-300 query-advertisement pairs per day.[2] Finally, one expert judge annotated each query-ad pair. In total, the expert judges assessed a total of 170K query-ad pairs using the 6-point Likert scale described in the previous section.

4.5 Feature Engineering

We leverage different features that we think should be able to capture part of the pairwise relationship existing between a query and the three different signals available in the textual advertisement, that is the title, abstract, and display URL. In [14] the authors introduce 19 features, namely the query length, and 6×3 features for each zone the ad namely word overlap (unigram and bigram), character overlap (unigram and bigram), cosine similarity, and ordered bigram overlap. In this work, inspired by the web search literature, we introduce a set of 185 features:

Common counts (12 features): This set of features counts the number of common unigrams, bigrams and q-grams (i.e. text substrings of length q = 4) between the query and each advertisement component. Note that we are considering an additional advertisement component that is the concatenation of the three advertisement components;

Jaccard (12 features): This set of features computes the Jaccard similarity between unigrams, bigrams and q-grams between the query and each advertisement component;

Length (10 features): This set of features counts the number of unigrams and the number of characters of query and each advertisement component;

[2]The threshold of 150-300 judgments per editor per day was set to ensure a high quality of annotation.

Algorithm 1 Computation of Hash Embedding Features

Input: query q, advertisement a ($title\_abstract\_displayURL$), and space length k
Output: hash embedding h[k]

for each t_q in q **do**
  for each t_a in a **do**
    $index = hash(t_q, t_a) \bmod k$
    if $index < 0$ **then**
      $index \leftarrow index + k$
    end if
    $weight = hash(t_q, t_a)$
    if $weight > 0$ **then**
      $weight \leftarrow 1$
    else
      $weight \leftarrow -1$
    end if
    $h[index] \leftarrow h[index] + weight$
  end for
end for
return $h[k]$

Cosine (4 features): This set of features computes the cosine similarity between the query vector and each advertisement component vector. Each vector is represented as the TF-IDF of its unigrams;

BM25 (4 features): This set of features computes the BM25 similarity between the query vector and each advertisement component vector;

Brand (8 features): This set of features computes, based on a brand dictionary built offline, the number of common brands, and the Jaccard brand similarity between query and each advertisement component;

LSI (4 features): This set of features computes the cosine similarity between the LSI query representation and each LSI advertisement component representation, where a vector of 100 dimensions represents each term.. These representations are build offline on a large data set of textual advertisements. Specifically, the term-document (i.e., concatenation of the title, abstract and display URL of advertisement) matrix is built using terms that appear at least 15 times in a random sample of 5 million unique advertisements;

Semantic coherence (3 features): This set of features compute statistics (i.e, max, min and mean) of the LSI cosine similarities between the three components of an advertisement (i.e. title, abstract and display URL). Intuitively, to be relevant an advertisement should have a description, and a display URL that are coherent with its title; and

Hash embedding (128 features): This set of features computes, for each pair of unigrams that contains one unigram from the query and one unigram from the concatenation of the three advertisement components, an h-index. The h-index corresponds to the hash encoding of the concatenated unigrams modulo the space length k. Algorithm 1 illustrates how the hashing embedding features are computed. This set features is inspired from the feature hashing trick introduced in [17], and adapted in our case to pairwise features. The space length k is fixed to 128 in this work.

4.6 Model Validation

We argue that by building a predictive model on a much richer text-based feature set we cannot only obtain a more accurate model in terms of AUC, but also can avoid making use of the click history associated to the advertisement. Indeed, unlike [8], we argue that using the click history does not bring any additional improvement if the click-oblivious text-based features are well selected. To validate our hypothesis, we sample a set of query-advertisement pairs that have more than 5000 impressions during a period of one month, from the 170K query-advertisement pairs used in our study (as described in Section 4.1), to extract their click-through rate values. This set contains 28K query-advertisement pairs, for which we also compute the 19 basic text features as described in [14], and the 185 features we propose in Section 4.5. Our objective is to train a classifier that can distinguish irrelevant advertisements from the rest of the advertisements. Hence, the final editorial relevance score is mapped to a binary score, -1 if annotated as Bad, and +1 otherwise. Nevertheless, we still make use of the editorial scale by weighting the samples using their editorial score. We test two state-of-the-art binary classifier models: logistic regression and random forest. We evaluate their performance using an adapted 10-fold cross validation. Note that we ensure that a query in the test set never appears in the training set to avoid any overfitting at the query level.

Table 1: Averaged 10-fold cross validation AUC of Logistic Regression (LR), and Random Forest (RF) models trained on 28K editorially annotated query-advertisement pairs.

Model	19 features [14]	19 features + CTR [8]	185 features
LR	0.569	0.573	0.651
RF	0.566	0.595	0.675

We test the proposed 185 text-based features on the task of relevance prediction against the 19 text-based features baseline ([14]), and the 19 text-based features supplemented with the click history feature ([8]). As reported in Table 1, although using click through rate does improves the performance of the basic baseline models trained with 19 text-based features, the models trained with our 185 text-based features still outperform the baseline models that uses 19 features and click feature by 13.5%. Notice that we can slightly improve our feature set by making use of the click history. However as pointed out in [8], adding click history in the relevance model lead to large restrictions on the applicability of the model. Indeed, by making use of click history, the model cannot be applied for filtering not-yet-seen advertisements (Section 5.1), and for recovering cold advertisements that do not have click history but have high relevance (see Section 5.3). Actually, all challenging use cases, where click history is not available cannot be targeted. These challenging problems are exactly the ones we target in the remainder of this paper.

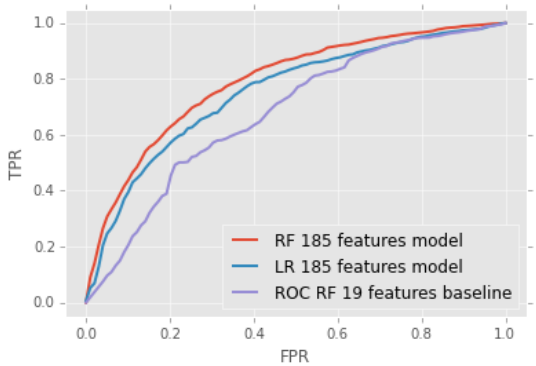
In Table 2, we report the 10 most important features as well as the most important features from each feature group as described in Section 4.5. We notice that apart from the most important feature, i.e. the cosine similarity that is also part of the 19 basic features [14], Jaccard, BM25, LSI and Semantic coherence are all very important features that were not previously considered. Moreover, the features computed based on q-grams are among the top-10 important features and are more important than the ones from the same group but computed based on unigrams or bigrams. These advanced text-based features are the key to the performance improvements reported in Table 1.

Table 2: Feature importance of the Random Forest relevance model.

1	COSINE_TITLE	1.000
2	Q_GRAMS_JACCARD_ALL	0.987
3	LSI_URL	0.983
4	Q_GRAMS_JACCARD_TITLE	0.964
5	LSI_TITLE	0.958
6	BM25_TITLE	0.956
7	Q_GRAMS_COUNT_ALL	0.849
8	BM25_ALL	0.797
9	LSI_ALL	0.763
10	SEMANTIC_COHERENCE_AVG	0.713
13	LSI_DESCRIPTION	0.679
14	NUMBER_CHARS_TITLE	0.670
27	NUMBER_UNIGRAMS_ALL	0.428
39	BRANDS_JACCARD_ALL	0.167
40	HASH_EMBEDDING_15	0.163

5. EXPERIMENTS WITH RELEVANCE

In this section, we discuss four applications of the proposed relevance model in sponsored search. We train a Random Forest model (RF) with the 185 text-based features (Section 4.5) using the 170K editorially judged query-advertisement pairs (Section 4.1). We also compare its performance against a Logistic Regression model (LR) trained using the same features and same dataset, and the baseline model using 19 text-based features [14] and the same dataset. The three algorithms are tuned on an independent validation set. Figure 5 shows the ROC curves of the three models, as well as their corresponding AUC values. We observe that the Random Forest model using 185 text-based features improves the AUC of the baseline approach using 19 text-based features by 14.5%. As the Random Forest model results in the best performance, we rely on it in the rest of this section to discuss the applications of relevance model in sponsored search.



(a) ROC curves

Model	*avg* AUC
RF	0.764
LR	0.741
19-feature baseline [14]	0.667

(b) Average AUC

Figure 5: Performance of relevance model.

5.1 Filtering Irrelevant Advertisements

One straightforward application of a relevance model consist of pre-filtering (i.e. before auctions) query-advertisements candidates. By so doing, we aim at improving the user experience by filtering irrelevant advertisements. This also helps to reduce the amount of candidates that have to be processed by the following modules in the advertisement serving system and thus improves the overall efficiency of the system.

In this scenario, the objective is to filter as much irrelevant advertisements as possible while making sure only a limited number of relevant advertisements[3] are filtered in the process. To this end, we select the threshold to filter irrelevant advertisements with regards to the acceptable false negative (FN) rate (i.e. fraction of relevant advertisements filtered by mistake). This choice is motivated by the fact that relevant advertisements are those attracting user clicks and are thus important for both user experience and search revenue.

In this experiment, we set the threshold for each model in such a way that at most 2% relevant get filtered. Table 3 reports the detailed performance of our Random Forest model, the Logistic regression and the 19-feature baseline with respect to each relevance category. Interestingly, we observe that for all the three models, the majority of the 2% query-advertisement pairs that were wrongly filtered are actually labeled as "Barely Relevant" and "Somewhat Relevant", and no "Perfectly Relevant" advertisements get filtered. This conveys the benefits of working with a finer grade relevance guideline (Section 4.3), especially in the irrelevant region.

We then run an A/B test in the search engine from which we sample our training data, using the Random Forest model to assess the impact of the relevance-based filtering on ad coverage, ad CTR and search revenue (i.e., RPM). Ad coverage measures the fraction of queries that can get at least one advertisement shown in the north of their search result pages. The control and test buckets both represent 5% of the overall query traffic of the search engine and were running for a week. We report in Table 3 the changes in the three metrics of the test bucket compared to the control bucket. We observe that we can boost the CTR of Exact match and Broad match by 4.07% and 2.71%, respectively, for a decrease of about 5% of ad coverage. Interestingly, the impact on revenue is only about 1.71% for Exact match, and there is almost no negative impact on Broad match. This reveals that we have room to increase the filtering threshold for Broad match in order to get more irrelevant advertisements filtered. All the numbers reported in Table 3 are statistically significant according to a Student's t-test with p-value lower than 0.05. This experiment confirms that using relevance as a filter can help search engine to show fewer irrelevant advertisements and improve user engagement without losing much its revenue.

5.2 Improving the Click Model

In this section, we hypothesize that relevance can help to improve the accuracy of the click model. The underlying intuition is that relevant advertisements are more likely to attract user clicks. Moreover, when an advertisement has little click history, it is difficult for the click model to accurately predict its probability of being clicked given a query

[3]Here we use the term "relevant" to refer to any advertisement that is not "Irrelevant" as defined in Section 4.3.

Table 3: % of filtered advertisements in each editorial category. Thresholds set to have at most 2% of relevant advertisements filtered.

Model	Irrelevant	Barely Relevant	Somewhat Relevant	Relevant	Highly Relevant	Perfectly Relevant
RF	12.47	1.32	0.8	0.07	.0.01	0
LR	7.78	1.06	0.72	0.18	.0.09	0
Baseline	7.72	1.19	0.65	0.14	.0.05	0

Table 4: Relevance as a filter bucket. The differences reported are statistically significant according to a t-test with p-value lower than 0.05.

Metric	Relative change w.r.t. control bucket	
	Exact match	Broad match
Ad coverage	-5.07%	-5.21%
CTR	+4.07%	+2.71%
Revenue	-1.71%	-0.00%

Table 5: Relevance as click model feature bucket. The differences that are statistically significant according to a t-test of p-value lower than 0.05 are marked with *.

Metric	Relative change w.r.t. control bucket
Ad coverage	+1.45%*
Click-yield	+3.02%*
CTR	+2.09%*
RPM	+0.52%

(i.e., clickability). However, as relevance is independent of click history, it may be helpful for the click model to improve its prediction accuracy for such advertisements.

To validate this hypothesis, we use the predicted relevance score as a feature with all the other features originally used in the click model to train a new click model. We then run an A/B test in the same search engine to compare the performance of the new click model and that of the original click model. The control and test buckets both represent 5% of the overall query traffic of the search engine and were running for a week. Table 5 reports the change of the key performance metrics of the test bucket (i.e., new click model having relevance score as feature) compared to the control bucket (i.e., original click model without relevance score as a feature).

We observe that when the new model is used, 1.45% more queries are served with at least one advertisement in the north of their search result pages (i.e., Ad coverage) and 3.02% more advertisements shown in the north of a search result page are clicked (i.e., Click-yield). This leads to a CTR increase of 2.09%. These numbers are statistically significant according to a Student's t-test with p-value lower than 0.05. This confirms that using relevance as a feature helps the click model to improve the accuracy of its prediction by serving advertisements to users that are more likely to attract clicks. We also observe a slight increase of revenue per mille when the new click model is used. Although this number is not statistically significant, it still shows that it is promising to leverage relevance in the click model to improve user satisfaction with the search engine as well as search revenue.

5.3 Recovering Filtered Cold advertisements

As we have described above in Section 5.1, one of the possible uses of relevance is to filter out advertisements that are not retained to be relevant to users. This is clearly the most straightforward way to exploit the relevance score between a query and an advertisement. A different possibility is to use relevance to recover query-advertisement pairs that have been incorrectly filtered out for reasons different from relevance score being too low. Filtering errors of query-advertisement pairs can happen in many occasions. One notable example of such cases is when a query-advertisement pair has a very low probability of being clicked (i.e., clickability). However, there are cases where the reliability of the click model is not very high as shown in Figure 2.

In such cases we may risk to filter out important query-advertisement pairs that are very relevant to user queries.

To address this problem, we propose to rely on relevance to recover the advertisements that are likely to be not filtered. Specifically, we first check whether the amount of "*historical*" information about a query-advertisement pair is enough to trust the click model. If this is not the case, we check whether the relevance between the query and the advertisement is above a pre-determined threshold. If the relevance is higher than the threshold, we "recover" the advertisement by moving it from the filtered list to the candidate list that contains the advertisements to be ranked by the system.

In order to take the steps we have just described, we need to determine: i) what is the right amount of historical information we need to trust the click model, and ii) what is the right threshold for the relevance score to recover an advertisement.

Regarding to the first problem, we consider clickability as a probability density function of a Bernoulli variable X, which is equal to 1 when the user clicks on the advertisement and 0 otherwise. We then resort to use Hoeffding Inequality[4] [9] to compute the sample size we need to make sure that clickability (that we consider a rare event) can be reliably estimated at a $(1-\alpha) = 0.95$ confidence level and within an additive error $\epsilon = 0.005$. Sample size, in this particular case, represents the number of impressions we need to consider to make sure we are confident to be within the requested error range. By applying the aforementioned bound we have that we need at least $n = 73,788$ impressions as given by the following inequality:

$$n \geq ln\left(\frac{2}{\alpha}\right)\frac{1}{2\epsilon^2} \tag{2}$$

Therefore, whenever the predicted value for CTR is obtained for samples observed less than n times we resort to use the relevance score to decide whether to recover the advertisement or not.

This leads us to the second problem: what is the right threshold for the relevance score in order to consider the advertisement for the recovery? We consider, also in this

[4]Hoeffding Inequality is very conservative and the lower bound on the sample size computed using this inequality is not tight. For the sake of this study we can rely on it.

Table 6: Relevance as relevant advertisement recoverer bucket. The differences that are statistically significant according to a t-test of p-value lower than 0.05 are marked with *.

Metric	Relative change w.r.t. control bucket
Ad coverage	+1.02%*
Click yield	+1.40%*
CTR	+0.06%
RPM	+3.70%*

case, a very conservative argument consisting in selecting the relevance score threshold that will cause the recovery of 10% of advertisements that were filtered out because of low clickability. According to this criterion we set the relevance score threshold to 0.65 that corresponded to recovering about 9.65% of impressions.

We test online, through an A/B test, the effect of the recovery strategy described above. We direct 5% of the traffic to our test treatment and another 5% to the control treatment during 11 days. Table 6 reports the key metrics improvements of the test bucket compared to the control bucket. Results show that critical metrics go up. In particular, the click-yield metric (the number of queries having at least a click on an advertisement) increased by 1.40% relative to control and, more importantly, the revenue increased by 3.71% relative to control. The impact on users is also impressive if we consider that 12% of the impressions are generated by the recovered advertisements. Finally, it is worth noticing that the impacted advertisers are advertisers whose quality is generally high and who usually post the large amount of new advertisements frequently. On those cases, relying only on historical information for those advertisements, thus, can be limiting because of the limited amount of historical information on which the click model can be trained. The relevance score does not make use of any historical data thus is not affected by this kind of cold-start problem.

5.4 Reranking Advertisements Using Relevance

The advertisements that pass the filters or get recovered by relevance are typically ranked by the expected revenue per mille (eRPM, Formula (1)) before serving to users. Another possible way of using relevance is to leverage relevance score in the ranking function. The objective is to improve user perceived relevance of the served advertisements while generating high revenue for the search engine.

We rely on a basic strategy to leverage relevance score in the ranking function. Thus, the relevance score is used to weight the expected revenue per mille for each ad. Intuitively, advertisement with higher relevance is more likely to attract click and thus its expected revenue per mille is more likely to be accurate. Formally, we define the relevance-boosted ranking score as follows:

$$\begin{aligned} \text{eRPM}_A^R &= \mathbb{P}(R=1|A) \times \text{eRPM}_A \\ &= \mathbb{P}(R=1|A) \times \mathbb{P}(C=1|A) \times \text{bid}_A \end{aligned} \quad (3)$$

We re-rank the advertisements shown in the north of each search result page using the relevance-boosted ranking score. We use the Normalized Discounted Cumulative Gain (NDCG) metric to evaluate the quality of a ranking. For ranking r of n north advertisements, we define $rel_i(A)$ of the advertisement A at position i ($1 \leq i \leq n$) of ranking r as 1 if A is clicked and as 0 otherwise. The Discounted Cumulative Gain (DCG) of ranking r is computed as $DCG(r) = \sum_{i=1}^{N} rel_i(A)$. Ideally, advertisements in a search result page that are clicked by user should appear on top of the ranked list. Given the ideal ranking r^* that has all the clicked advertisements ranked in the top positions,[5] the NDCG of a ranking r is computed as

$$NDCG(r) = \frac{DCG(r)}{DCG(r^*)} \quad (4)$$

Table 7: Quality of relevance-boosted ranking.

	Average NDCG
Original eRPM-based Ranking	0.858
Relevance-boosted Ranking	0.870
Relevance-based Ranking	0.819

We compare the relevance-boosted ranking against the original eRPM-based ranking and a relevance-based ranking that ranks the advertisements solely using their relevance scores.

We evaluate the performance of the relevance-boosted ranking using a random sample of 2 millions advertisement rankings. Each advertisement ranking contains at least two ads appearing in the north of the corresponding search result page and at least one of them is clicked.

Table 7 shows the average NCDG values over all the 2 million rankings for the three ranking approaches. We observe that leveraging relevance score in the original eRPM-based ranking improves the average NCDG by 1.40%. We also observe that using relevance score alone to rank advertisements results in much lower average NDCG. This reveals that advertisements relevant to user query in a broad sense (i.e., having relatively low relevance score) may still attract user clicks, confirming the rationality of eRPM-based ranking. Yet, incorporating relevance score into the ranking function clearly helps boosting the clicked ads in higher positions of search result pages and thus improving user experience.

Figure 6 shows the distribution of NDCG values for the 2 million rankings in our experiments. Since no more than 4 advertisements are usually shown in the north of a search result page, and we consider rankings with at least 2 advertisements, we only observe 5 different NCDG values. Compared to the original eRPM-based ranking, relevance-boosted ranking is able to bring more advertisements that attract user clicks to higher positions and to boost the corresponding NDCG values.

To test the statistical significance of the difference between the NDCG values of the relevance-boosted ranking and the original eRPM-based ranking, we use Wilcoxon signed-rank test [2] in our experiments. This is because the NDCG values do not follow a Gaussian distribution required by Student's t-test as shown in Figure 6. Table 8 reports the p-value of the statistical significance test as well as the fraction of rankings that are different from its original eRPM-based ranking after applying the relevance-boosted ranking or the relevance-based ranking. This result confirms the improvement in NDCG by leveraging relevance score in the original eRPM ranking is statistically significant, and relevance can be used to improve traditional eRPM-based advertisement ranking.

[5]In case of more than one advertisements are clicked, the advertisement generating more revenue is ranked higher.

Table 8: Difference between relevance-boosted ranking and original eRPM-based ranking.

	p-value	% different rankings
Relevance-boosted Ranking	<0.01	15.39%
Relevance-based Ranking	<0.01	51.71%

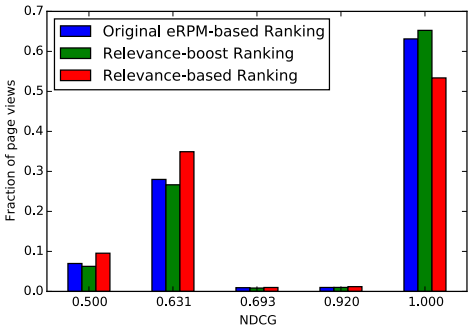


Figure 6: NDCG distribution.

6. CONCLUSIONS

We present in this paper a new relevance model that can accurately predict query-advertisement relevance only using text-based features. We show through experiments on large-scale datasets that using advanced text features outperforms the state-of-the-art relevance model that uses basic text features and click through rate. More importantly, our model is flexible to be applied for any query-advertisement pair that has little or even no click history. Furthermore, we explore four important application scenarios of the proposed relevance model in sponsored search, i.e., irrelevant advertisements filtering, click model improving, relevant cold advertisements recovery, and relevance-boosted advertisement reranking. We demonstrate through either offline experiment or online A/B test that the proposed relevance model can help a popular commercial search engine to significantly improve its user experience and its search revenue.

7. REFERENCES

[1] K. Dave and V. Varma. Computational advertising: Techniques for targeting relevant ads. *Found. Trends Inf. Retr.*, 8(4-5):263–418, Oct. 2014.

[2] J. Demšar. Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res. (JMLR)*, 7:1–30, 2006.

[3] P. Gill, V. Erramilli, A. Chaintreau, B. Krishnamurthy, K. Papagiannaki, and P. Rodriguez. Follow the money: Understanding economics of online aggregation and advertising. In *Proceedings of the 2013 Conference on Internet Measurement Conference*, pages 141–148, 2013.

[4] T. Graepel, J. Q. Candela, T. Borchert, and R. Herbrich. Web-scale bayesian click-through rate prediction for sponsored search advertising in microsoft's bing search engine. In *Proceedings of the 27th International Conference on Machine Learning*, pages 13–20, 2010.

[5] M. Grbovic, N. Djuric, V. Radosavljevic, F. Silvestri, R. Baeza-Yates, A. Feng, E. Ordentlich, L. Yang, and G. Owens. Scalable semantic matching of queries to ads in sponsored search advertising. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 375–384, 2016.

[6] M. Grbovic, N. Djuric, V. Radosavljevic, F. Silvestri, and N. Bhamidipati. Context- and content-aware embeddings for query rewriting in sponsored search. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 383–392, 2015.

[7] X. He, J. Pan, O. Jin, T. Xu, B. Liu, T. Xu, Y. Shi, A. Atallah, R. Herbrich, S. Bowers, et al. Practical lessons from predicting clicks on ads at facebook. In *Proceedings of the Eighth International Workshop on Data Mining for Online Advertising*, pages 1–9, 2014.

[8] D. Hillard, S. Schroedl, E. Manavoglu, H. Raghavan, and C. Leggetter. Improving ad relevance in sponsored search. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, pages 361–370, 2010.

[9] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

[10] Z. Jiang, S. Gao, and W. Dai. Research on ctr prediction for contextual advertising based on deep architecture model. *Journal of Control Engineering and Applied Informatics*, 18(1):11–19, 2016.

[11] S. Lahaie, D. M. Pennock, A. Saberi, and R. V. Vohra. Sponsored search auctions. *Algorithmic game theory*, pages 699–716, 2007.

[12] H. B. McMahan, G. Holt, D. Sculley, M. Young, D. Ebner, J. Grady, L. Nie, T. Phillips, E. Davydov, D. Golovin, et al. Ad click prediction: a view from the trenches. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1222–1230, 2013.

[13] A. Mehta, A. Saberi, U. Vazirani, and V. Vazirani. Adwords and generalized on-line matching. In *Proceedings of the 46th Annual IEEE Symposium on Foundations of Computer Science*, pages 264–273, 2005.

[14] H. Raghavan and D. Hillard. A relevance model based filter for improving ad quality. In *Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 762–763, 2009.

[15] H. Raghavan and R. Iyer. Probabilistic first pass retrieval for search advertising: From theory to practice. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, pages 1019–1028, 2010.

[16] T. Saracevic. Relevance reconsidered. In *Proceedings of the Second Conference on Conception of Library and Information Science*, pages 201–218, 1996.

[17] K. Weinberger, A. Dasgupta, J. Langford, A. Smola, and J. Attenberg. Feature hashing for large scale multitask learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1113–1120, 2009.

[18] S. Zhai, K.-h. Chang, R. Zhang, and Z. Zhang. Attention based recurrent neural networks for online advertising. In *Proceedings of the 25th International Conference Companion on World Wide Web*, pages 141–142, 2016.