# Interpreting Advertiser Intent in Sponsored Search

Bhanu C. Vattikonda, Santhosh Kodipaka*, Hongyan Zhou*,
Vacha Dave*, Saikat Guha†, and Alex C. Snoeren
University of California, San Diego
*Microsoft, Redmond, Washington,
†Microsoft Research, Bangalore, India

## ABSTRACT

Search engines derive revenue by displaying sponsored results along with organic results in response to user queries. In general, search engines run a per-query, on-line auction amongst interested advertisers to select sponsored results to display. In doing so, they must carefully balance the revenue derived from sponsored results against potential degradation in user experience due to less-relevant results. Hence, major search engines attempt to analyze the relevance of potential sponsored results to the user's query using supervised learning algorithms. Past work has employed a bag-of-words approach using features extracted from both the query and potential sponsored result to train the ranker.

We show that using features that capture the advertiser's intent can significantly improve the performance of relevance ranking. In particular, we consider the ad keyword the advertiser submits as part of the auction process as a direct expression of intent. We leverage the search engine itself to interpret the ad keyword by submitting the ad keyword as an independent query and incorporating the results as features when determining the relevance of the advertiser's sponsored result to the user's original query. We achieve $43.2\%$ improvement in precision-recall AUC over the best previously published baseline and $2.7\%$ improvement in the production system of a large search engine.

## Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Relevance feedback; I.5.4 [**Applications**]: Text processing

## General Terms

Algorithms, Supervised learning, Experimentation

## Keywords

Sponsored search; ad relevance; online advertising


*KDD '15,* August 11–14, 2015, Sydney, NSW, Australia

ACM 978-1-4503-3664-2/15/08$15.00
DOI: http://dx.doi.org/10.1145/2783258.2788566.

## 1. INTRODUCTION

Search engines deliver two types of results in response to a user query: organic and sponsored search results. Organic results are identified from around the web based on their relevance to the query. Sponsored results—i.e., advertisements—on the other hand, are selected from submissions made by advertisers interested in attracting the user. These additional results are chosen based on a combination of relevance of the ad to the query and the expected revenue the search engine will derive from displaying the ad [22].

When providing advertisements to the search engine, an advertiser specifies for which user queries, e.g., "shirts", "polos", "jersey", etc. it would like to show each ad. These words, referred to as ad keywords, are the basic mechanism through which advertisers can target their ads. Often ads are shown to users when the ad keyword matches the query—resulting in what is known as an exact match. But, it is challenging for an advertiser to enumerate all the queries for which they would like to advertise. Hence, many search engines provide an option of matching the ad keywords with a broader range of relevant user queries, e.g., alternative spellings, synonyms, etc., potentially resulting in what we term a broad match. Typically, when an ad is chosen and displayed to the user, the advertiser only pays the search engine if the user clicks on the ad, incentivizing the search engine to keep even broad matches relevant to the user's query.

Yet sponsored results are generally perceived to degrade user experience on search engines [23]. An aggressive pursuit to maximize revenue from each search impression could therefore hurt the user experience—and, in the long run, the search engine's popularity and profitability. Thus, in some cases, it may be desirable to show few or even no ads if they do not meet some minimum relevance threshold [9]. For example, if the query is "weather", an ad for "cold weather jackets" might occasionally generate revenue, it is not likely if the user is simply seeking the current temperature—and especially if the temperature is currently warm. In this case not showing the ad would be the prudent choice. Another canonical example where sponsored results are frequently ill advised is a so-called navigational query like "Macys", where ads other than those from the retail chain Macy's could elicit a negative user response.

Hence, one of the key challenges for search engines is to quantify the relevance of ads to a user's query. Our partner search engine considers an ad relevant to the query if the following four components are aligned [5]: *i*) query; i.e., what the user is looking for, *ii*) ad creative; what is being promised to the user, *iii*) ad landing page; the web page actually delivered to the user if the ad is clicked, and *iv*) ad keyword; which indicates the type of traffic the advertiser seeks to attract.

Interpreting a user's query is hard mainly because it is very short: 2.5-words long on average [33]. Over time search engines have in-

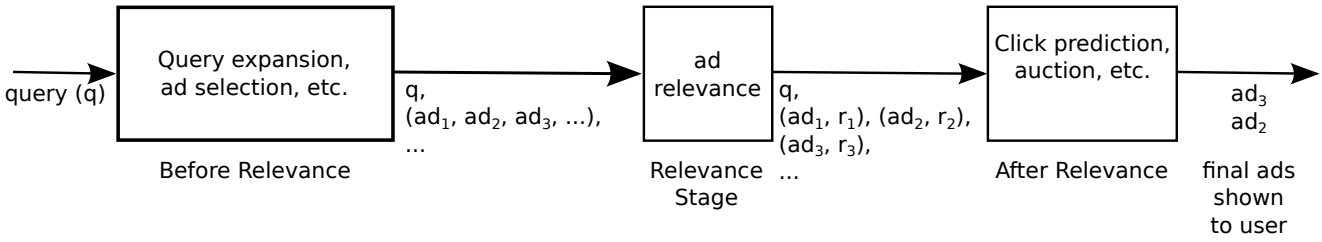


**Figure 1:** Abstract pipeline for selecting search ads. In this paper we focus on the relevance stage.

corporated large amounts of associated metadata such as the user's search pattern within a session [6] and click-through data [21], as well as employed various other query augmentation techniques [7, 25, 26, 28, 37] in attempts to accurately interpret user queries.

Similarly, various techniques have been used to understand the creative and landing page to improve ad selection [14]. Such approaches are effective because the creative is often a good reflection of what is being advertised and the landing page offers a rich set of features. However, the creative itself offers very little information (a typical creative is a few tens of characters, see Figure 2) and landing pages are known to be noisy [14]. These approaches are even more challenging to apply in the case of broad match since the query and advertisement may not be textually similar. For example, an ad bidding on "sneakers" might be quite relevant for the query "shoes" but there may be little textual similarity between them.

In this work, we complement prior approaches by interpreting the ad keyword as well. Unlike the creative and landing page, both of which are to be displayed to the end user, the ad keyword represents an unconstrained opportunity for the advertiser to be direct about their desires without concern of offending or dissuading the user. Hence, we argue that it represents a very strong signal that should be mined to the fullest extent. We build on the fact that—as discussed above—search engines are good at interpreting a query. In particular, we determine advertiser intent by submitting the ad keyword to the search engine and use organic results that the search engine returns to provide additional context with which to interpret the ad keyword. Specifically, given an advertisement to be scored for relevance against a particular query—we denote this as a (query, ad) pair in the remainder of the paper—-we send the ad keyword associated with the ad to the search engine and use the top organic results returned to get additional information about the ad keyword. We then use features extracted from both these results and the organic results for query itself to measure the similarity between the advertiser's and user's intents.

We consider introducing two complimentary sets of features: 54 features that can be generated using just the query issued by the user and another 21 that require information capturing user intent, which we get by using organic results generated for the user query. We evaluate the benefits of adding each of these feature sets by comparing the performance of the resulting ranker to the best previously published baseline [29] and the production system at a large search engine. We achieve a $43.2\%$ improvement in precision-recall area under the curve (AUC) over the baseline and $2.7\%$ improvement over the highly engineered production system.

## 2. BACKGROUND

In the sponsored search model advertisers provide an ad comprising of *i*) the creative, a short textual ad that is shown to the user, see Figure 2, *ii*) ad keyword, user queries the advertiser would like to target, *iii*) match type, how the search engine is allowed to match


**Figure 2:** Creative has a title, description and display URL.

the ad keyword to user queries, *iv*) destination URL, the web page to which the user should be redirected upon clicking the ad, and *v*) bid value, the cost that advertiser is willing to pay for a click.

### 2.1 Ads pipeline

When the user makes a query, the search engine has to decide which ads (if any) should be displayed along with the organic results. While the production pipeline has many stages, in this paper we abstract it into three stages as shown in Figure 1 with our primary contribution being to the relevance stage. Before the relevance stage, the query (q) is expanded to create expanded queries which are related to the original query depending on the match type indicated by the advertiser [10]. Each expanded query is then used to select candidate ads ($ad_1$, $ad_2$, ...) to be further evaluated. The goal at this stage is to identify all ads that are potentially related to the set of expanded queries.

In the relevance stage each ad picked thus far in the pipeline is evaluated for relevance to the original query. While the previous stages choose ads that are related to the expanded queries, this stage performs deeper inspection of the relevance of each ad ($ad_1$, $ad_2$, ...) to query. The output of this stage is a score for each ad ($ad_1$, $ad_2$, ...) on how relevant it is to query (i.e., generate $r_1$, $r_2$, ...). For example, an ad for "nike shirts" would be scored higher for the query "shirts" than would an ad for "jackets". To do so, the relevance stage trains a learning ranker [35] using a labeled training set of (query, ad) pairs and features computed for each (query, ad) pair. The trained ranker is then deployed to measure the relevance of an ad to the query. In this work we study features that can be used to improve the performance of the ranker.

After the relevance stage, the ads pipeline involves estimating the probability (click-through rate) that an ad would be clicked and conducting the second-price auction. The relevance of ad to query is a factor when estimating the click probability [1]. The probability estimate allows the search engine to calculate the expected revenue to be derived by showing a particular ad. These predictions are made using information about the ad from previous impressions of the ad [31]. For rare or new ads where such information is not available, information from semantically similar ads is used [16]. In both the cases, the relevance of ad to query can be used as one of the features to predict the click-through-rates. So, while click-through-rate estimates do filter out ads, accurate relevance scores complement these efforts. Subsequently, most search engines rank

ads by the product of click probability and advertiser bid value in an attempt to maximize expected revenue for the second-price auction [10, 20].

In this paper we evaluate the gains that can be achieved using two different sets of features. The first set of features use just the query that has been submitted by the user and advertiser intent captured using the ad keyword associated with the ad. The second set of features capture the similarity between user and advertiser intent by further considering the organic results the search engine generates in response to the user query, leading to additional improvements in relevance ranking.

## 2.2 Performance metrics

Search engines use metrics like share of queries with ads, ads per query, clicks per search impression, and clicks per ad [20] to evaluate the performance of the overall ads pipeline. A range of factors including user click probability and bid values play a role in the final decision to show an ad. Because we are focused exclusively on changes to the relevance stage—and, more to the point, do not have access to the various additional inputs used in latter stages of the pipeline—we instead measure the performance of the ranker and its feature set in terms of precision-recall values on a hold out validation set.

# 3. RELATED WORK

As discussed above, a lot of work has been done to interpret and expand the user query [10, 21]. User click behavior [15, 21, 32] and electronic dictionaries [34] have been used to enhance the query and expand it. The expanded queries are then used to retrieve relevant ads from the corpus of ads. These techniques rely on the fact that relevance of words to a particular query is correlated to the user click behavior or the semantic similarity between words.

Query expansion techniques have also used categorization of the query and topical information to achieve improvements [11, 24, 32] in ranking documents. These approaches use human-judged datasets to classify web pages into hierarchical categories. These categories are then used to find ads that may not be textually similar but belong to same category as the query. Bennett et al. [8] use documents classified under the Open Directory Project (ODP) [2] to train a classifier and show that using such techniques can improve ranking of relevant documents. Broder et al. [10] use search engine results to create an augmented query and then select ads using the augmented query.

Blind relevance feedback techniques have also been used to expand the query and retrieve relevant documents. These techniques [10, 18, 26, 36], like ours, assume that the top results returned by search engine are relevant to the query issued. Such cross-corpus learning techniques have been used to transfer knowledge from one task to another [17, 30].

The most relevant and related work is by Hillard et al. [20], addressing the problem of relevance using human-judged datasets. The authors use translation models to predict click-through rates from click logs. The predicted click-through rates are then used along with baseline features, to train a classifier to predict ad relevance. They also highlight challenges in predicting relevance of an ad to a query given the relatively short nature of queries and creatives.

Other work [20, 35] has explored the use of different supervised learning algorithms to improve the accuracy and performance of the relevance rankers. In this paper, we work with the ranker used by a large search engine and explore the advantages of using new features. As He et al. [19] point out, identifying the right features for the ranker is important. The focus of our work, then, is to illustrate the advantages of using features extracted from the ad keyword associated with an ad.

# 4. MOTIVATION

The goal of the relevance stage is to compute the relevance of the ad to query in each (query, ad) pair that has been selected by the previous stages. While the stages prior to the relevance stage focus on casting a wide net to rapidly identify as many related ads as possible, the relevance stage uses a broader range of features to measure the relevance of ad to the query.

## 4.1 Capturing advertiser intent

The learning ranker used to compute a score measuring the relevance of ad to query in a (query, ad) pair is trained using a set of features computed for each pair. For the task of feature computation, the key fields available in the ad are: *i*) creative (Figure 2), *ii*) ad keyword and *iii*) landing page. Features that are currently used by the production system include text similarity features between query and these fields along with other external sources of information — including the click-through rate of the ad from past impressions [5].

We expand the ad keyword and use the resulting features to measure relevance of ad to query. Our key insight here is that the ad keyword captures advertiser intent more accurately than the creative itself. The ad keyword is the only field in the entire ads pipeline through which an advertiser can explicitly express the type of traffic that they would like to attract. Other attributes of the ad, like the creative and landing page, are seen by the user which could prevent the advertiser from freely expressing their intent. Hillard et al. [20] observe that, for example, an ad for "limo rentals" would be quite relevant to a user query for "prom dresses". An advertiser might, thus, list an advertisement for "limo rentals" and bid for the keyword "prom dresses". In the absence of an understanding of the ad keyword, such an ad would be considered completely irrelevant to the user query "prom dresses". If one had a way to identify that that prom dresses and limos are frequently used together, however, better relevance scores could be computed improving the quality of ads delivered.

We choose to solve this problem at the relevance stage because it is expensive to perform a deep evaluation of all possible variations of the query computed by the query expansion algorithms employed in the stages prior to the relevance stage. However, ad keywords associated with the ads provide us with a defined set of keywords on which deeper analysis can be performed.

## 4.2 Broad match opportunity

The importance of understanding the ad keyword is highlighted in the case of broad match when the query is expanded before being matched against the ad keyword, increasing the likelihood that the user's original intent may not align with the advertiser's. To illustrate this, we compare the performance of a previously published baseline relevance ranker on (query, ad) pairs matched using exact match to those matched using broad match. The performance of the relevance ranker is evaluated using precision/recall values over a hold out validation set. This approach allows for evaluation of the ranker independent of other factors like click probability and bid values which play a significant role in the final decision to show an ad to the user.

Figure 3 plots the precision/recall curve obtained by scoring (query, ad) pairs in the validation set using a relevance ranker trained on baseline features [29] (detailed in Section 5.3). High precision values indicate that a large fraction of the ads being selected are relevant — enhancing the user experience. Whereas, high recall

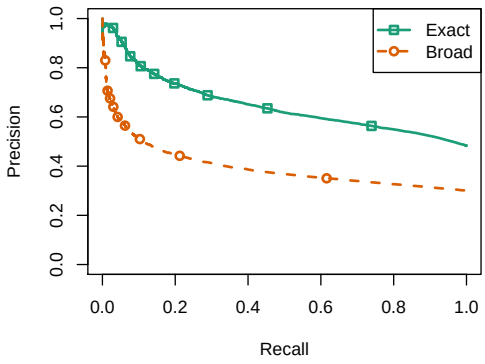


**Figure 3:** Ads chosen through exact match can be scored more accurately with previously published baseline features than those chosen through broad match.

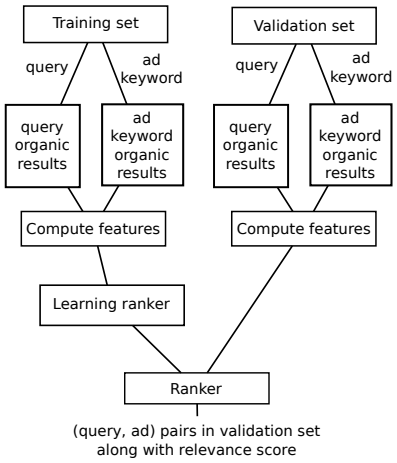


**Figure 4:** Feature construction and system evaluation overview.

values indicate that a greater number of relevant ads are being selected — improving revenue opportunity for the search engine. Unsurprisingly, the baseline features that measure similarity between only the query and the ad to determine relevance do not work as well in the case of broad match.

# 5. METHODOLOGY

In this section we present our methodology for using the capabilities of the search engine to create features that represent the advertiser intent. Using the search engine to interpret a query, which is usually short, allows us to leverage years of research. Using the organic web search results corresponding to the ad keyword provides us with detailed information about advertiser intent. We use these features to improve the performance of a learning ranker used to measure the relevance of an ad to a query. For comparison purposes we build upon a baseline ranker using the 19 features described by Hillard et al. [29].

This section starts with a description of the datasets, learning ranker, and baseline features that we use. We then introduce our additional features, starting with features that can be extracted using just the user query along with organic results for the ad keyword. We then explore features of the organic results for the query which act as a proxy for user intent information allowing us to capture the overlap between advertiser and user intent.

## 5.1 Data overview

The key datasets in our system are the training and validation sets comprising of (query, ad) pairs. Each ad described in Section 2 has an associated ad keyword indicating the type of queries from which the advertiser would like to attract traffic. As shown in Figure 4, in the first phase, we obtain the top 40 organic results associated with the query and the ad keyword in an approach similar to the one taken by Hillard et al. [10].

Each of the 40 organic results returned for a query submitted to the search engine has the following fields associated with it: *i*) title of the web page as shown in the web results, *ii*) snippet, a small piece of text displayed on the results page providing a view into the result page itself, *iii*) description, a small piece of text from the web page which most accurately describes the web page, *iv*) ODP category, the ODP [2] category to which the result belongs, and *v*) URL of the result.

To clean up these fields, we remove stop words [3] and stem the title, snippet, and description of each result using the Porter stemmer [27]. We then concatenate all the titles, snippets, descriptions of results associated with each query to create a bag-of-words representation.

## 5.2 Ranker

For each (query, ad) pair in the training and validation sets, we compute features as described in the following sections. We train the LambdaMART learning ranker [12] on the features obtained over the training dataset. LambdaMART has been shown to be very effective in solving real-world ranking problems [8, 13]. LambdaMART is known to be robust to features that take a range of values and produces a tree-based model. In our evaluation, the algorithm is trained at a learning rate of 0.12, with 120 leaves and 2,000 trees. The model produced by the ranker can be used to determine a ranked list of the features on which the ranker was trained. We use this attribute of LambdaMART to identify the importance of features we discuss in this paper.

## 5.3 Baseline features

We use the features described by Hillard et al. [29] as a baseline against which to compare the gains offered by the additional features. Hillard proposes 19 features: query length and $6 \times 3$ features obtained by computing the following: *i*) word unigram overlap, *ii*) word bigram overlap, *iii*) character unigram overlap, *iv*) character bigram overlap, *v*) ordered word bigram overlap, and *vi*) cosine similarity between the query and each of the title, description, and display URL of the creative.

Each of the overlap features is the overlap coefficient of the corresponding sets, computed as:

$$\text{overlap}(X, Y) = \frac{|X \cap Y|}{\min(|X|, |Y|)}.$$

For example, the word unigram overlap coefficient between "black shoes" and "shoes at contoso inc" would be 0.5.

## 5.4 Query features

The first set of features that we introduce rely on using only the information that can be computed based upon the ad and the query in each (query, ad) pair. In the next section we discuss more advanced features that can be computed using organic results for the query which allows us to better measure the similarity between user and advertiser intent.

| Feature Type | Details | Count |
|---|---|---|
| Creative | ad title ∩ bk. titles | 6 |
| | ad text ∩ bk. desc. | 6 |
| | ad text ∩ bk. snip. | 6 |
| Landing Page | title ∩ bk. titles | 6 |
| | snippet ∩ bk. desc. | 6 |
| | snippet ∩ bk. snip. | 6 |
| Query | query ∩ bk. titles | 6 |
| | query ∩ bk. desc. | 6 |
| | query ∩ bk. snip. | 6 |
| Total | | 54 |

**Table 1:** Query features computed using organic results for the ad keyword.

For each ad, we compute features to determine whether the ad creative and landing page are consistent with the ad keyword the advertiser supplies. Specifically, we compute the same six similarity features as Hillard et al. [29], but between the organic results returned for the ad keyword and aspects of the ad creative and landing page. For the creative, we compare the ad title to the search result titles, and the ad text to both the search result description and snippets. For the landing page, we compare its title to the search result titles, and the snippet to both result descriptions and snippets.

We further compute features that measure the similarity of query to the results of searching for the ad keyword. For this, we compute the same six similarity features, but this time between the query and the titles, snippets and descriptions associated with the ad keyword search results, respectively. These features are easy to implement because organic results for ad keyword can be precomputed, and when the query is received in the online system, feature construction is a matter of computing overlap features. While in each case we use the same overlap and cosine similarity features as in Section 5.3, there is no limitation against using other similarity measures like Jacquard index or edit distance. In total, we add the $54$ features shown Table 1. We call these query features.

## 5.5 Query search features

We also consider features that can be computed if the ads pipeline can interpret user intent in the same way the search engine does to generate organic results for query. We use the organic results generated for query as a proxy to capture user intent in much the same way as we use organic results for the ad keyword to capture advertiser intent. Once we have the organic results for the query, we compute six overlap features for each pair of titles, snippets and descriptions obtained from results of query and ad keyword.

### *5.5.1 Category feature*

Each web page in the search index is classified using ODP data [2] into categories by an internal classification engine at indexing time as described by Bennett et al. [8]. Each organic result (which is chosen from the index) is thus classified into one of the 219 categories at the top two levels of the hierarchy. For the query and ad keyword, we obtain the categories to which the corresponding organic results belong. We then compute the cosine similarity between the categories of the two organic results:

$$\text{score} = \frac{\sum_c n_{qc} \times n_{kc}}{\sqrt{\sum_c n_{qc}^2} \times \sqrt{\sum_c n_{kc}^2}},$$

| Feature Type | Details | Count |
|---|---|---|
| Query Search | query titles ∩ bk. titles | 6 |
| | query desc. ∩ bk. desc. | 6 |
| | query snip. ∩ bk. snip. | 6 |
| Category overlap | query ∩ bk. categories | 1 |
| Domain count | domain in query results | 1 |
| | domain in bk. results | 1 |
| Total | | 21 |

**Table 2:** Query search features constructed using organic results for query and ad keyword.

where, $n_{qc}$ and $n_{kc}$ are the number of times a result belonging to category $c$ is present in organic results for the query and ad keyword respectively.

As has been argued by Broder et al. [10], this feature allows us to identify scenarios when the query and ad keyword might not have a strong overlap but are relevant to each other because they belong to the same category. For example, for the query "shoes", the ad keyword "sneakers" does not result in a text overlap, but is very relevant.

### *5.5.2 Domain features*

The last set of features that we introduce captures the presence of the ad domain (e.g., `contoso.com`) in the organic results for query and ad keyword. The ad domain of an ad is determined to be relevant to a query if the ad domain is present in the organic results for the query. Similarly, the the presence of the ad domain in organic results for the ad keyword associated with the ad indicates that the ad domain is relevant to the type of traffic the advertiser wants to attract by bidding on the particular ad keyword. We introduce two features to capture the relevance of ad domain to query and the ad keyword. Specifically, the features are computed as number of times the ad domain is present in organic results for both the query and ad keyword.

In sum, we call these additional $18 + 1 + 2 = 21$ features query search features and summarize them in Table 2. Together with the 54 features computed in Section 5.4 they form a total of 75 features that we consider.

# 6. EVALUATION

In this section we quantify the improvements in relevance ranking obtained by incorporating advertiser intent. We present the gains in precision and recall over both a published baseline [29] and the production system for a large search engine.

The baseline system [29] is rudimentary and captures only the similarity between query and the ad creative for each (query, ad) pair. However, as far as we are aware this is the best-performing published system in this space. The features that we introduce use much richer information from the organic results to capture advertiser and user intent. As a result, we achieve extraordinary gains over the baseline. Our benefits over the production system—which uses hundreds of features—are less dramatic, but still significant in practice.

## 6.1 Datasets

The learning ranker that we use is trained using a sample of 1.28 million hand-scored (query, ad) pairs drawn from the ground-truth the production system uses. The scores ranges between one and five with five representing high relevance between the query and the ad. For offline testing of the model, the holdout validation dataset has

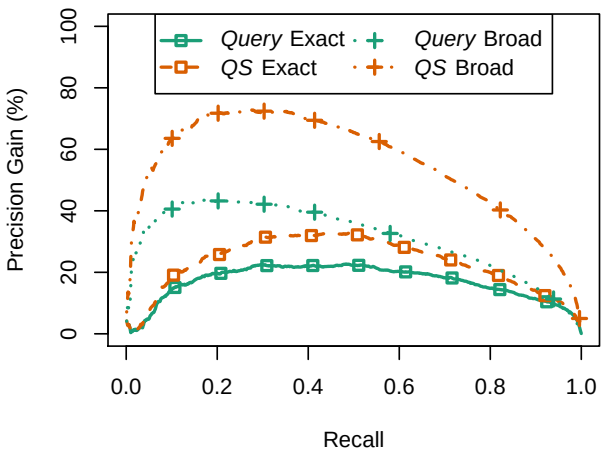


**Figure 5:** Improvement in precision for different recall values using the new ranker.

| | | Ground-Truth Scores | | | |
|---|---|---|---|---|---|
| | | All | 3+ | 4+ | 5 |
| Query Features | All | 27.5 | 70.3 | 60.1 | 21.6 |
| | Exact | 16.6 | 37.6 | 35.7 | 11.5 |
| | Broad | 32.2 | 90.3 | 82.6 | 58.8 |
| Query Search Features | All | 43.2 | 103.5 | 122.7 | 79.6 |
| | Exact | 22.7 | 50.0 | 67.9 | 52.2 |
| | Broad | 55.7 | 165.3 | 228.6 | 145.8 |

**Table 3:** Relative (%) improvement in precision-recall AUC over baseline for different types of ads.

320,000 similarly sampled (query, ad) pair scores. The training and validation datasets are retrieved from the ad corpus using information retrieval methods used by the selection stage [10]. They contain queries from all search frequency deciles.

We obtain the organic results corresponding to the query and the ad keyword by submitting each of them to the search engine. We also use the ODP [2] categorization of the organic results returned for query and ad keyword.

## 6.2 Baseline comparison

We start by evaluating the gains that the new features provide over the baseline [29]. For now, we consider an ad to be not relevant to the query if the (query, ad) pair is judged to be a one. We return to consider more stringent cutoffs in Section 6.2.3.

### *6.2.1 Query features*

As discussed in Section 5, for each (query, ad) pair in the training and the validation sets, we add 54 new features. Among these, 18 features compute the similarity of query to the title, snippets and descriptions associated with organic results for the ad keyword. Similarity between creative, landing page and ad keyword organic results is captured in another 36 of these features. We compare the performance of a ranker trained with these features to a ranker trained using only the 19 baseline features.

The "Query Exact" and "Query Broad" lines in Figure 5 show the relative improvement in precision at different recall values over the baseline obtained by using the ranker trained on query features over (query, ad) pairs matched through exact and broad match, respectively. The relative change in the area under curve for the precision-recall curves is presented in the All column of Table 3.

The results show that adding information from organic results for the ad keyword provides a large improvement in precision over the baseline. Also, note that the improvement is higher for (query, ad) pairs matched through broad match. Intuitively, the improvement is because information from organic results for the ad keyword increases the possibility of a match between the query and ad when the ad is relevant to the query. In the case of exact match, the baseline features already capture the overlap between the ad and the query because the creative likely contains the ad keyword and hence contains the query.

### *6.2.2 Query search features*

In this section, we measure the benefit of adding new features which capture the similarity between user and advertiser intent by using organic results for the query and ad keyword. We introduce a total of 21 features which capture the similarity between user intent and the advertiser intent along with 54 features introduced in the previous section.

As described in Section 5 we use query organic results as a proxy for interpreting the user intent. Note that in an online system generating both organic and sponsored results, the results for query themselves may not be needed, instead techniques used to process the query and identify organic results would be enough to interpret the query.

The organic results of a query give us titles, snippets and the descriptions associated with the query. For each of these fields we compute six features which measure similarity to the corresponding field from organic results for ad keyword. This gives us a total of $6 \times 3$ new features. In addition to this, we also add one feature which captures the similarity between the categories of results for the query and ad keyword. Two additional features use organic results to capture how relevant the ad domain itself is to the query and ad keyword.

The "QS Exact" and "QS Broad" lines in Figure 5 show the improvement in precision at different recall values using the model obtained by training the ranker with query search features for exact and broad match types, respectively. These results show that using search results for the ad keyword to interpret advertiser intent provides a large improvement in the accuracy of the relevance ranker over using just the baseline features. Again, the relative change in precision-recall AUC is presented in the All column of Table 3.

### *6.2.3 Identifying good ads*

While achieving high overall precision-recall numbers is important to distinguish between relevant and irrelevant ads, a good relevance ranker should be especially adept at identifying high-quality ads accurately. So, techniques which lead to gains in overall precision should not negatively impact the ability of the search engine to identify ads scored three or higher. The ability to identify good ads accurately is a desirable feature for the model used by search engines because it allows search engine to show good ads to the users and not just suppress bad ads. While suppressing bad ads is good for user experience, a ranking model which does not identify good ads would lead to lower revenue.

We measure the ability of the new features to distinguish between ads scored three or higher and ads scored lower than three. We consider ad scored one or two as irrelevant to the corresponding query in the (query, ad) pairs. The relative improvement in precision-recall AUC over the baseline model is shown by the 3+ column in Table 3. We perform a similar analysis for ads scored

| | | All | 3+ | 4+ | 5 |
|---|---|---|---|---|---|
| Query Features | All | 0.5 | 1.0 | 0.8 | 0.6 |
| | Exact | 0.4 | 0.6 | 0.1 | 0.1 |
| | Broad | 0.6 | 1.8 | 2.7 | 1.9 |
| Query Search Features | All | 2.7 | 3.7 | 5.7 | 8.9 |
| | Exact | 0.7 | 1.6 | 2.6 | 5.4 |
| | Broad | 3.9 | 7.3 | 12.3 | 13.2 |

**Table 4:** Relative (%) improvement in precision-recall AUC over the production ranker for different types of ads.

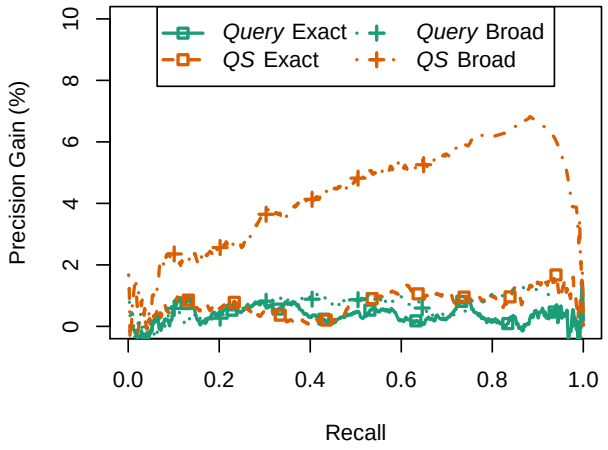


**Figure 6:** Gains in precision with the introduction of features capturing advertiser intent.

four or higher, and five by considering the remaining ads to be irrelevant respectively.

We see that the understanding of advertiser intent works especially well for identifying high-quality ads. Ads scored four or better and those scored three or better see greater improvement than the overall pool of ads. This behavior is expected because the ad keyword organic results of good ads are more likely to have pattern overlap with the query organic results.

## 6.3 Production comparison

The results in the previous sections demonstrate that the new features provide significant gains over a published baseline proposed by Hillard et al. [29]. It stands to reason, however, that highly engineered production systems employed by major search engines exhibit better performance and may, in fact, already consider many of the features we suggest. In this section, we consider the gains our new features bring to the production pipeline of a large search engine. We add the above mentioned features to the existing production features and train the production ranker on the combined feature set. The production features capture a variety of attributes about the query and ad, including quality of the ad domain, historical click-through-rates of the ad and landing page attributes [5].

Precision gain over the production system at different recall values with the introduction of query features for exact and broad match are shown using "Query Exact" and "Query Broad" lines in Figure 6. The results show that there is a small improvement in precision-recall curves due to the introduction of query features. The relative change in the precision-recall AUC is presented in the All column of Table 4. The gains are smaller than those over baseline but are nevertheless significant in a production system. As before, the new features perform best for highly-relevant ads in broad searches.

| Ranker | Precision | Recall | Max F-Score |
|---|---|---|---|
| Query | -1% | +2.3% | +0.3% |
| Query Search | +1.7% | +2.9% | +2.2% |

**Table 5:** Gains achieved in precision, recall and max F-Score compared to the production system of a large search engine.

Moreover, we find that the production ranker improves significantly with the addition of query search features. The improvement in precision at different recall values is shown in Figure 6 by the lines labeled "QS Exact" and "QS Broad". The relative improvement in AUC is presented in the All column of Table 4. As before, the new features work even better for high-quality ads.

To capture improvement in the accuracy of the ranker, we measure precision-recall values at max F-score. Table 5 shows the precision and recall values for the production ranker and the ranker trained using query search features at max F-score.

## 6.4 Feature importance

The value of the features we introduce can also be seen in the importance given to them by the LambdaMART ranker. The tree based model that is produced after training the ranker allows us to determine the ranked list of features. The importance of query search features is very clear in the ranking of the features. We find that the tree created by training the ranker on a combination of existing features and query search features ranks the following as the top three features: *i*) ad domain count in query organic results, *ii*) ordered bigram overlap between snippets of organic results for query and ad keyword and *iii*) ad domain count in ad keyword organic results. These features rank higher than many other highly engineered features.

The benefits of adding query features are smaller than those when compared to the baseline as we in Section 6. But, the value of interpreting the ad keyword is reflected in the ranking of the new features. Two of the 54 features are among the top 30 features in the final tree produced by the ranker. These are: *i*) word unigram overlap between query and snippets in organic results for ad keyword and *ii*) order word bigrams between query and titles of the organic results for ad keyword.

# 7. DISCUSSION

The gains we see over the production system are naturally lower than the gains over the baseline. However, they are significant in production [4]. Moreover, the features that we propose can be obtained using the datasets and learning experience already at the disposal of the search engine.

Quantifying the impact of these improvements on the revenue of the search engine and user experience is complicated. We see significant improvements in ranking accuracy. However, it may be possible that the ads which have been more accurately scored will not be shown to the user for a host of other reasons including low click probability, low bid values by the advertiser, and so on. In this case, improvements to the relevance ranker would not enhance the user experience. Increase in recall values, however, would lead to the identification of more ads that are truly relevant to the user query, leading to greater competition in the auction—and higher revenue for the search engine.

An immediate extension of our work is to study the benefits of incorporating these features over a slice of live traffic handled by the production system. Such a study would allow us to measure the

impact of new features on metrics (e.g., share of queries with ads, ads per query, clicks per search impression and CTR) that have a direct bearing on user experience and search engine revenue.

# 8. CONCLUSION

The ad keyword is the only field in a sponsored search ad that allows an advertiser to express the type of traffic that they would like to attract. At the same time the ad keyword, like the user query, is very short which makes the task of interpreting advertiser intent hard. In this work, we leverage the ability of modern search engines to interpret the intent behind a user's query to similarly understand the advertiser's intent as conveyed in the ad keyword.

We make three main contributions in this paper. First, we show that using organic engine results to expand the ad keyword provides us with a rich source of information from which we can interpret advertiser intent. Second, we identify the features to be extracted from these organic results which can be used to improve the relevance ranker. Among these, 54 features can be implemented with few changes to the existing system and another 21 features would require user intent information as well. Finally, we evaluate the benefits of these features using training and validation datasets of 1.28M and 320,000 samples sampled from a corpus of ground-truth (query, ad) pair scores respectively. We show that using features which capture user and advertiser intent leads to 43.2% improvement in precision-recall AUC over the baseline and a 2.7% improvement over the production ranker for a large search engine.

## Acknowledgments

We thank Greg Buehrer for initial discussions on the idea of using web relevance feedback for improving ad relevance. We also thank the anonymous reviewers for their feedback. This work was funded in part by the National Science Foundation through grant CNS-1237264.

# 9. REFERENCES

[1] Click Modeling in Search Advertising: Challenges & Solutions. http://advertise.bingads.microsoft.com/en-us/editorial-relevance-quality-guidelines.

[2] DMOZ - the Open Directory Project. http://www.dmoz.org.

[3] English Stop Words (CSV). http://www.textfixer.com/resources/common-english-words.txt.

[4] Personal communications with Bing Ads Relevance Team.

[5] Relevance and quality guidelines - Bing Ads. http://advertise.bingads.microsoft.com/en-us/editorial-relevance-quality-guidelines.

[6] E. Agichtein, E. Brill, and S. Dumais. Improving Web Search Ranking by Incorporating User Behavior Information. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '06, pages 19–26, 2006.

[7] L. Ballesteros and W. B. Croft. Phrasal Translation and Query Expansion Techniques for Cross-language Information Retrieval. In *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '97, pages 84–91, 1997.

[8] P. N. Bennett, K. Svore, and S. T. Dumais. Classification-Enhanced Ranking. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 111–120, 2010.

[9] A. Broder, M. Ciaramita, M. Fontoura, E. Gabrilovich, V. Josifovski, D. Metzler, V. Murdock, and V. Plachouras. To Swing or Not to Swing: Learning when (Not) to Advertise. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, CIKM '08, pages 1003–1012, 2008.

[10] A. Z. Broder, P. Ciccolo, M. Fontoura, E. Gabrilovich, V. Josifovski, and L. Riedel. Search Advertising Using Web Relevance Feedback. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, CIKM '08, pages 1013–1022, 2008.

[11] A. Z. Broder, M. Fontoura, E. Gabrilovich, A. Joshi, V. Josifovski, and T. Zhang. Robust Classification of Rare Queries Using Web Knowledge. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '07, pages 231–238, 2007.

[12] C. Burges, R. Ragno, and Q. Le. Learning to Rank with Non-Smooth Cost Functions. In *Advances in Neural Information Processing Systems 19*. MIT Press, Cambridge, MA, January 2007.

[13] C. J. Burges. From RankNet to LambdaRank to LambdaMART: An Overview. Technical Report MSR-TR-2010-82, June 2010.

[14] Y. Choi, M. Fontoura, E. Gabrilovich, V. Josifovski, M. Mediano, and B. Pang. Using Landing Pages for Sponsored Search Ad Selection. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 251–260, 2010.

[15] M. Ciaramita, V. Murdock, and V. Plachouras. Online Learning from Click Data for Sponsored Search. In *Proceedings of the 17th International Conference on World Wide Web*, WWW '08, pages 227–236, 2008.

[16] K. S. Dave and V. Varma. Learning the Click-through Rate for Rare/New Ads from Similar Ads. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '10, pages 897–898, 2010.

[17] C. B. Do and A. Y. Ng. Transfer learning for text classification. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *Advances in Neural Information Processing Systems 18*, pages 299–306. MIT Press, 2006.

[18] E. N. Efthimiadis and P. V. Biron. UCLA-Okapi at TREC-2: Query Expansion Experiments. In *In Proceedings of the Second Text Retrieval Conference*, pages 500–215, 1994.

[19] X. He, J. Pan, O. Jin, T. Xu, B. Liu, T. Xu, Y. Shi, A. Atallah, R. Herbrich, S. Bowers, and J. Q. n. Candela. Practical Lessons from Predicting Clicks on Ads at Facebook. In *Proceedings of the Eighth International Workshop on Data Mining for Online Advertising*, ADKDD'14, pages 5:1–5:9, 2014.

[20] 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*, WSDM '10, pages 361–370, 2010.

[21] K. Hui, B. Gao, B. He, and T.-J. Luo. Sponsored Search Ad Selection by Keyword Structure Analysis. In *Proceedings of the European Conference on Information Retrieval*, ECIR '13, pages 230–241, 2013.

[22] B. J. Jansen and T. Mullen. Sponsored search: an overview of the concept, history, and technology. *International Journal of Electronic Business*, 6(2):114–131, 2008.

[23] B. J. Jansen and M. Resnick. An examination of searcher's perceptions of nonsponsored and sponsored links during ecommerce Web searching. *Journal of the American Society for Information Science and Technology*, 57(14):1949–1961, 2006.

[24] P. Kowalczyk, I. Zukerman, and M. Niemann. Analyzing the Effect of Query Class on Document Retrieval Performance. In G. Webb and X. Yu, editors, *AI 2004: Advances in Artificial Intelligence*, volume 3339 of *Lecture Notes in Computer Science*, pages 550–561. Springer Berlin Heidelberg, 2005.

[25] D. Metzler, S. Dumais, and C. Meek. Similarity Measures for Short Segments of Text. In G. Amati, C. Carpineto, and G. Romano, editors, *Advances in Information Retrieval*, volume 4425 of *Lecture Notes in Computer Science*, pages 16–27. Springer Berlin Heidelberg, 2007.

[26] M. Mitra, A. Singhal, and C. Buckley. Improving Automatic Query Expansion. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, pages 206–214, 1998.

[27] M. F. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.

[28] F. Radlinski, A. Broder, P. Ciccolo, E. Gabrilovich, V. Josifovski, and L. Riedel. Optimizing Relevance and Revenue in Ad Search: A Query Substitution Approach. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '08, pages 403–410, 2008.

[29] 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*, SIGIR '09, pages 762–763, 2009.

[30] R. Raina, A. Y. Ng, and D. Koller. Constructing Informative Priors Using Transfer Learning. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pages 713–720, 2006.

[31] M. Richardson, E. Dominowska, and R. Ragno. Predicting Clicks: Estimating the Click-through Rate for New Ads. In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, pages 521–530, 2007.

[32] D. Sculley, R. G. Malkin, S. Basu, and R. J. Bayardo. Predicting Bounce Rates in Sponsored Search Advertisements. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, pages 1325–1334, 2009.

[33] Y. Song, H. Ma, H. Wang, and K. Wang. Exploring and Exploiting User Search Behavior on Mobile and Tablet Devices to Improve Search Relevance. In *Proceedings of the 22nd International Conference on World Wide Web*, WWW '13, May 2013.

[34] E. M. Voorhees. Query Expansion Using Lexical-semantic Relations. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '94, pages 61–69, 1994.

[35] Q. Wu, C. Burges, K. Svore, and J. Gao. Adapting boosting for information retrieval measures. *Information Retrieval*, 13(3):254–270, June 2010.

[36] J. Xu and W. B. Croft. Improving the Effectiveness of Information Retrieval with Local Context Analysis. *ACM Trans. Inf. Syst.*, 18(1):79–112, Jan. 2000.

[37] W. V. Zhang, X. He, B. Rey, and R. Jones. Query Rewriting Using Active Learning for Sponsored Search. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '07, pages 853–854, 2007.