Evaluating Vector-Space and Probabilistic Models for Query to Ad Matching

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

Rukmini Iyer
Yahoo! Inc
Great America Parkay
Santa Clara, CA, 95054
riyer@yahoo-inc.com

ABSTRACT

In this work, we evaluate variants of several information retrieval models from the classic BM25 model to Language Modeling approaches for retrieving relevant textual advertisements for Sponsored Search. Within the language modeling framework, we explore implicit query expansion via translation tables derived from multiple sources and propose a novel method for directly estimating the probability that an advertisement is clicked for a given query. We also investigate explicit query expansion using regular web search results for sponsored search using the vector space framework. We find that web-based expansions result in significant improvement in Mean Average Precision.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

General Terms

Information Retrieval, Vector Space Models, Language Models, Translation Models

1. INTRODUCTION

The primary source of revenue for major search engines is through advertising. The online ad spend of advertisers has been growing significantly over the past few years [1]. In the popular auction model used by search engines, an advertisers bids on a keyword such as *used cars*. When a user types in the query *used cars*, this particular advertiser's ads will be among the candidate set of ads that can be displayed alongside the search results. If an advertiser opts in for "advance match", his ad may also be shown for queries such as *cheap cars* or *old cars*. Once the ads are part of the candidate set, they are then ranked by a product of relevance of the ad to the query and the bid. If a user clicks on the ad, the advertiser pays the search engine for the click. The cost paid is determined by the bid and relevance of the ad shown immediately below the given ad, following the framework of a generalized second price auction [11]. Since the advertiser pays only when a user clicks on the ad, this monetization model is often called "pay-per-click" marketing. If the relevance score is assumed to be a measure of the probability that an ad is clicked for a given query, or the estimated click-through-rate of the ad for that query, ranking by bid $\times$ relevance is optimal for maximizing revenue.

Although sponsored search is a relatively new area of research, matching an ad to a query poses problems quite similar to those addressed by the information retrieval and web search community for many years. However, there are some key differences between web search and sponsored search. One of the primary differences is that the collection of web documents is significantly larger than the advertiser database, and retrieving candidate ads for tail queries using advanced match is a very important area of research for sponsored search. Another big difference from web search is the fact that the user model is different. Many queries do not have commercial intent; displaying ads on a query like "formula for mutual information" may hurt user experience and occupy real-estate on the search results page in a spot where a more relevant web-search result might exist. Therefore, in sponsored search, we prefer not to show any ads when the estimate of click-through-rate and/or relevance of the ad is low. Using the same user experience argument, for a navigational query [4] like "bestbuy.com", we would rather show only the most relevant exact ad if that ad existed in the advertiser database. We refer the reader to the study of Jansen and Resnick [15] for further details on user perceptions of sponsored search. In web-search, determining how many candidates to retrieve and display is not as much of an issue as the generally accepted user model is one where users read the page in sequence and exit the search session when their information need is satisfied. While all of the above problems are interesting areas of research, we restrict ourselves to the following scope.

Scope of this work:

In this paper, we are concerned with only those ads that have opted in for advanced match. We intend to compare various well known information retrieval methods for (advance) matching queries to ads. While in practice several features associated with the observed historical click-through-rate of an ad might be used in addition to word-overlap features (see eg. the work of Richardson et al [30]), in this paper we consider mainly the problem of using traditional information retrieval features to determine relevance of an ad

to a query. We seek to evaluate whether techniques that work well in classic information retrieval will work for the problem of advance matching queries to ads. In addition, we propose a novel variant of translation models in information retrieval, where the relevance score is an estimate of the click-through rate of an ad to a query. We discuss how we would incorporate our methods into more feature based methods in Section 7.

In the following section, we describe the structure of an advertisement and the related terminology. In section 3, we describe the various models that we apply to the problem of advertisement retrieval. Section 4 outlines our experimental setup and evaluation procedure. In section 5, we report initial results and analyze the differences across the various approaches. Next, in section 6, we place our work in the context of other work in sponsored search and information retrieval. Finally, we derive conclusions from our current set of experiments and explore future directions in section 7.

The main contributions of this work are: (1) A study of information retrieval techniques for sponsored search with insights into when and why certain types of techniques will and will not work (2) A proposal for a new click-based translation model for sponsored search.

2. AN ADVERTISEMENT

In this section, we describe the template of an advertisement using the specific example of Yahoo!'s Panama platform. The other search engines are similar. We use an illustrative but fictitious example of an advertiser who sells all kinds of shoes on the internet. This advertiser typically will have an **account** with the search engine, and would run many **campaigns**. A campaign can consist of many ad groups each of which in turn consist of a set of related keywords for a campaign. [1]

Bidded terms or keywords: For each adgroup, there is a set of keywords that the advertiser bids on, e.g., *sports shoes*, *stilletoes*, *canvas shoes* etc.

Creative: A creative is associated with an adgroup and is composed of a title, a description and a display url. Advertisers may chose to use a template for an ad. The template may have a title like *Buy* {*keyword:cheap shoes*}, an abstract - *Find* {*keyword:shoes of all styles*} *at low prices* and a display url *cheapshooz.com*. The portion between curly braces can be substituted by alternate text (henceforth called alt text) corresponding to a bidded term. So for a bidded tem *sports shoes*, if the advertiser has specified the alt text as *sneakers*, the title will be converted to *Buy Sneakers*. Similar is the case for the abstract. The default text in the template is used in case the ad exceeds a certain length after the template is filled out.

Matchtype: An advertiser can choose to use "standard" or "advanced" match for the keywords or adgroup. For example, if the advertiser choses to use only standard match for the keyword "sports shoes", his ad may be shown for that exact query. Whereas, if he enables the keyword to be advance matched, the search engine can show the same ad for the queries "running shoes" or "track shoes".

Bid: Associated with each keyword is a bid. The final ranking displayed on the search engine is a product of the *bid* and the *relevance* of the ad to the query. Relevance can be assumed to be a surrogate for the expected click through rate (CTR) of the ad. Hence ranking by the product of relevance and bid is an attempt to maximize revenue for the search engine.

Landing Page: Clicking on an ad will lead the user to the landing page of the advertiser which can also be very informative of the relevance of the ad to the query. Note that the landing page is typically the "document" used in web search - the title and creative for displaying a web result are typically auto-generated by a model at runtime.

[1] `http://help.yahoo.com/help/l/us/yahoo/ysm/sps/manage/mngca/import_spreadsheet.html`

3. MODELS

A document $\mathcal{D}$ in our index is a unique creative composed of 5 zones: the unfilled templates of the title and description, the display url, the bidded terms and the alt text. Let $z = 1...5$ represent an index into each of the above mentioned five zones respectively. In this work, a query $\mathcal{Q}$ is represented as a bag of words $q_1, q_2, ..., q_n$. For all our models, the similarity of a query to a document ($S(\mathcal{D}, \mathcal{Q})$) is a linear combination of the similarity of the query to the individual zone. In other words,

$$S(\mathcal{D}, \mathcal{Q}) = \sum_z w_z S(\mathcal{D}_z, \mathcal{Q})$$

where $\mathcal{D}_z$ represents a zone of a document $\mathcal{D}$ and w_z is the weight attributed to the zone. Henceforth, we use $\mathcal{D}$ to mean $\mathcal{D}_z$ for better readability.

3.1 BM25

We evaluate the classic BM25 model [31] which is an approximation of the 2-poisson model and is often considered the state-of-the-art for many information retrieval tasks [34]. The model, as we used it, is outlined below:

$$S(\mathcal{D}, \mathcal{Q}) = \sum_i IDF(q_i) \frac{tf_{q_i,\mathcal{D}}(k_1+1)}{tf_{q_i,\mathcal{D}} + k_1(1 - b + b\frac{L_\mathcal{D}}{avg\_dl})} \quad (1)$$

$$IDF = log\frac{N - n_i + 0.5}{n_i + 0.5} \quad (2)$$

$$n_i = \text{number of documents containing } q_i$$

$$N = \text{collection size}$$

$$tf_{q_i,\mathcal{D}} = \text{term frequency of } q_i \text{ in } \mathcal{D}$$

$$L_\mathcal{D} = \text{length of document } \mathcal{D}$$

The parameters k and b are tuned empirically.

3.2 Term Presence Absence Model (PA)

We also explored a simple model that ignored TF and considered only the IDF part of the above equation. We call this simple model the term presence absence model; this model is likely to place greater emphasis on a document that has all the terms in a query than the above BM25 model.

3.3 Combination Vector Space Model (PA+-BM25)

Experiments on our training data with this model showed increased precision and decreased recall due to the PA model. This lead us to try out a linear weighted combination of the BM25 and this model as well (PA+BM25). We discuss all these models and their advantages and disadvantages in greater detail in section 5.

3.4 Query Expansion using the Web as an external resource

External resources for query expansion have proven useful for several information retrieval tasks [9, 18, 7] as well as sponsored search [28]. Since web queries are very short and often only 2-3 words in length, expansion helps add context, as well as add synonyms to the original query.

We tried a simple query expansion approach using the web. We queried our native search engine for the top 10 results. We concatenated the bag-of-words of the query-biased summaries of the top 10 results, and retained the top 5 most frequent terms as query expansion terms. We expect that the addition of terms such as "bank of america" and "bankofamerica" to the query "boa". We use the expanded query to retrieve documents in the vector space framework outlined in sections 3.1 and 3.2 above.

3.5 Language Models

The language modeling framework makes the assumption that a user has an idea of what the "perfect" document for his or her information need will look like. The user samples from this perfect document to generate the query $\mathcal{Q}$. The task of the system then is to estimate the document closest to the ideal document for the query $\mathcal{Q}$.

$$\underset{\mathcal{D}}{\operatorname{argmax}}\, P(\mathcal{D}|\mathcal{Q}) = \underset{\mathcal{D}}{\operatorname{argmax}}\, \frac{P(\mathcal{Q}|\mathcal{D})P(\mathcal{D})}{P(\mathcal{Q})}$$

In the query likelihood model we rank documents by $P(\mathcal{D}|\mathcal{Q})$. Typically the terms $P(\mathcal{D})$ and $P(\mathcal{Q})$ are ignored leading to ranking by $P(\mathcal{Q}|\mathcal{D})$. Like the Ponte and Croft model [26], we model $P(\mathcal{Q}|\mathcal{D})$ as an i.i.d sampling of the query words from a document model as $P(\mathcal{Q}|\mathcal{D}) = \prod_{i=1}^{n} P(q_i|\mathcal{D})$ where,

$$\begin{aligned} P(q_i|\mathcal{D}) &= \lambda P_{mle}(q_i|\mathcal{D}) + (1-\lambda)P_B(q_i|\mathcal{D}) \\ P_{mle}(q_i|D) &= \frac{tf_{q_i,\mathcal{D}}}{|\mathcal{D}|} \\ P_B(q_i|C) &= \frac{\sum_j^N tf_{q_i,j}}{\sum_i^{|V|}\sum_j^N tf_{q_i,j}} \end{aligned}$$

Note that, if we use a context-dependent formulation, or n-grams where n > 1, we will inherit some of the qualities of the PA model. Using bigram or even trigram context has two distinct advantages: (a) documents which have more of the query terms in close proximity will be preferred, and (b) the back-off probability will help enforce phrases. In this paper, however, we only report results using unigram models.

3.6 Translation Models

Berger and Lafferty [3] proposed modeling the query as a translation of a document. As described in section 3.5, the user has a notion of an ideal document. In this model, the query formulation process can be viewed as a translation of the ideal document into a query through a noisy channel. The translation process accounts for deletion and substitution of terms from the ideal document in the query. While the basic language modeling framework described above does not allow for query expansion, this model does. More recently this model has shown success in several information retrieval tasks such as sentence retrieval and FAQ retrieval [18, 24] where the "lexical gap" between the query and document is high and the documents are short.

We add a twist to the original model by trying to estimate the probability that the document will be clicked for a query-ad pair as follows. If C is a binary random variable that takes the value 1 to indicate that a click is observed and 0 to indicate that it is not, in this model we aim to rank documents by the $P(C = 1|\mathcal{D}, \mathcal{Q})$:

$$P(C|\mathcal{D},\mathcal{Q}) = \frac{P(\mathcal{Q}|\mathcal{D},C)P(C|\mathcal{D})}{P(\mathcal{Q}|\mathcal{D})} \quad (3)$$

where

$$P(\mathcal{Q}|\mathcal{D},C) = \prod_{i=1}^{n} P(q_i|\mathcal{D},C) \quad (4)$$

$$(5)$$

$P(q_i|\mathcal{D}, C)$ can be estimated in a number of ways. We chose the following mixture model:

$$\begin{aligned} P(q_i|\mathcal{D},C) &= \lambda_1 P_{mle}(q_i|\mathcal{D}) \\ &\quad +\lambda_2 P_B(q_i|\mathcal{D}) + \lambda_3 P_{TM}(q_i|\mathcal{D},C) \end{aligned} \quad (6)$$

The first two components of $P(q_i|\mathcal{D})$ are as in section 3.5 and in estimating them we do not consider the conditional factor of the clicks. The third component, viz, P_{TM} or the translation model can be expanded as follows:

$$\begin{aligned} P_{TM}(q_i|\mathcal{D},C) &= \sum_j^{|\mathcal{D}|} P(q_i|t_j,C)P(t_j|\mathcal{D},C) \quad (7) \\ P(t_j|\mathcal{D}) &= \sum P_{mle}(t_j|\mathcal{D}) \end{aligned}$$

The key to the translation model is in estimating the translation tables which associates a probability $p(q_i|t_j, C)$ for a word pair q_i, t_j where q_i may correspond to the token "shoes" and t_j correspond to the token "sneakers". Note that self translations are also modeled, i.e. we can have $p(q_i|t_j, C)$ where $q_i = t_j$. In this way, the model assigns a non-zero probability mass to those ads for which "translations" or synonyms (t_j) of the query term q_i occur in the ad. There are several sources for deriving the translation tables: from clicked query-ad pairs, web search results, wikipedia, user sessions, etc. Smoothing the translation probability across multiple sources provides robustness and diversity of translations. As described in section 3.5, using n-gram probabilities, where n > 1, enforce term proximity and automatically capture multi-word phrases.

To implement Equation 3, we need to model two additional components: $P(C|\mathcal{D})$ and $P(\mathcal{Q}|\mathcal{D})$. $P(C|\mathcal{D})$ can be considered to be a quality score for an ad independent of the query, which can be estimated from syntactic and semantic features and the prior historical click-through-rate of the ad. $P(\mathcal{Q}|\mathcal{D})$ plays the role of IDF in the vector space approach; we estimate the statistics for this component from all ads displayed for all queries, not just the clicked query-ad pairs. The denominator in Equation 3 can be used to discriminate the clicked ads from the non-clicked ads given a query.

Note that in this paper, we only focus on deriving unigram translation tables from clicked ads. We leave the estimation of the complete $P(C|\mathcal{D}, \mathcal{Q})$ and the use of n-gram probabilities for future work.

4. EXPERIMENTAL SETUP

In this section, we describe the tools, training and test data and the evaluation.

$p(q_i = \text{yoga}\|t_j)$		$p(q_i = \text{cyst}\|t_j)$		$p(q_i = \text{acetaminophen}\|t_j)$	
ilchi	0.500	dermoid	0.466	antipyret	0.250
dahn	0.453	pilonidal	0.465	paracetemol	0.153
iyengar	0.439	bartholin	0.440	overdose	0.068
ashtanga	0.400	epidermoid	0.416	analgesic	0.037
astanga	0.384	ganglion	0.361	acetylcysteine	0.033
kriya	0.355	epiderm	0.273	pathophysiology	0.016
asana	0.354	sebaceous	0.242	caplet	0.010
hatha	0.320	popliteal	0.158	hydrocodon	0.007

Table 1: Example translation tables (from Web Search). The table shows the top terms sorted by $p(q_i|t_j)$ for 3 different t_j.

4.1 Tools

We indexed only those ads for which the advertiser had opted in for advanced match. We use a similar infrastructure to the work of Broder et al [7], i.e., we use Hadoop grid computing infrastructure to preprocess the ads and build an inverted index [13] and use the WAND algorithm to retrieve ads [6]. Stemming was done using the Porter stemmer[27]. Since our methods score each zone differently, we maintained a separate postings list for each zone. Hence, TF, IDF and other statistics can be computed for each zone. URL segmentation was done using a simple unigram model whose vocabulary was trained on a web document collection and a decoder that used a dynamic programing algorithm to retrieve the best segmentation for a given URL.

4.2 Data

Our training and test data comprised of query-ad pairs that had been judged by trained editors on a 5 point scale - *Perfect, Excellent, Good, Fair* and *Bad*. The editors only looked at the creative and not the landing page while making their judgment. The editors were trained for the task and were instructed to reserve the judgment "Perfect" to those query-ad pairs where the query has an unambiguous target (eg. "abc.com") and the ad's display url corresponds exactly to that target. This is typically true only for navigational queries. The judgment "Fair" was reserved to those query-ad pairs for which it was not completely obvious that the user would be able to find what he or she was looking for after a click, but there was a reasonable chance of doing so. The judgment criterion was quite similar to the work of Metzler and Dumais [22].

We had a set of about 47000 query-ad pairs for about 1000 unique queries that had been judged for a different system that we used as our development training set. We tuned some of the parameters for the different models on this training set. We indexed all the ads that had opted in for advanced match and that were present in our advertiser database on one day in April. We retrieved ads using BM25, PA, LM, TM, TM(Web) and the PA+BM25 models on 317 queries sampled from a query log of the first 2 weeks of April from a major search engine. We used the median score of a method as a threshold to filter out query-ad pairs so as to decrease the effort of the editors. In all we had about 10000 query ad pairs to be judged. In our final evaluation data set the number of Perfect, Excellent, Good, Fair and Bad judgments were 5, 21, 335, 1648 and 7735 respectively. 87 queries had no relevant documents (examples are *how to make a pinata* and *human body system*) and only 76 queries had greater than 10 relevant documents. In our final evaluation, we considered un-judged documents retrieved by an algorithm as non-relevant.

4.2.1 Estimating the translation tables

In this work we estimated the translation tables ($p(q_i|t_j, C)$ in Equation 7) from two sources of data.

1. Using Sponsored Search Click Data: We used a month's worth of sponsored search click data that was available to us from November of last year. The data consists of tuples of the form <query, ad, click>, where *click* indicated whether a click was observed (click = 1) for that query and ad pair or not. Clicks were spam filtered. We considered all lines where a click was observed and estimated the probability of $P(q_i|t_j)$ to be $P(q_i|t_j, click = 1)$ as follows.

$$P(q_i|t_j) = \frac{\#(q_i \in \text{query \& } t_j \in \text{ad \& click=1})}{\#(t_j \in \text{ad \& click=1})}$$

   A query and ad pair had to be clicked a minimum number of times before it was used in the computation.

2. Using Web Search Logs: Web search engines typically show a ranked listing of results for a query. The listings usually comprise of a title, an abstract of the page and a url (similar to an ad's title-creative-url). We used the Yahoo! API and obtained the summaries for the top 10 results for the top 200K unique queries on our search engine. We estimated $P(q_i|t_j, C)$ in a manner similar to what is described above, except that we used information from the web-page abstract shown on the "organic" or web-search results page and ignored the click information (we did not have the click information for the search engine and we believe it is reasonable to assume that the search engine did quite well for popular queries). In other words $P(q_i|t_j, C) \sim P(q_i|t_j)$.

Table 1 shows example translation probabilities learned using the second method listed above. We deliberately chose non-commercial examples since we did not want to label specific brands or product names and model numbers, which often show up as the top translations for many commercial queries. As can be seen from the examples, many of the expansions explore facets of a query.

The probabilities $P(q_i|t_j, C)$ can also be estimated from other sources of click data: eg., reformulations of a query where the reformulated query had a web result clicked on and so on. If these additional sources of information are available, we believe they can prove to be quite useful.

4.3 Evaluation Measures

We evaluate our work using standard measures from information retrieval and web-search such as precision at different ranks and recall. The precision at rank r, P_r is defined as the fraction of relevant documents in ranks 1 to r that are considered relevant. Recall is defined as the fraction of documents that are relevant to the query that are retrieved. Average precision of a query is computed by averaging the precision at different points of recall. In other words:

$$AP_q = \frac{\sum_{r=1}^{M} P_r \times R(r)}{\text{no. of rel docs for q}} \quad (8)$$

where M is the number of retrieved documents and $R(r)$ is a binary variable which takes on the value 1 if there is a relevant document at rank r and 0 otherwise. Mean Average Precision is the average of the average precision scores computed for all queries. Since MAP requires binary relevance judgments, we report results for two cases: (1) where every document judged "Good" or better was judged relevant and (2) where every document judged "Fair" or better was judged relevant . Anything un-judged was marked as "Bad".

Since our judgments are obtained on a five point scale, we can also compute metrics based on discounted cumulative gain (DCG) [17], a measure which aims at measuring user experience. The DCG of a query is given by:

$$DCG = \frac{\sum_{i}^{M} \frac{\text{score}(label_i)}{log_2(i+1)}}{M} \quad (9)$$

where M is the number of retrieved documents for that query and $label_i$ is the judgment of a document at rank i. The scores assigned to Perfect, Excellent, Good, Fair and Bad documents are 10, 7, 3 and 0.5 respectively. The normalized DCG (nDCG) metric for a query normalizes the DCG by the maximum value of DCG that is possible for that particular query. In addition to the average DCG of the retrieved results we report DCG at ranks 1 and 5.

5. RESULTS AND DISCUSSION

Results from our experiments are given in Table 2. Of the first two models, BM25 and PA, the BM25 model retrieves many more ads that are "Good" or better (higher recall), but the PA model does better in ranking the good ones at the top of the ranked list. Many ads can contain several 100s of bidded terms for the same creative for eg., "running shoes", "jogging shoes", "walking shoes", "comfortable walking shoes" and so on. Thus a word like "shoes" can get repeated a few 100 times inflating the TF component of the above model quite significantly. For creatives such as these the PA model does much better, but for longer queries the BM25 model appeared to do better. The model with the weighted combination of these two (PA+BM25) has a weight of 0.333 for the PA model and 0.666 for the BM25 model. Of the vector space models without expansion this model seems to perform the best.

Of the language modeling methods, the translation models are expected to decrease precision and increase recall. The translation models learned from querying the search engine [TM(web)] significantly increases the recall as expected. However, the translation models that learn from click data (TM) did not improve recall as much. The best performing model is the model that performs expansion based on the abstracts of the web-results (last column in Table 2).

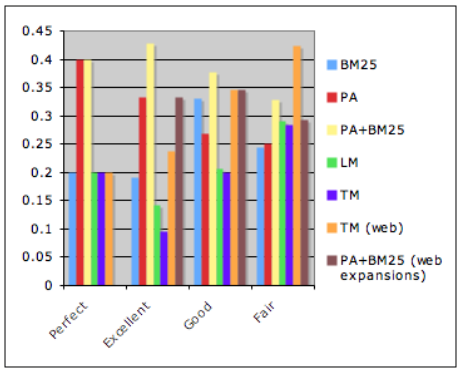

Figure 1: Fraction of "Perfect", "Excellent", "Good" and "Fair" results retrieved by each method. Note that there are only 5 "Perfect" results, 21 "Excellent" ones, and 335 and 1648 "Good" and "Fair" ones respectively.

Although there is some non significant decrease in precision and nDCG, the overall improvement in mean average precision is statistically significant. The loss in DCG comes mainly from the top ranks where the decrease is statistically significant.

The numbers in Table 2 are all micro-averaged. If we consider macro-averaged recall as shown in figure 1, we notice that the translation model using web based expansions has high recall, especially for "Fair" documents. This means that there are certain queries for which the translation models perform significantly better. An example is *amtrak and schedule* to which the addition of the terms *rail,vacation* and *adventure* and the url token *amtrak.com* results in significant increase in recall.

The translation model learned from click data did not perform as well as expected. Upon analysis, we found that many of the translations learned were from a query-word to a segment of the display url. Since the data used to train the translation probabilities was a few months older than the index on which retrieval was performed, many of the the ads containing the same tokens in the display url did not exist in the current database. Few expansions actually triggered and hence the TM model is quite similar to LM in performance. We believe that cleaner dictionaries that "expire" less easily may be constructed by considering only the query and ad-title and/or query and bidded term. Another reason for the TM model not performing well may lie in the fact that we did not do any rank normalization when we learn the translation probabilities from clicked query-ad pairs. Since exact matches typically show up higher in the display, it is likely that we are learning more self translation probabilities than synonym translations. On the other hand, with the web data, we had 10 results per query to expand with and the dictionaries learned were less noisy and had more synonym translations.

The translation model that learned from web-result abstracts does not perform as well as the PA+BM25 model that expands on the web abstracts. The principle difference here is that the translation dictionaries in the language modeling framework were not query specific. This led to much quicker "query drift". For example the query *baja* occurs in many queries that are in the context of *baja*

	BM25	PA	LM	TM	TM (Web)	PA+ BM25 (unigrams)	PA+ BM25 (web expansions)
DCG_1	0.494	**0.591**	0.452	0.448	0.398	0.522	0.334
DCG_5	1.074	**1.134**	1.049	0.987	0.940	1.095	0.667
DCG	1.456	1.406	1.494	1.438	1.415	**1.570**	1.154
NDCG	0.357	0.331	0.362	0.351	0.361	**0.384**	0.324
Good or Better is relevant							
P_1	0.218	0.276	0.206	0.206	0.178	0.237	**0.288**
P_5	0.147	0.136	0.146	0.130	0.126	0.138	**0.169**
Recall	0.318	0.241	0.296	0.268	0.301	0.338	**0.407**
MAP	0.156	0.153	0.137	0.130	0.116	0.163	**0.236**
Fair or Better is relevant:							
P_1	**0.459**	0.391	0.412	0.401	0.367	0.457	0.386
P_5	0.296	0.291	**0.312**	0.303	0.283	0.309	0.300
Recall	0.289	0.273	0.335	0.321	0.361	0.320	**0.381**
MAP	0.198	0.178	0.210	0.202	0.199	0.213	**0.246**

Table 2: Results of 7 different models on the task of retrieving ads relevant to a set of queries. Bolded values indicating the best performing system for a given metric and underlined values indicate statistical significance (at the 95% level of confidence) as compared to the baseline (BM25)

motorsports, leading to a translation dictionary that corresponds to this theme. However expanding on this theme for a query like *baja fresh* will change the context of the original query significantly. There are many solutions to work around this problem: one way is to construct query specific translation dictionaries using content of the landing page of the clicked url in the same way that the web-abstracts were used. This approach can be expensive from a storage perspective and may not work for truly tail queries. Another approach would be to embed context into the translations, explicitly via query segmentation into phrases for phrase to phrase translation tables [2], or implicitly via use of n-gram contexts in the translation probabilities as proposed earlier in this paper. Also, although we filtered out low frequency terms (t_j) in computing $P(q_i|t_j)$, the filtering was probably insufficient, leading to some incorrect high probability translations. Modeling deletions and insertions via more advanced translation models like IBM Model-3 can also be beneficial.

There are also certain classes of queries for which unigram models do not perform well. Particularly notable in our analysis were queries with geographic intent and people names. It is obvious that showing ads for *pizza in springfield, illinois* is probably not acceptable for a user located in Springfield, MA. Significant number of web-search queries have geographic intent and they should probably be handled seperately [20, 12]. Likewise showing ads for *jeniffer lopez* for the query *Jeniffer Howard* (a politician) is not acceptable. Some of these errors may be controlled by using phrase based models or using entirely different retrieval models for certain classes of queries [16].

We also played around with alternate forms of smoothing like Dirichlet smoothing that overcomes some of the document length normalization issues (note that the zone corresponding to the bidded terms can have significant variance in length). We did not see significant improvement in performance as compared to the Jelinek Mercer Smoothing method whose results we reported.

6. RELATED WORK

The information retrieval community has studied the problem of matching queries to relevant document documents for several years [34]. Queries can be long like in the TREC adhoc retrieval tasks unambiguous natural language questions or web-queries. Likewise retrieval on several types of collections have been studied. Perhaps the most relevant areas within information retrieval for this work are the following areas: the classic ad-hoc retrieval task, web document retrieval, retrieval of small snippets of text and online advertising, a fairly nascent field.

Work in online advertising focuses on two main areas: contextual advertising and sponsored search. Contextual advertising mainly concerns itself with the placement of ads on publisher pages. Since publisher pages are rich in content, a rich set of features can typically be extracted from the webpage and used to find relevant ads [5, 33]. The Sponsored Search problem on the other hand suffers from the same problem as web-search – that the queries are short and have little context. Exacerbating the problem is the fact that the document is short with little context as well. One way of overcoming this problem is though "query rewriting" techniques. The transformed query is then used for retrieval. Models to predict query rewriting techniques may be learned from query logs [19, 28]. Alternately some techniques, including ones explored in this paper, expand the query using the *organic search* or web-search results [7]. A third source of data that we do not use in this paper, but has been proven useful for sponsored search is the historical click-through-rate (CTR) of a query-ad pair in predicting its relevance to a query. We envision our system to be a two-stage one where the first stage relies less on history and the second "re-ordering" stage may use historical CTR in addition to word-overlap features [8]. This allows new advertisers who have never been shown for a given query to have a chance to be show up on the search results page. CTR information can also be incorporated into the term $P(C|\mathcal{D})$ in Eq 3.

Query expansion is generally accepted as beneficial for information retrieval. Expansion may be through pseudo rele-

vance feedback [36, 21] or interactive techniques [14]. Using the web as an external resource has proven beneficial for sponsored search and other information retrieval tasks [7, 35, 9, 10, 32] and particularly when the snippets of text that need to be matched are short [22, 32]. Ribeiro-Neto et al [29] found expanding the content of publisher pages to be useful to the problem of contextual advertising. In the field of machine learning the addition of unlabeled data to improve classification accuracy has received special attention through the fields of unsupervised and semi-supervised learning [23].

Jeon[18] and Murdock [24] recently used the translation models of Berger and Lafferty [3] for query expansion for two new tasks and found considerable improvement. While Jeon was attempting a Q&A retrieval task, Murdock was attempting a sentence retrieval task. Given the short length of sentences Murdock naturally found expansion beneficial. Murdock et al [25] also applied the translation model approach to contextual advertising. They computed translation models between publisher pages and landing pages by using a parallel corpus determined by human judgments. We presented a variation of the translation models in which the translation dictionaries use click information. Our dictionaries can easily be computed from search engine logs and therefore our method is more robust to seasonal variations in the vocabulary of commercial terms.

The work of Zhou et al [38] attempts to model $P(C|\mathcal{Q},\mathcal{D})$ in by factoring out the query into units. In their method, the query $\mathcal{Q}$ is broken into units or phrases for which high CTR for the document $\mathcal{D}$ is seen. The probability $P(C|\mathcal{Q},\mathcal{D})$ is then computed as the product of the observed CTR of the sub-phrases when issued as individual queries. However, their method does not incorporate query expansion.

7. CONCLUSIONS AND FUTURE WORK

In this paper we explored several models of information retrieval from the classic BM25 model to the translation models. The BM25F model has shown to be effective for retrieval on certain semi-structured documents like HTML pages. As a next step, we would also like to train the BM25F model that uses a separate b and f parameter for each zone and an additional saturating parameter [37]. Since the different zones in the ad creative have different properties such as length and the way in which terms repeat, we expect that tuning parameters per zone would improve performance.

The translation models in general improved recall. Expanding the original query with web search results significantly improves performance; using sponsored search clicked query-ad pairs showed less benefit. However, we believe that there is plenty of room for improvement of the translation model. We used very simple co-occurence probabilities for our translation dictionaries. The next step would be to use more sophisticated translation models including phrase based and/or ones with n-gram contexts. The translation models also lend themselves naturally to using a mixture modeling framework, where the mixtures can be over different corpora from which the translation tables are derived, or different query slices that determine semantic categorization of queries. We have not looked into using query reformulation data for learning translations, or into using non-Yahoo sources of data. Combining translation tables from different sources will improve smoothing (reinforce translations coming from multiple sources) and increase the coverage of the translation models over a larger fraction of the tail queries. Slicing queries based on query intent and/or based on query clusters may also improve targeting the translation models to a specific query ("baja fresh" and "baja motorsports" fall in different semantic categories).

We proposed a new variant of the translation model which aims at capturing the probability that a given ad will be clicked for a query. However, in this paper, we have not fully explored this model. We would like the use historical CTR of an ad and the query, combined with textual features as a prior in equation 3. Rank-normalizing the impact of clicks on ads may also play an important role in learning more synonym translations rather than self translations.

Finally, we did not explore feature-based models for retrieval in this paper. For instance, we can use the scores from our information retrieval models in models that learn from click logs like those of Richardson et al [30] and Ciaramita et al [8] either as a prior or as a feature in a machine learning model. These models are can be trained and evaluated on clickthrough-logs. We however, believe such models that are trained and optimized to perform well on historical data should be used in a re-ranking step as opposed to the initial retrieval since such models are heavily biased towards what has been seen, resulting in fewer opportunities for new advertisers to be shown on the page. Using a model that relies on historical clicks of ads for re-ranking, but not for initial retrieval would provide new advertisers the opportunity to be shown on the page, and improve their ranking if clicks are observed.

While the problem of sponsored search is relatively new and offers several new areas of exploration, we believe that several tools and techniques developed for several information retrieval tasks may be directly applied with small modifications and enhancements for the new task.

Acknowledgments

We would like to thank our colleagues in the Sponsored Search team at Yahoo! and the reviewers of this paper for useful feedback.

8. REFERENCES

[1] `www.emarketer.com/Article.aspx?id=1006319`.

[2] L. Ballesteros and W. B. Croft. Phrasal translation and query expansion techniques for cross-language information retrieval. *SIGIR Forum*, 31(SI):84–91.

[3] A. Berger and J. Lafferty. Information retrieval as statistical translation. In *SIGIR '99: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 222–229, New York, NY, USA, 1999. ACM.

[4] A. Broder. A taxonomy of web search. *SIGIR Forum*, 36(2):3–10, 2002.

[5] A. Broder, M. Fontoura, V. Josifovski, and L. Riedel. A semantic approach to contextual advertising. In *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 559–566, New York, NY, USA, 2007. ACM.

[6] A. Z. Broder, D. Carmel, M. Herscovici, A. Soffer, and J. Zien. Efficient query evaluation using a two-level retrieval process. In *CIKM '03: Proceedings of the twelfth international conference on Information and*

knowledge management, pages 426–434, New York, NY, USA, 2003. ACM.

[7] A. Z. Broder, M. Fontoura, E. Gabrilovich, A. Joshi, V. Josifovski, and T. Zhang. Robust classification of rare queries using web knowledge. In *SIGIR '07*, pages 231–238, New York, NY, USA, 2007. ACM.

[8] M. Ciaramita, V. Murdock, and V. Plachouras. Online learning from click data for sponsored search. In *WWW '08: Proceedings of the 16th international conference on World Wide Web*, 2008.

[9] F. Diaz and D. Metzler. Improving the estimation of relevance models using large external corpora. In *SIGIR '06*, pages 154–161, New York, NY, USA, 2006. ACM.

[10] S. Dumais, M. Banko, E. Brill, J. Lin, and A. Ng. Web question answering: is more always better? In *SIGIR '02*, pages 291–298, New York, NY, USA, 2002. ACM.

[11] B. Edelman, M. Ostrovsky, and M. Schwarz. Internet advertising and the generalized second-price auction: Selling billions of dollars worth of keywords. *American Economic Review*, 97(1):242–259, March 2007.

[12] Q. Gan, J. Attenberg, A. Markowetz, and T. Suel. Analysis of geographic queries in a search engine log. In *LOCWEB '08: Proceedings of the first international workshop on Location and the web*, pages 49–56, New York, NY, USA, 2008. ACM.

[13] Hadoop. `http://hadoop.apache.org/core/`.

[14] D. Harman. Towards interactive query expansion. In *SIGIR*, pages 321–331, 1988.

[15] B. Jansen and M. Resnick. Examining searcher perceptions of and interactions with sponsored results. In *Workshop on Sponsored Search Auctions*, 2005.

[16] B. J. Jansen, D. L. Booth, and A. Spink. Determining the user intent of web search engine queries. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 1149–1150, New York, NY, USA, 2007. ACM.

[17] K. Järvelin and J. Kekäläinen. Ir evaluation methods for retrieving highly relevant documents. In *SIGIR '00*, pages 41–48, New York, NY, USA, 2000. ACM.

[18] J. Jeon. *Searching Question and Answer Archives*. PhD thesis, University of Massachusetts, Amherst, 2007.

[19] R. Jones, B. Rey, O. Madani, and W. Greiner. Generating query substitutions. In *WWW '06: Proceedings of the 15th international conference on World Wide Web*, pages 387–396, New York, NY, USA, 2006. ACM.

[20] R. Jones, W. V. Zhang, B. Rey, P. Jhala, and E. Stipp. Geographic intention and modification in web search. *International Journal of Geographical Information Science*, 22(3):229–246, 2008.

[21] V. Lavrenko and W. B. Croft. Relevance based language models. In *SIGIR '01: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 120–127, New York, NY, USA, 2001. ACM.

[22] D. Metzler, S. Dumais, and C. Meek. Similarity measures for short segments of text. *Advances in Information Retrieval*, pages 16–27, 2007.

[23] T. M. Mitchell. *Machine Learning*. McGraw-Hill Higher Education, 1997.

[24] V. Murdock. *Aspects of Sentence Retrieval*. PhD thesis, University of Massachusetts, Amherst, 2007.

[25] V. Murdock, M. Ciaramita, and V. Plachouras. A noisy-channel approach to contextual advertising. In *ADKDD '07: Proceedings of the 1st international workshop on Data mining and audience intelligence for advertising*, pages 21–27, New York, NY, USA, 2007. ACM.

[26] J. M. Ponte and W. B. Croft. A language modeling approach to information retrieval. In *SIGIR '98*, pages 275–281, New York, NY, USA, 1998. ACM.

[27] M. 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 *SIGIR '08*, 2008.

[29] B. Ribeiro-Neto, M. Cristo, P. B. Golgher, and E. S. de Moura. Impedance coupling in content-targeted advertising. In *SIGIR '05*, pages 496–503, New York, NY, USA, 2005. ACM.

[30] M. Richardson, E. Dominowska, and R. Ragno. Predicting clicks: estimating the click-through rate for new ads. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 521–530, New York, NY, USA, 2007. ACM.

[31] S. E. Robertson and S. Walker. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *SIGIR '94*, pages 232–241, New York, NY, USA, 1994. Springer-Verlag New York, Inc.

[32] M. Sahami and T. D. Heilman. A web-based kernel function for measuring the similarity of short text snippets. In *WWW '06: Proceedings of the 15th international conference on World Wide Web*, pages 377–386, New York, NY, USA, 2006. ACM.

[33] W. tau Yih, J. Goodman, and V. R. Carvalho. Finding advertising keywords on web pages. In *WWW '06: Proceedings of the 15th international conference on World Wide Web*, pages 213–222, New York, NY, USA, 2006. ACM.

[34] TREC. `http://trec.nist.gov`.

[35] E. Voorhees. Overview of the trec 2004 robust retrieval track. In *Proceedings of the 14th Text REtrieval Conference*, 2005.

[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, 2000.

[37] H. Zaragoza, N. Craswell, M. Taylor, S. Saria, and S. Robertson. Microsoft cambridge at trec-13: Web and hard tracks. In *TREC-2004*, 2004.

[38] D. Zhou, L. Bolelli, J. Li, C. L. Giles, and H. Zha. Learning user clicks in web search. In *IJCAI*, pages 1162–1167, 2007.