Clicked Phrase Document Expansion for Sponsored Search Ad Retrieval

Dustin Hillard
Yahoo! Inc
Great America Parkway
Santa Clara, CA, 95054
dhillard@yahoo-inc.com

Chris Leggetter
Yahoo! Inc
Great America Parkway
Santa Clara, CA, 95054
cjl@yahoo-inc.com

ABSTRACT

We present a document expansion approach that uses Conditional Random Field (CRF) segmentation to automatically extract salient phrases from ad titles. We then supplement the ad document with query segments that are probable translations of the document phrases, as learned from a large commercial search engine's click logs. Our approach provides a significant improvement in DCG and interpolated precision and recall on a large set of human labeled query-ad pairs.

Categories and Subject Descriptors

H.3.3 [**Information Systems**]: Information Search and Retrieval—*General*

General Terms

Algorithms, Experimentation

1. INTRODUCTION

The primary source of revenue for major search engines is textual advertising. A search engine typically displays sponsored listings along with organic web-search results in response to user queries. The revenue model for these listings is "pay-per-click" where the advertiser pays the search engine only if the advertisement is clicked. A search engine typically decides which ads to show (and in what order) by optimizing revenue based on the probability that an ad will be clicked, combined with the cost of the ad [5].

The advertiser "targets" specific keyword markets by bidding on search queries. An advertising **campaign** consists of many **ad groups** where each ad group in turn consists of a set of **bidded phrases** or keywords that the advertiser bids on. A **creative** is associated with an ad group and is composed of a **title**, a **description** and a **display URL**. The title is typically 3-4 words in length and the description has about 10-15 words.

An advertiser can choose to use **standard** or **advanced** match for the keywords in an ad group. For example, enabling only standard match for the keyword "sports shoes", will result in the corresponding creative being shown only for that exact query. Whereas, if the keyword is enabled for advance match, the search engine can show the same ad for the related queries "running shoes" or "track shoes". A **bid** is associated with each keyword and a second price auction model determines how much the advertiser pays the search engine for the click [3].

Most search engines typically take a two pronged approach to the problem: (1) finding relevant ads for a query, and (2) estimating CTR for the retrieved ads and appropriately ranking those ads for display on the search page. In this work we are largely concerned with the first point of discovering ads for a query.

Finding ads relevant to a query is an information retrieval problem and the nature of the queries makes the problem very similar to web-search, with some key differences. The collection of web documents is significantly larger than the advertiser database and retrieving candidate ads for infrequent queries is a very important area of research for sponsored search. A relevant sponsored search result may also match the query in a broader sense than when compared to web search: an ad for "limo rentals" could be a good match to a query for "prom tux", although the match would be too coarse for a typical web search.

In this paper we present an approach for document expansion that improves retrieval of sponsored results by mining user click logs for high click-rate phrase translations. We expand ad documents by automatically extracting title phrases and then including query segments that have historically high click rates in the context of an ad's title phrases.

2. PHRASE-BASED DOCUMENT EXPANSION

Our document expansion approach consists of three main components: CRF segmentation of the user queries and ad titles, collection of click-rate statistics for query-title segment pairs, and expanding ad documents by inserting high click-rate query phrases.

Linear chain CRFs have been used successfully for many sequential labeling tasks like segmentation and part-of-speech tagging, as well as in web search[4]. CRFs are particularly attractive because we can use arbitrary feature functions on the observations. Let $Q = q_1, q_2...q_3$ denote an input query and $seg = y_1, y_2...y_n$ denote the corresponding state sequence. Each y_i falls into one of two categories *begin-segment* and *end-segment*. The conditional probability $P(seg|Q)$ is given as:

$$p(seg|Q;\Lambda) = \frac{1}{Z(Q;\Lambda)} exp\{\sum_k \lambda_k \sum_i^n f_k(y_{i-1}, y_i, Q, i)\} \quad (1)$$

where $f_k(y_{i-1}, y_i, Q, i)$ is a feature function and $\Lambda = \{\lambda_i\}$ are

SIGIR'10, July 19–23, 2010, Geneva, Switzerland.
ACM 978-1-60558-896-4/10/07.

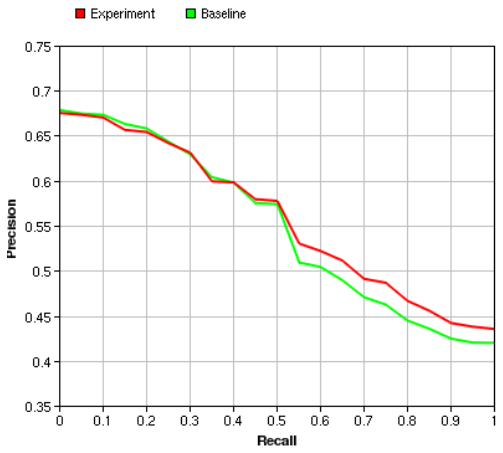

Figure 1: Interpolated Precision Recall

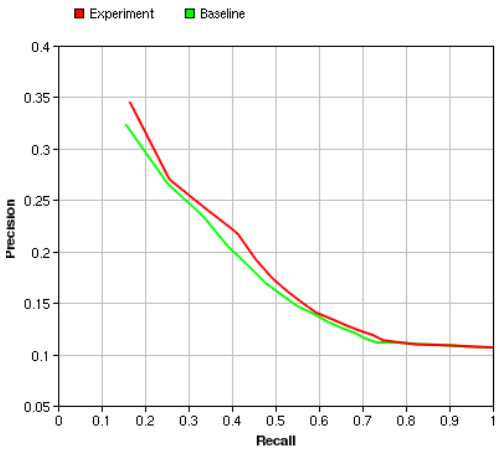

Figure 2: Precision Recall

the learned feature weights. We have several features based on dictionaries of people names, celebrity names, stopword lists and the top phrases in the query logs. We use the CRF++ toolkit [1] to train the model. The model is trained in a supervised fashion on 6000 queries annotated with phrases by humans.

We then apply this segmentation model to two months of sponsored search web logs, obtaining the top three hypotheses for every query and title. For each unique query-title segment pair we count the number of user clicks and expected clicks, where expected clicks is a position normalized count of impressions [6]. The result is a translation table of click-rates for all observed query-title segment pairs.

Finally, for each title segment in an ad we augment our ad document by adding all query segments that occurred in the context of that title segment as confident translations (we threshold on a click-rate of 1.5 and translation probability of 0.05, optimized on a held-out data set). In order to weight the query segments according to their click-rate we multiply their TF contribution by their click-rate in that context, so that additional weight is given to segments with high click rates. The final index contains ad documents expanded with weighted query segments that either reinforce current document phrases or add additional related phrases. Our approach is related to much previous work in using search logs as implicit relevance feedback (such as [2]), but our novel contribution is to focus on CRF extracted phrase terms and normalize clicks by expected clicks.

3. RESULTS

Our baseline retrieval system is a TF-IDF based ad retrieval system similar to [1]. We index our advertiser database, where each ad document consists of a title, description, URL, and bid terms. Our experiments compare the standard baseline system to the document expansion approach described in Section 2.

The test data contains 1k unique queries, which were selected based on a stratified sample of search engine traffic that represents all ten search frequency deciles. An average of 20 results per query were retrieved for the baseline and experimental system (the retrieved documents from the two systems overlapped by about 60%). A total of about 30k unique query-ad pairs were judged by human editors on a standard five point relevance scale. The document expansion experiment provides a 8% relative gain for DCG@1, and 3% relative gain for DCG@3.

Figure 1 shows a 2% absolute gain in recall for the region of higher recall performance. Figure 2 additionally shows the document expansion has improved the classification accuracy in terms of detecting Bad ad documents. Classification is particularly important for sponsored search because the retrieval stage typically presents an unordered candidate set that is later re-ranked by a click model.

[1] `http://crfpp.sourceforge.net/`

4. CONCLUSION

We develop an approach to document expansion that utilizes automatic CRF phrase extraction, leverages normalized click-rates from web logs, and augments ad documents with weighted query phrases. The approach provides gains in both relevance ranking and classification of relevant ads.

5. REFERENCES

[1] A. Z. Broder, D. Carmel, M. Herscovici, A. Soffer, and J. Zien. Efficient query evaluation using a two-level retrieval process. In *CIKM*. ACM, 2003.

[2] H. Cui, J.-R. Wen, J.-Y. Nie, and W.-Y. Ma. Probabilistic query expansion using query logs. In *WWW*, 2002.

[3] 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), March 2007.

[4] X. Li, Y.-Y. Wang, and A. Acero. Extracting structured information from user queries with semi-supervised conditional random fields. In *SIGIR*, pages 572–579, 2009.

[5] M. Richardson, E. Dominowska, and R. Ragno. Predicting clicks: estimating the click-through rate for new ads. In *WWW*, 2007.

[6] W. V. Zhang and R. Jones. Comparing click logs and editorial labels for training query rewriting. In *WWW 2007 Workshop on Query Log Analysis: Social And Technological Challenges*, 2007.