Ads Keyword Rewriting Using Search Engine Results

Javad Azimi
Microsoft Inc.
Sunnyvale, CA, USA
jaazimi@microsoft.com

Adnan Alam
Microsoft Inc.
Redmond, WA, USA
adnanal@microsoft.com

Ruofei Zhang
Microsoft Inc.
Sunnyvale, CA, USA
bzhang@microsoft.com

ABSTRACT

Paid Search (PS) ads are one of the main revenue sources of online advertising companies where the goal is returning a set of relevant ads for a searched query in search engine websites such as Bing. Typical PS algorithms, return the ads which their Bided Keywords (BKs) are a subset of searched queries or relevant to them. However, there is a huge gap between BKs and searched queries as a considerable amount of BKs are rarely searched by the users. This is mostly due to the *rare* BKs provided by advertisers. In this paper, we propose an approach to rewrite the rare BKs to more commonly searched keywords, without compromising the original BKs intent, which increases the coverage and depth of PS ads and thus it delivers higher monetization power. In general, we first find the relevant web documents pertaining to the BKs and then extract common keywords using the web doc title and its summary snippets. Experimental results show the effectiveness of proposed algorithm in rewriting rare BKs and consequently providing us with a significant improvement in recall and thereby revenue.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; D.2.8 [**Machine Learning**]: [Performance Measures]

General Terms

Algorithm, Design.

Keywords

Paid Search, Rare Bided Keyword, Search Engine.

1. INTRODUCTION

Typical Paid Search (PS) algorithms return ads to the web search engines in which the returned ads Bided Keywords (BKs) are usually a subset of searched queries or relevant to them. The more similar the returned ads BKs are to the searched queries, the higher Click Through Rate (CTR) is expected which consequently optimizes the revenue [4]. However, we have found that a large portion of ads in our

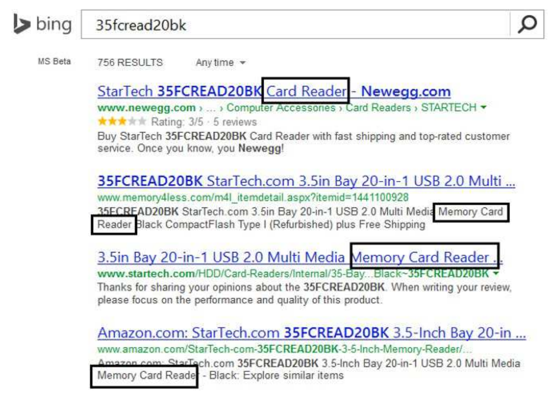

Figure 1: Bing search results for 35fcread20bk.

ads store are never been placed since their BKs are not existed in search queries. This phenomena is mostly due to the rare BKs provided by advertisers while their ads usually represent some popular products. For example, the BK *35fcread20bk* is one of the rare ads BK in our data set while it simply represents a multi media memory card reader.

The goal of this work is presenting an algorithm to rewrite the rare BKs to common keywords in order to increase their chances to be matched with more commonly searched user queries. In one of our implementations, given a rare BK, we query for the relevant Web docs using a web search engine. Alternative methods that consider advertiser meta-data (beyond the BKs) in ranking and retrieving relevant docs have also been tried. However, the process of retrieving relevant docs is not covered as part of this paper. This paper describes the approach we took to extract representative keywords from a given relevant web doc's title and snippets. For example, Figure 1 shows the top four Bing search results for the BK *35fcread20bk*. As the figures shows, the phrase *memory card reader* exists in most of the titles and snippets, and it looks a good representative rewriting keyword for *35fcread20bk*. Therefore, this ad can be considered for the queries about *memory card reader* which increases its selection chances significantly as *memory card reader* is a more common phrase in searched queries than *35fcread20bk*. Offline and Online experiments show the effectiveness of proposed algorithm in 1)rewriting the rare BKs to common keywords, 2)impressing a large number of ads with rare BKs, and 3)increasing the recall and revenue significantly.

Extracting keywords from a document is a well-studied problem in contextual advertising in which the document is assumed to follow some grammatical rules [2]. However, this assumption is not true in our problem as the returned titles and snippets are not usually following grammatical rules.

WWW 2015 Companion, May 18–22, 2015, Florence, Italy.
ACM 978-1-4503-3473-0/15/05.
http://dx.doi.org/10.1145/2740908.2742739.

There are also some studies expanding the rare queries to common queries [1] while we are tying to rewrite the ads rare BKs to common keywords for matching with queries.

1.1 Proposed Approach

In this section we describe our four steps algorithm which generates common keywords from a given rare Bk.

1.2 Scraping

First, the ads BKs that have not been selected by the current system, during the last three months, are selected as rare BK candidates. Note that, we filter out the ads that explicitly have location terms, such as cities name, in their BKs. Then, we scrape the web search engine Bing in order to get the search results for these BKs. For each BK, we only consider top 20 search title and snippet results to balance efficiency and quality.

1.3 N-gram Generation

In this step, we first extract all the n-grams for $n = 2, 3, 4$ and their frequencies from the title and snippet search results. Then, a frequency filtration is applied based on the number of tokens of original BK (l_{bk}), and generated n-grams l_2, l_3, l_4. For example, if $l_{bk} \leq 5$, then all n-grams with $l_2 \geq 8$, $l_3 \geq 6$, $l_4 \geq 5$ are returned as candidate keywords. The threshold numbers over the n-grams frequencies are linearly increased with respect to the length of the original BK. Note that, we only generate n-grams with $n \leq 4$ to increase the chance of being matched with searched queries as most of the search queries have less than five tokens.

1.4 Entity Filtration

The extracted n-grams are sometimes noisy as the correlation between tokens has been not considered yet. One approach to remove the noisy n-grams is filtering out the n-grams which do not preserve the bigram, $n = 2$, entities. A bigram is considered as entity if 1)it appears in more than 80% of title and snippet search results, and 2)its both tokens always comes together in all search title and snippet results. Note that, the entities of each BK are extracted only from its search results and independent from other BKs search results. This approach customizes the entity detection for each BK which increases the accuracy of selected entities. Finally, all candidate n-grams that break the entity tokens are removed from the candidate keywords. Table 1 shows a few BKs and their corresponding extracted entities.

Table 1: Bided Keyword and Extracted Entity

Original BK	Extracted Entity
bad credit rating auto loan finance	bad credit
buy world cup ticket 2014	world cup
afb veterans mesothelioma attorney	air force
35fcread20bk	memory card

1.5 DSSM Filtration

As another level of filtering irrelevant candidate keywords, we use a deep learning neural network model [3] to measure the similarity of candidate keywords and original BKs using one year historical clicked data. In general, [3] transforms a document to a low-dimension vector using a latent semantic model. We then measure the cosines similarity of two documents, original BK and generated candidate keyword, as their similarity. Then, the candidate keywords with similarity less than 0.75 with their corresponding original BK are filtered out from the candidate set. Experimental results show that around 25% of candidate keywords are removed at this step. The candidate keywords that pass this step are considered as representative keywords for the original BK.

2. EXPERIMENTS

In this section we evaluate the effectiveness of proposed approach using different measurements.

2.1 Manual Evaluation

First, we look at the generated keywords for some BKs in order to see the capability of proposed algorithm to cover different types of rare BKs. Table 2 represents some examples in which the first column is the original BK and the second column is the proposed algorithm generated keyword.

Table 2: Original BKs vs. Generated Keyword

Original BK	Generated Keyword
35fcread20bk	memory card reader
canine Iris melanoma	dogs eye cancer
casinoroulettegame	casino game roulette
buy www.seatgeek.com show ticket	seatgeek show ticket
007fragrances	james bond body sprays

2.2 Human Labeling Evaluation

We selected 2000 pairs of <original BK, generated keyword> randomly and asked human experts to label them as Bad, Fair, Good, Excellent and Not Sure. As the results, we had 74.50% good, 10.20% Excellent, 9.80% Bad, 5.00% Fair and 0.50% Not sure pairs. This shows that the proposed algorithm is very successful in rewriting the rare BKs to common keywords without loosing the original BK intent.

2.3 Online Evaluation

We also evaluated our algorithm using online traffic in Bing. The results show +1.27% improvement in RPM (Revenue Per Million Impression) comparing to the current system which is very significant.

2.4 Coverage Evaluation

Our further investigation shows that the additional online experiment revenue mostly comes from increasing the coverage of unimpressed ads. The results show that we are able to impress more than $5M$ unimpressed ads in a period of two weeks. These ads have been existed in our ad store for more than three months and did not have a chance to be impressed. This again indicates the effectiveness of proposed algorithm in generating common keywords from rare BKs.

3. CONCLUSION

In this paper we proposed a novel approach that rewrites the advertiser rare BKs to common keywords in order to match with user search queries. We first scrape Bing to get the web search engine results as a knowledge base for each BK and then extract some keywords from the search title and snippet results. Experimental results demonstrated the effectiveness of proposed approach in rewriting the rare BKs to common keyword preserving the intent of original BKs.

4. REFERENCES

[1] A. Broder, P. Ciccolo, E. Gabrilovich, V. Josifovski, D. Metzler, L. Riedel, and J. Yuan. Online expansion of rare queries for sponsored search. In *WWW*, 2009.

[2] K. S. Dave and V. Varma. Pattern based keyword extraction for contextual advertising. In *CIKM*, 2010.

[3] P.-S. Huang, X. He, J. Gao, L. Deng, A. Acero, and L. Heck. Learning deep structured semantic models for web search using clickthrough data. CIKM, 2013.

[4] S. Kiritchenko and M. Jiline. Keyword optimization in sponsored search via feature selection. In *JMLR*, 2008.