

Enhancing Mobile Search Using Web Search Log Data

Yoshiyuki Inagaki, Jiang Bian, Yi Chang
Yahoo! Labs
Sunnyvale, CA 94089
{inagakiy, jbian, yichang}@yahoo-inc.com

Motoko Maki
Yahoo! Inc.
Sunnyvale, CA 94089
motoko@ yahoo-inc.com

ABSTRACT

Mobile search is still in infancy compared with general purpose web search. With limited training data and weak relevance features, the ranking performance in mobile search is far from satisfactory. To address this problem, we propose to leverage the knowledge of Web search to enhance the ranking of mobile search. In this paper, we first develop an equivalent page conversion between Web search and mobile search, then we design a few novel ranking features, generated from the click-through data in web search, for estimating the relevance of mobile search. Large scale evaluations demonstrate that the knowledge from web search is quite effective for boosting the relevance of ranking on mobile search.

Categories and Subject Descriptors

H.3.3 [Information Systems]: Information Search and Retrieval - Retrieval functions

General Terms

Experiments. Measurements

Keywords

Mobile Search, Ranking, Ranking Features

1. INTRODUCTION

With rapid growth of mobile products in recent years, an increasing number of users become used to conducting search on mobile devices, therefore how to provide effective mobile search has become a critical research topic. Compared with traditional web search, there are a few unique characteristics for mobile search, which bring some new challenges. First, the texts in mobile pages are usually short and concise for the purpose of fitting in smaller screens of mobile devices. And, as Yamauchi et al. [5] pointed out, mobile users also prefer concise mobile content. This property of mobile search reduce the effectiveness of text matching features though they play an important role in web search. Second, anchor text features, which is another important category of features for web search, become not reliable in mobile search, because mobile pages contain much fewer anchor texts than Web pages. Moreover, click-through logs, known to be useful to estimate relevance in web search (e.g., [3]), is not rich enough in mobile search, partly due to *good abandonment* [4].

As mobile search is still in its infancy, we propose to take advantage of the rich knowledge from web search to improve mobile search. Although mobile search is quite different with web search, there is still a strong connection between them: If a web page should be ranked at high position for query on web search, its equivalent mobile page should be similarly ranked at high position for the same query on mobile search. For example, for query, "Toyota", if `http://toyota.jp/` is ranked at the top position, its equivalent mobile page, `http://toyota.jp/mb/m/top`, should be ranked at top in mobile search.

As mobile search is growing fast, an increasing amount of web pages already have their mobile equivalence, such as many corporate official homepages and portals of online games and social network service (SNS). In order to use such equivalence to benefit mobile search, it is necessary to accurately convert Web host/URL names to mobile host/URL names so as to promote corresponding mobile pages in mobile search. As shown by a couple of examples in Table 1, there is no systematic way for the host name conversion, but the domain name conversion is less complicated. In this paper, we propose to use domain name matching to find equivalent mobile pages for traditional web pages and boost those mobile pages whose equivalence have high relevance in web search for the same query. In next section, we will describe in details about how to convert web page into mobile ones, followed by creating novel features based on knowledge from click-through in web search to benefit ranking relevance in mobile search.

Table 1: Example: Name conversion between Web host names and mobile host names.

web host name	mobile host name
www.yahoo.co.jp	mobile.yahoo.co.jp
www.google.co.jp	m.google.co.jp
ja.wikipedia.org	mwkp.fresheye.com
www.grj.jp	www.grj-m.jp

2. TECHNICAL APPROACH

In this section, we first introduce our method for domain/URL conversion driven by click-through log in web search, then based on such conversion, we propose several new features that can be used to boost mobile search relevance.

To leverage knowledge from web search, we first accumulate a three-months click-through logs from a commercial web search. Then, we select a set of samples of $\langle \text{query}, \text{URL} \rangle$ pairs whose clicks are higher than a reasonable threshold values (which are to be empirically determined). From the selected URLs, domain names were extracted and converted into mobile domain names by looking up a domain name conversion dictionary whose entries look like examples in Table 1. If a web domain name does not match

in the name conversion dictionary, we use the web domain name and slightly shortened domain name, e.g. XXXX.jp instead of XXXX.co.jp, as potential mobile domain names.

We then find all web search queries that have also been issued in mobile search. Each of these queries is stored in a *query-domain* dictionary together with correlated both web and mobile domain names and click-through data corresponding to the original web domains. When a query is issued to a mobile search engine, the system will look up the query-domain dictionary: if matched, the query is annotated with a list of domains retrieved from the dictionary for the query and sent to the mobile search engine. The search engine will try to match the domain name of a retrieved mobile page to the one in the list of domains annotating the query. If there is a name match, we propose to use additional novel ranking features to estimate relevance of this mobile page.

In particular, we propose two new ranking features as described below. Let $dmatch1(q, u)$ and $dmatch2(q, u)$ be the new domain name matching features for a query, q , and a mobile URL, u . Let $dom(q)$ be a set of domain names retrieved from the dictionary for q . Moreover, let d_u be the domain name of u .

Pseudo-Navigational Indicator:

$$dmatch1(q, u) = \begin{cases} 1/n & : \text{if } d_u \in dom(q) \\ 0 & : \text{otherwise} \end{cases} \quad (1)$$

where n is the number of clicked domain names retrieved from the dictionary for the query.

Query-Domain Level Percent of Click:

$$dmatch2(q, u) = \begin{cases} DPCTCLK(q, d_u) & : \text{if } d_u \in dom(q) \\ 0 & : \text{otherwise} \end{cases} \quad (2)$$

where $DPCTCLK(q, d_u)$ is a domain-level percent of click for the query-domain pair, (q, d_u) .

The proposed new features are combined with other features in the machine learned framework for training a ranking algorithm. We employ a state-of-the-art algorithm Gradient Boosted Decision Tree [1] to learn the ranking function.

3. EXPERIMENTS

Since these two features are new and not available for ranking in the mobile products' setting, the experiments are purely based on evaluations on the offline data sets.

• Feature Generation

We gathered web search click-through logs from December, 2009, to February, 2010. And, training and testing data for mobile search were collected throughout the year 2009. We empirically set the minimum numbers of clicks and views of URLs for a query as 5 and 30 respectively to filter $(query, URL)$ pairs, and generated the dictionary of clicked domains and feature values. The generated features covered about 5.0% of the training and holdout data sets that are 150K query-URL pairs with 11460 queries and the 27K query-URL pairs with 1005 queries, respectively.

• Evaluation Metric

In our experiments, we use Normalized Discounted Cumulative Gain (NDCG)[2] to evaluate ranking performance. In the following, we only report $NDCG_5$.

• Results

We compare the ranking model that trained using our new features with the baseline model trained without using any new features. Figure 1 shows $NDCG_5$ of the baseline model as well as $rmNDCG_5$'s of three models that either used one of new features or used both.

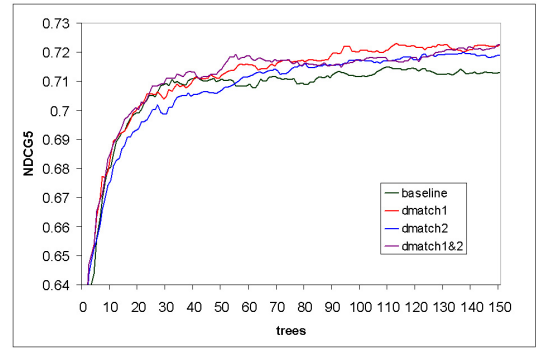


Figure 1: NDCG5 v.s. the number of trees in a machine learned model with mobile click feature.

Statistically significant $NDCG_5$ improvement was achieved if we use mobile click features that were generated from mobile click logs in the same manner as new features were generated (see Table 2). We will need to conduct online tests in order to see whether and how much new features will truly improve mobile search. It is usually the case that we introduce new rank features, new machine learned models often bring up new pages that have never been brought up.

Table 2: NDCG5 improvement over the baselines.

NDCG5	without mobile click	with mobile click
baseline	0.7087	0.7150
<i>dmatch1</i>	0.7124	0.7232
<i>dmatch2</i>	0.7146	0.7201
<i>dmatch1&2</i>	0.7137	0.7226

4. CONCLUSIONS AND FUTURE WORK

In this paper, we proposed and implemented rank features to enhance mobile search ranking by boosting the mobile pages equivalent to top ranked and well-clicked web pages. Our experiments show promising results on popular queries that are common to both web search and mobile search. However, due to the low coverage of such kind of queries, this method can not be extended to long tail queries. In future, we will investigate how to conduct search on long tail queries in mobile search by extending the idea of leveraging the knowledge from web search.

5. REFERENCES

- [1] J. H. Friedman. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5):1189–1232, 2001.
- [2] K. Jarvelin and J. Kekalainen. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems*, 20(4):422–446, 2002.
- [3] T. Joachims. Optimizing search engines using clickthrough data. In *Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (KDD)*, 2002.
- [4] J. Li, S. B. Huffman, and A. Tokuda. Good abandonment in mobile and pc internet search. In *Proceedings of the 32nd Annual ACM SIGIR Conference*, 2009.
- [5] K. Yamauchi, W. Chen, and D. Wei. A study on japanese mobile phone market and its applications. *the 4th International Conference on Computer and Information Technology*, pages 875–878, 2004.