Personalize Web Search Results with User's Location

Yumao Lu Fuchun Peng Xing Wei Benoit Dumoulin
Yahoo Inc.
701 First Avenue
Sunnyvale, CA, 94089
{yumaol,fuchun,xing,benoitd}@yahoo-inc.com

ABSTRACT

We build a probabilistic model to identify implicit local intent queries, and leverage user's physical location to improve Web search results for these queries. Evaluation on commercial search engine shows significant improvement on search relevance and user experience.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Algorithms, Experimentation

Keywords

query log analysis, personalized search

1. INTRODUCTION

There is a huge amount of searches on the Web has local intent, meaning that they are searching for things in a particular area, like restaurant, job listings, shopping center, etc. Some queries have explicit location information in the query like "*pizza hut palo alto*", while many others do not but still expect search engines to return localized search results, like "*laundry service*". This kind of queries is considered as implicit local intent queries. For such queries, users expect personalized search results [2] that are customized to their locations. Thus, identifying implicit local intent queries and finding out the location information for the users are particularly useful to improve user search experience.

Yi et al. [3] uses language modeling approach to identify implicit local intent queries. In our work we also use language modeling approach, however instead of building a language model for each city, we only build one single language model for all locations to avoid sparsity. We then obtain user's location directly from IP address mapping. What further makes our work unique is that we integrate the identified implicit location information into ranking directly to improve search relevance and prove the practical impact of this work. As far as we know, there is little work has been published in this front.

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

2. IMPLICIT LOCAL INTENT DISCOVERY

Let Q denotes the general query set. A conditional random field (CRF) based named entity tagger is used to tag all the queries in Q. We select all queries that contain a location from Q to form a new query set Q_L. We then remove all location components from queries in Q_L and form an artificial query set Q_{LC}. We define the probability that a query q has implicit local intent by

$$P(\text{implicit local intent}|q) = \frac{P_{Q_{LC}}(q)}{P_Q(q)}, \quad (1)$$

where $P_{Q_{LC}}(q)$ and $P_Q(q)$ can be estimated using n-gram language model from corpus Q_{LC} and Q respectively.

Equation (1) however does not always give an accurate estimation on $P(\text{implicit local intent}|q)$ since locations are often used as constraints for queries that do not have local intent. For example, query "*John San Jose mypace*" implies the user is looking for some one named *John* in *San Jose* from *myspace*. A popular domain name "*myspace*" is a strong indicator for such case. We create a feature called highest domain rank for each query

$$R_d^m(q) = \min_{s \in q} R_d(s) \quad (2)$$

where s is a substring of query q and $R_d s$ is the rank of domain s, which is ranked by the accumulated number of clicks on the documents that belong to domain s. The lower value $R_d(s)$ is, the more popular domain s is. In implementation, we only keep tracking top 1000 domains for simplicity.

A implicit local query may not contain general local intent which should return same results even when queries are from different locations. For example, "*Google headquarter office*" contains implicit local query but the intent is restricted to a specific location as there is only one Google headquarter office. To capture the general local intent, the entropy of the conditional probability distribution $P(l|q)$, where l is a specific location associated with query q that is identified from user IP address (described later). To calculate $P(l|q)$, we first normalize set Q_L to set $\bar{Q_L}$ such as all locations are disambiguated and canonicalized. For example, "*CA*" is canonicalized to "*California*", "*la*" is canonicalized to "*Los Angeles*" or "*Louisiana*", "*New York*" is canonicalized to "*New York City*" or "*New York State*" based on the context. The conditional probability distribution $P(l|q)$ can be then estimated through n-gram language model estimated in corpus $\bar{Q_L}$:

$$P(l|q) = \frac{P_{\bar{Q_L}}(l,q)}{P_{\bar{Q_L}}(q)}, \quad (3)$$

where $P_{\bar{Q_L}}(l, q)$ is the probability of location l and a query q co-occur estimated in corpus $\bar{Q_L}$. The entropy $E(q)$ is defined as

$$E(q) = -\sum_{l} P(l|q) \log p(l|q). \quad (4)$$

If $E(q)$ is high, the query q is more likely to be associated with many locations with similar possibilities; otherwise, q is biased towards certain locations.

Gaussian kernel support vector machine is used as the classifier. We first train a weak classifier with 5000 editorially labeled random query set. For each query, we generate features based on (1) (2) and (4). The resulting weak classifier based on the 5000 training data is used to label 100,000 queries. Queries that lie between the margin (with non-zero α values are selected for the next batch of editorial test. The classifier is then retrained with combined labeled data.

After a query is classified as having implicit general local intent, we use user's IP address to obtain the city name and zip code where the query is sent from. For ambiguous city names that exist in multiple states (such as *Oakland*) or that have different meanings (such as *Mountain View*), we add use user location's state name for disambiguation.

3. PERSONALIZE WEB SEARCH RESULTS

3.1 Personalization by Query Rewriting

To personalize search results given user's location, one intuitive way is to expand the original query by user's location. For example, if the original query is "*Italian restaurants*" and the user's location has been determined to be "*San Francisco*", a new search query may be formed as "*Italian restaurants San Francisco*" by appending the determined location to the original search query. The new search query is then issued to the search engine to obtain the search result for the user. Because the location "*San Francisco*" is now included in the new search query, based on which the search result is identified, the search engine is more likely to find Italian restaurants located in San Francisco. This approach, however, suffers from two sources of errors: (1) the local intent might not be the only intent of the query or even may not exist due to limited precision of the general implicit local intent classifier; (2) the user's location may be determined wrongly.

3.2 Personalization by Re-ranking

As a more conservative approach, document re-ranking is proposed to leverage user's location. We first extract top K documents with the original query and then adjust the search results by increasing the ranks of those documents that match the user location. Consequently, the Web documents that match the user's location are ranked higher than those documents that do not match the user's location. We also differentiate and weight user location matches for different sections in top documents. We re-rank those document based on their current rank score and text matching features that tell if user location and its variation exists in certain document sections. The re-rank score $s^r(q, d, l)$ for query q, Web document d and user location l can be expressed by

$$s^r(q, d, l) = s(q, d) + \sum_{i} w_i I_i(d, l), \quad (5)$$

where $s(q, d)$ is the original rank score before personalization, $I_i(d, l)$ is an indicator function that tells if user location l exist in document section i and w_i is the weighting parameter for document section i. Supervised learning is used to estimate the parameter w_i to maximize the relevance after personalization.

4. EXPERIMENTS

We evaluate our personalization system by both editorial relevance test based on Discounted Cumulative Gain (DCG) and online bucket test for user experience evaluation based on click-through rate (CTR), two commonly used metrics to evaluate search engine relevance and user experience [1].

We apply both query expansion and document re-ranking to the queries that are classified as general implicit local queries. We sample 1300 random personalized queries and submit the top 5 Web search results for each approaches together with test queries and the user location for each query to trained editors for relevance judgment. If a test query does not contain general implicit local intent, the user location information will be ignored in the judgment; otherwise, the user location will be considered in the relevance judgment. The baseline is one of the top commercial search engines. We can see from Table 1 personalized Web search results dramatically improved search relevance. In online test, we also observe that user experience measured by CTR has been improved significantly by 1.4% with $p - value < 0.05$ for affected queries.

Table 1: Relevance Impact with Personalization

Search Engine	DCG	Improv.	p-value
baseline	2.71	-	-
Query Expansion	2.96	9.2%	0.000
Re-rank	3.29	21.4%	0.000

5. CONCLUSION AND FUTURE RESEARCH

We successfully improved relevance the user experience over one of the top commercial search engine by personalizing search results using user's physical location. The current model covers 2.6% search traffic. We plan to increase the coverage substantially by incorporating more salient features and better language models. The current re-ranking algorithm is rather simple and can be substantially improved by jointly considering location matching features and other ranking features together in the phase of re-rank.

6. REFERENCES

[1] B. Carterette and R. Jones. Evaluating search engines by modeling the relationship between relevance and clicks. In *NIPS*, 2007.

[2] A. Micarelli, F. Gasparetti, F. Sciarrone, and S. Gauch. Personalized search on the world wide web. *The Adaptive Web*, 4321 of Lecture Notes in Computer Science:195 – 230, 2007.

[3] X. Yi, H. Raghavan, and C. Leggetter. Discovering Users' Specific Geo Intention in Web Search. In *WWW*, 2009.