Identifying Regional Sensitive Queries in Web Search

Srinivas Vadrevu, Ya Zhang, Belle Tseng, Gordon Sun, Xin Li
Yahoo! Inc.
701 First Ave.
Sunnyvale, CA 94089
{svadrevu, yazhang, belle, gzsun, xinli}@yahoo-inc.com

ABSTRACT

In Web search ranking, the expected results for some queries could vary greatly depending upon location of the user. We name such queries regional sensitive queries. Identifying regional sensitivity of queries is important to meet users' needs. The objective of this work is to identify whether a user expects only regional results for a query. We present three novel features generated from search logs and build a meta query classifier to identify regional sensitive query. Experimental results show that the proposed method achieves high accuracy in identifying regional sensitive queries.

**Categories and Subject Descriptors:** H.3.m [Information Search and Retrieval]: Miscellaneous
General Terms: Algorithms, Experimentation
Keywords: query classification, regional sensitive, search

1. INTRODUCTION

With the fast penetration of the Internet and the World Wide Web throughout the world, the number of search users has increased dramatically from many geographic locations. Search engines are now facing the problem of 'localization' of search results, i.e. displaying regional results at higher rank when the query is *regional sensitive*. A query is considered regional sensitive if local results are expected from the query instead of global results. For example, queries such as 'weather' and 'train tickets' have very strong desire for regional content and hence are classified as regional sensitive queries. For regional sensitive queries, the expected search results depend upon the location of the user. Figure 1 shows an example where the query 'train tickets' is issued from different countries and how varying results are expected for users in each of the country. Thus it is evident that the users expect to see more regional results when the query is regional sensitive. On the other hand, queries like 'funny pictures' and 'garden design' that bear no regional intention are considered global queries. Here the word 're- gional' may be interpreted at different granularity such as country and state. While the technique presented here is in- dependent of the actual meaning of 'regional', we focus our experiments and analysis at the country level in this paper.

Identification of user intent has been an important part of many web applications [1]. In the case of localization of search results, recognizing regional sensitive queries is crit-
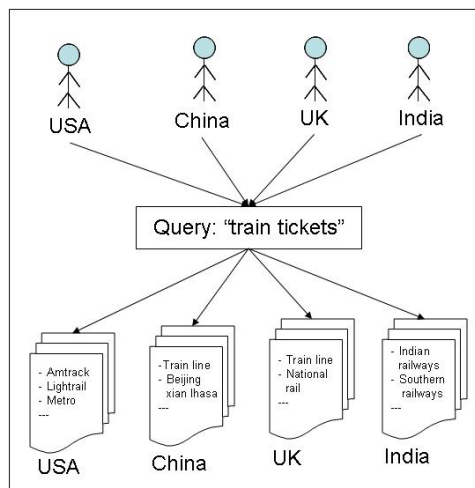
Figure 1: A scenario where a regional sensitive query such as 'train tickets' shows different search results based on the geographical location of the user.

ical for determining whether to provide generic search results or local results. If all queries are treated the same, a generic query like 'funny pictures' may have only local results returned which may be less relevant to what the user is looking for. Currently major commercial search engines offer users the options to restrict search results to a specific region/language combinations through advanced search options. While it partially solves the localization problem, more input is required from the users. One desirable property of a search engine is to automatically recognize users' intention on the fly and return relevant results accordingly.

There is a wealth of information that we can use to identify regional sensitive queries. User behaviors on search engines, e.g. query and click logs, represent valuable resources for identifying user intent [2]. In this paper, we present several approaches to utilize this information to identify regional sensitive queries and describe a meta-classifier to classify queries as regional sensitive or not.

2. IDENTIFICATION OF QUERIES WITH REGIONAL INTENT

We present three features to identify the regional sensitivity of queries from query and click logs of the search engine. We also discuss how to combine these features with a meta-classifier to identify regional sensitive queries.

2.1 Location Likelihood

In this approach, we utilize the co-occurrence of query terms with locations to identify the regional sensitivity of a query. Usually ambiguous regional sensitive queries like 'train tickets' become obvious regional queries when a location name is embedded in them. For example if a user A types 'train tickets in new york', then it is very clear that the user A is looking to buy train tickets in the location of new york. One advantage of the Web is the vast amount of users typing similar queries without the location information, which can be used to identify queries without explicit location information.

Suppose that $\mathcal{D}$ is a dictionary of all location names in a given region and $\mathcal{L}$ is a query log that contains all the query strings that several users typed in a given region. The location likelihood of an n-gram (a contiguous substring with n terms) within a query is a simple ratio of the number of times the n-gram appears with location to the number of times the n-gram appears overall. For a given query Q, the location likelihood of Q can be computed as a weighted sum of the location likelihoods of the n-grams within the query.

$$ll(Q) = \sum_{\forall ngrams, N_i \in Q} len(N_i) * ll(N_i), \; where$$

$$ll(N) = \frac{\# \; times \; N \; occurs \; with \; any \; l \in \mathcal{D} \; in \; \mathcal{L}}{total \; \# \; times \; N \; occurs \; in \; \mathcal{L}}$$

where N_i is an n-gram within the query. Note that the location likelihood of the n-gram within a query is weighted with its length, so that longer n-grams get higher weight in identifying location sensitivity.

2.2 Utilizing User Click Patterns

User click embeds a wealth of information about users' intention. Clicks have been explored in many ways to improve search engine relevance [1]. Based on user click behavior for each query, we construct a set of regional sensitive features. **Local switch rate** (swr) is defined as the user switch rate from all-the-web to country search only for the same query in a query session log. **Regional click rate** (rcr) is defined as the ratio of the number of regional results being clicked over the number of non-regional results being clicked for a given query.

The feature *swr*, extracted by performing session analysis, usually accurately reflected the users regional intent. However, for a given regional intended query, the users only make a switch if the top results returned are not what the users intended for. Hence, this swr feature has a very low coverage. On the other hand, the other feature, *rcr*, utilizes all clicks available to generate the regional sensitive features. While the feature has a coverage of almost all queries in the query log, the accuracy drops slightly.

2.3 Regional Sensitive Meta Query Classifier

The meta classifier for regional sensitivity is obtained by linearly combining the individual features. The weights for the linear combination are learnt by optimizing the area under ROC curve for each of the individual features to learn the optimum thresholds and using these thresholds to combine them. The resulting meta classifier scores are normalized in order to facilitate comparison against the features themselves.

query	ll	swr	rcr	$metaclass$
hotels	0.42	0.08	0.36	0.54
buy	0.23	0.04	0.04	0.29
rent car	0.37	0.12	0.35	0.53
engineering college	0.46	0.04	0.71	0.68
restaurants	0.40	0.07	0.49	0.57
zee news	1.00	0.00	1.00	0.31
mathematics	0.05	0.02	0.04	0.10
bbcnews	0.00	0.00	0.00	0.009
overall precision	0.92	0.87	0.90	0.90
overall recall	0.54	0.29	0.62	0.66
overall f-measure	0.69	0.46	0.74	0.78
accuracy	0.69	0.46	0.78	0.88

Table 1: Regional sensitive feature values and the meta classifier results.

3. EXPERIMENTAL RESULTS

To evaluate the features, we obtained a hand labeled set of 100 queries that were randomly sampled from the query logs of a commercial search engine with millions of query volume. A binary value is assigned to each query to indicate whether it is regional sensitive or not. The criterion to determine whether a given query is regional sensitive is based on whether the user expects to see regional content for this query in the search engine.

Table 1 shows the individual feature values and the meta classifier output for certain queries in the query log. Note that each of the individual features capture certain class of regional sensitive queries, whereas the meta classifier is able to identify all regional sensitive queries with high accuracy. The *ll* feature is able to identify individual n-grams like 'buy' and 'zee news' correctly, whereas the regional click rate is able to capture common queries like 'restaurants' and 'engineering college' effectively.

Table 1 also shows the overall precision, recall and f-measure values for each of the individual features and the meta-classifier. Our results are obtained by labeling a set of 100 randomly sample queries. The precision is set at about 90% for all methods and other metrics have been calculated correspondingly. Among the individual features, regional click rate (*rcr*) performs the best with 78% accuracy. The meta query classifier, however is able to beat all the individual features and achieves highest accuracy.

4. CONCLUSIONS AND FUTURE WORK

We presented three individual features and a meta query classifier to identify the regional sensitivity of queries. We capture user behaviors in a commercial search engine from query and click logs to compute the regional sensitivity of queries. Our experimental results indicate that our methods can scale up to millions of query volume and can achieve high precision in identifying regional sensitive queries. The regional sensitivity of queries can be used as a rank feature in Web search engine and sponsored search.

5. REFERENCES

[1] H. Dai, L. Zhao, Z. Nie, J. Wen, L. Wang, and Y. Li. Detecting online commercial intention (oci). In *Proceedings of the 15th International Conference on World Wide Web*, 2006.

[2] B. J. Jansen, D. L. Booth, and A. Spink. Determining the user intent of web search engine queries. In *Proceedings of the 16th international conference on World Wide Web*, 2007.