Modeling and Visualizing Geo-Sensitive Queries Based on User Clicks

Ziming Zhuang
Pennsylvania State University
University Park, PA, USA
zzhuang@ist.psu.edu

Cliff Brunk
Yahoo! Applied Research
Santa Clara, CA, USA
brunk@yahoo-inc.com

C. Lee Giles
Pennsylvania State University
University Park, PA, USA
giles@ist.psu.edu

ABSTRACT

The number of search queries that are associated with geographical locations, either explicitly or implicitly, has been quadrupled in recent years. For such *geo-sensitive* queries, the ability to accurately infer users' geographical preference greatly enhances their search experience. By mining past user clicks and constructing a geographical click probability distribution model, we address two important issues in spatial Web search: how do we determine whether a search query is geo-sensitive, and how do we detect, disambiguate, and visualize the associated geographical location(s). We present our empirical study on a large-scale dataset with about 9,000 unique queries randomly drawn from the logs of a popular commercial search engine *Yahoo! Search*, and about 430 million user clicks on 1.6M unique Web pages over an eight-month period. Our classification method achieved recall of 0.98 and precision of 0.75 in identifying geo-sensitive search queries. We also present our preliminary findings in using geographical click probability distributions to cluster search results for queries with geographical ambiguities.

Categories and Subject Descriptors

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

General Terms

Algorithms, Experimentation

Keywords

geographic information retrieval, geo-sensitivity, geographical query, classification

1. INTRODUCTION

It is estimated that 12% to 16% of the Web search queries are *local* queries [8]: queries that can be associated with a geographical locality. With the recent proliferation of such *local* queries, most commercial search engines offer the capability to conduct *Local Search* [1, 2, 3], which usually requires the user to specify a location qualifier (e.g. *"San Francisco"*) in addition to the search query (e.g. *"car dealers"*). For queries without a clearly specified location qualifier (e.g. *"pizza delivery stanford"*), search engines need to accurately infer the user's geographical preference (*"stanford"* in the example).

In this paper we tackle a more challenging scenario: when users submit queries with spatial ambiguities. For example, a query *"bay bridge"* may refer to the San Francisco-Oakland Bay Bridge in California, the Chesapeake Bay Bridge in Maryland, or the Escambia Bay Bridge in Florida. Similarly, a query *"chicago pizza"* may refer to *chicago style pizza* or *pizza stores in Chicago*. From these and many other ambiguous queries, it is obvious that syntactic information derived from the query terms alone is not sufficient to infer the user's geographical preference.

We propose to exploit the geographical probability distributions of user clicks as an additional perspective for locality inference. Specifically, we address the following two problems:

- Given a search query, how do we algorithmically determine whether it is geo-sensitive, and derive its geographical locality from past user clicks?
- Given a query with geographical ambiguities, how do we disambiguate and cluster the search results based on the locality information inferred from user clicks?

Using textual syntactics to infer locality information has been studied in recent years. A number of topics have been investigated, including the characteristics of the search terms in geographical queries [9, 10], geographical query expansion and rewriting [6, 7, 14], and geotagging web resources [4, 5, 11, 13]. More relevant to this study are [8, 12]. However, our proposals are significantly different in the methodologies and the formulation of the solution.

The rest of the paper is organized as follows. We first briefly introduce the datasets in Section 2. We then propose the definition of *query geo-sensitivity*, examine and visualize the geographical patterns of user clicks, and present the algorithm for identifying geo-sensitive queries in Section 3. In Section 4 we discuss the classification results from an empirical evaluation. We also present in Section 5 some preliminary findings of using geographical click distributions to disambiguate and cluster search results. We conclude our paper with directions of future work in Section 6.

LocWeb 2008, April 22, 2008, Beijing, China.
ACM 978-1-60558-160-6/08/04 . . . $5.00.

Table 1: Definitions of Query Geo-Sensitivity

Sensitivity Category	Definition	Examples
Explicit	Queries about local business, organizations, product, and information of a particular location, which explicitly specify a geographical location.	"georgia department of revenue", "nissan dealer in bay area", "Seattle tax rate"
Implicit	Queries that implicitly specify a location with a popular or famous local business or landmark.	"dulles airport", "broward community college", "aspen grove shopping center", "dumbarton bridge"
Local	Queries that are not specific to a particular location but local information is implicitly preferred; typically contain the name or type of a business, service, or organization without a specific geographical location.	"dentist", "toyota dealer", "AAA branch"
Non-Sensitive	Generic or navigational queries, and queries about information interesting to users regardless of their physical locations.	"yahoo", "google", "disneyland", "new york times"

2. DATASETS

We were granted access to the **anonymized** logs of a popular commercial search engine *Yahoo! Search*. From the search query log, we sampled a set of 8.9K queries which were believed to be representative and covering a broad range of search topics. We then aggregated about 430M user clicks on these queries from the click logs of Feb. – Sept. 2007, which contained about 1.6M unique (query, URL) pairs. Each record consisted of the following fields: query, URL, rank of the URL in the search results, IP address, and the aggregated number of clicks which originated from this address. An example is (*"american funds"*, *"http://adviser.americanfunds.com"*, 29, 4aedd5d1, 33) Note that we collapse multiple clicks from the same search session into one.

After filtering the IP addresses of popular proxy servers and large ISP hubs, we used a proprietary IP lookup database to map each IP address to a geographical location (i.e., latitude, longitude, city/town, county, state, zipcode, and country). In the previous example *"american funds"*, the geographical location is (35.152, -90.035, Memphis, Shelby, TN, 38105, USA). we denote this dataset by $\mathcal{C}$ in the rest of the paper.

3. MODELING AND CLASSIFYING GEO-SENSITIVITY

We first present our definition of *query geo-sensitivity*, then visualize and observe the characteristics of the spatial distribution of user clicks for the geo-sensitive queries, and discuss how we algorithmically derive the geo-sensitivity for queries under a classification scheme.

3.1 Geo-Sensitivity Definition

The geo-sensitivity of a query denotes that, to answer the query, webpages that either have explicit / implicit association with certain geographical location(s) or are considered more relevant to users in certain geographical location(s) will be considered more *relevant*. To make it more concrete, we define four categories of geo-sensitivity for queries in Table 1.

For the rest of this paper, we will follow these definitions. If a search query falls into either the *explicit* or the *implicit* sensitivity category, it is referred to as a *Geo-Sensitive Query* (*GSQ*), otherwise a *Non-Geo-Sensitive Query* (*NGSQ*).

3.2 Click Distributions and Geo-sensitivity

In Section 1, we already discussed the deficiency of using only the syntactics of a query to infer its geo-sensitivity. Assume that a user submits a query and the search engine returns a list of webpages, he/she is more likely to click on webpages that appear *more relevant*. For *GSQs*, whether a webpage is *geographically* relevant is an important factor of the overall relevance perceived by the user. Thus, given a search query, once we aggregate a significant number of user clicks and conduct a reverse-lookup of the IP addresses to find out where they come from, it is possible to learn about the locality preference of this query by *"following the herd"*. For example, a *GSQ "broward community college"* received the majority of user clicks originated from the Fort Lauderdale, FL area, which could indicate that this query is *more geographically sensitive* to the Fort Lauderdale area than anywhere else.

Hypothesis 1. *The geographical distributions of user clicks for* NGSQs *and* GSQs *have different patterns: the aggregated click distribution of* NGSQs *resembles the user population distribution, while the distribution of* GSQs *differs from the population distribution.* ■

To examine the above hypothesis, we visualize the aggregated click records in $\mathcal{C}$ using *click maps*. Here, the *click map* of a query visualizes the geographical locations of search users who have issued the query and also clicked on some of the results. Each red dot on the click map of a query represents the aggregated user clicks originated from the corresponding geographical location.

Figure 1(a) shows a click map of all 8.9K search queries in $\mathcal{C}$, in which the geographical distribution of user clicks mostly follows the population distribution in the U.S. Compared with Figure 1(b) which shows a click map of the query *"Google"*, it is obvious that the two resemble each other, indicating that the query *"Google"* is not *geographically sensitive* to a particular location.

In contrast, Figure 1(c) and 1(d) show the click maps of two *geo-sensitive* queries *"Sun Country Airlines"* and *"93.7 Houston"*, respectively. The spatial distributions of the user clicks for these two queries both display a strong deviation from the population distribution, and have unusually high density in some regions: *"Sun Country Airlines"* received more than 30% of its clicks coming from Minneapolis, MN (a hub of the airline), and *"93.7 Houston"* – a local radio-station in Houston – received more than 80% of its clicks from the Houston, TX area.

As we examined the click maps of more search queries, it was consistent that the geographical distribution of user clicks were usually quite different between *GSQs* and *NGSQs*.

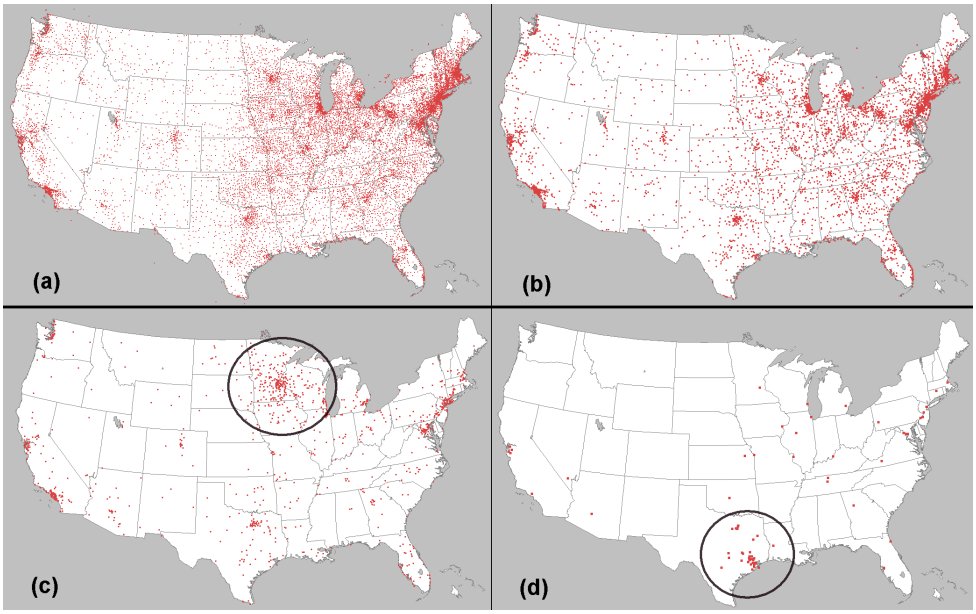

Figure 1: (a) The overall click map of the 8.9K search queries in $\mathcal{C}$. (b) The click map of query *"Google"*. (c) The click map of query *"Sun Country Airlines"*. (d) The click map of query *"93.7 Houston"*. Regions with unusually high density of clicks for the last two queries are indicated by the black circle.

3.3 Geo-Sensitivity Classification

We see the distinctive patterns of geographical click distributions of *GSQs*. Accepting Hypothesis 1, can we differentiate *GSQs* from *NGSQs*, and infer the locality preference of a query based on its geographical click distribution?

Without loss of generality, the task to identify the *geo-sensitivity* of a search query can be cast into a binary classification problem, formally defined as follows:

Definition 1 *Given a set of queries $\mathcal{Q} = \{q\}$, classify q into one of the two classes: (1) a class of* geo-sensitive queries *GSQ ($GSQ \subseteq \mathcal{Q}$), and (2) a class of* non-geo-sensitive queries *$NGSQ$ ($NGSQ = \mathcal{Q} - GSQ$).* □

Let $\mathbf{q} = \langle d_1, d_2, \cdots, d_k \rangle$ denote a query q. $\mathbf{q}$ is a k-dimension vector, and each dimension d_i represents *the probability of receiving user clicks from a geographical region r_i*. The most appropriate granularity of r_i (i.e. how large / small is the size of the region) can be iteratively learned or chosen arbitrarily.

Furthermore, let vector $\mathbf{p} = \langle d_1^p, d_2^p, \cdots, d_k^p \rangle$ denote the user population, with each dimension d_i^p represents the population in a geographical region r_i same as in $\mathbf{q}$. By accepting Hypothesis 1, we chose to approximate the population distribution with the aggregated click probability distribution of all the queries in $NGSQ$, i.e., the vector summation of all vectors $\mathbf{q}' \in NGSQ$ and normalized:

$$\mathbf{p} = \frac{\sum \mathbf{q}'}{|NGSQ|}, \forall \mathbf{q}' \in NGSQ \quad (1)$$

.

Given a query q, we can then measure the distance $\mathcal{D}$ between its geographical click probability distribution $\mathbf{q}$ and the user population distribution $\mathbf{p}$ approximated by Equation 1:

$$\mathcal{D} = \mathbf{dist}(\mathbf{p}, \mathbf{q}) \quad (2)$$

where **dist** can be an appropriate standard distance measure of choice, e.g. *chi-squared* χ^2, *KL Divergence*. And based on $\mathcal{D}$ we can decide to classify q as either GSQ or $NGSQ$.

4. EVALUATION AND DISCUSSIONS

To establish the ground-truth for evaluating the classification accuracy, a test-set of 1,000 queries were randomly sampled from the search logs, and each query in the test-set was reviewed by a team of professional editors and classified according to the definitions outlined in Table 1.

For this paper, we chose to limit the scale of experiments to geographical locations within the conterminous United States. Based on the classification performance in pilot tests comparing a number of various granularities, we fixed the region granularity as a State. Thus, $\mathbf{q}$ is a 48-dimension vector, and each dimension represents the probability of user clicks from the corresponding State. Same for $\mathbf{p}$. We also chose *chi-squared* as the distance metric.

We first calculated $\mathbf{p}$ on the aforementioned 8.9K queries training dataset $\mathcal{C}$. Then, we applied our proposed classification method on the 1K queries test-set. Evaluated on the ground-truth editorial classifications, our proposal obtained quite promising accuracy:

$$\textbf{Recall} = 0.983$$

$$\textbf{Precision} = 0.753$$

The classifier correctly labeled more than 98% of the *GSQs*, and had only about 25% false positive rate. Table 4 shows the top 10 queries that were classified as either *GSQ* or *NGSQ*, order by the distance $\mathcal{D}$. It is worth noting that among the queries labeled as *GSQ*, *"kcci"*, the acronym of a local TV station in Des Moines, IA, would not have been easily recognized as *geographically sensitive* had we relied only on the syntactic information of the query alone.

Class Label	Queries
GSQ	alaska dmv, rapid city journal, metro community college, kcci, utah power, state of maine, fosters daily democrat, idaholottery, alcorn state university, salt lake tribune newspaper
NGSQ	yahoo, ups, food network, adobe, hotels, itunes, google search, dell, microsoft, nero

Table 2: Classification results: Top 10 queries classified as either *GSQ* or *NGSQ*, in descending order of the distance reported.

5. DISAMBIGUATING AND CLUSTERING SEARCH RESULTS

We briefly discuss some preliminary findings in using geographical click distribution to cluster and disambiguate search results for ambiguous queries. We report our experiments on a query *"washington jobs"*, which is ambiguous in terms of the user's geographical preference: is it about jobs in the Washington state or Washington DC?

Our goal was to cluster the search results such that webpages in the same cluster were relevant w.r.t. the same geographical region. First, we aggregated past user clicks per URL, and modeled each click as a vertex on a two-dimension Euclidean space, denoted by (*latitude*, *longitude*) in the geographical space from where this click came from. Then we constructed an N nearest-neighbor graph G and repeatedly split G into K-clusters using the *Min-Max Cut* partitioning algorithm. After the vertexes (clicks) were all clustered, we clustered a given webpage based on its likelihood of receiving clicks in each of the click clusters. Eventually, each webpage in the search results was represented by a probability distribution over the K clusters.

Figure 2 visualizes the result of the document clustering process. Each dot represents a webpage returned by the search engine, and different colors denote different clusters. Webpages relevant to different geographical regions are mostly separated, and those that are relevant to the same region are grouped together into the same cluster.

6. CONCLUSION AND FUTURE WORK

We proposed to use the geographical probability distribution of user clicks as additional heuristics to accurately identify search queries that are *geo-sensitive*. The geographical preference of the user could then be modeled and detected. Our proposal demonstrated its effectiveness in empirical evaluations, showing promising recall and precision metrics in terms of query classification accuracy. We also presented briefly our findings in using the geographical click distributions for search results clustering.

In the current vector-based discrete representation of the geographical space, there is no consideration for proximity among dimensions due to arbitrarily-defined boundaries. For example, the distances between California and Texas and between California and Virginia are considered equal. We are currently exploiting continuous representations of the geographical space which would take into account such proximity metrics. We also plan to further study the impact of different geographical granularity choices.

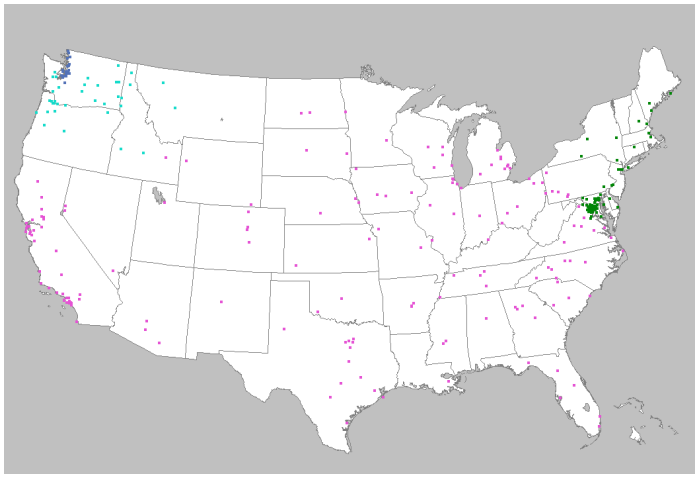

Figure 2: Visualization of search results clustering for ambiguous query *"washington jobs"*. Dots in different colors denote different clusters, e.g. the green dots denote webpages relevant to Washington DC, and the blue dots relevant to the State of Washington.

7. REFERENCES

[1] Google local search. http://maps.google.com/.

[2] Microsoft live search local. http://maps.live.com/localsearch/.

[3] Yahoo local search. http://local.yahoo.com/.

[4] E. Amitay, N. Har'El, R. Sivan, and A. Soffer. Web-a-where: geotagging web content. In *Proc. of the 27th ACM SIGIR conference on Research and development in information retrieval*, pages 273–280, 2004.

[5] K. A. V. Borges, A. H. F. Laender, C. B. Medeiros, and J. Clodoveu A. Davis. Discovering geographic locations in web pages using urban addresses. In *Proc. of the 4th ACM workshop on Geographical information retrieval*, pages 31–36, 2007.

[6] N. Cardoso and M. J. Silva. Query expansion through geographical feature types. In *Proc. of the 4th ACM workshop on Geographical information retrieval*, pages 55–60, 2007.

[7] G. Fu, C. B. Jones, and A. B. Abdelmoty. Ontology-based spatial query expansion in information retrieval. In *Lecture Notes in Computer Science*, pages 1466–1482, 2005.

[8] L. Gravano, V. Hatzivassiloglou, and R. Lichtenstein. Categorizing web queries according to geographical locality. In *Proc. of the twelfth international conference on Information and knowledge management*, pages 325–333, 2003.

[9] M. Sanderson and Y. Han. Search words and geography. In *Proc. of the 4th ACM workshop on Geographical information retrieval*, pages 13–14, 2007.

[10] M. Sanderson and J. Kohler. Analyzing geographic queries, 2004.

[11] C. Wang, X. Xie, L. Wang, Y. Lu, and W.-Y. Ma. Web resource geographic location classification and detection. In *Proc. of the 14th International Conference on World Wide Web*, pages 1138–1139, 2005.

[12] L. Wang, C. Wang, X. Xie, J. Forman, Y. Lu, W.-Y. Ma, and Y. Li. Detecting dominant locations from search queries. In *Proc. of the 28th ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 424–431, 2005.

[13] Q. Zhang, X. Xie, L. Wang, L. Yue, and W.-Y. Ma. Detecting geographical serving area of web resources. In *Proc. of the 3th ACM workshop on Geographical information retrieval*, 2006.

[14] V. W. Zhang, B. Rey, E. Stipp, and R. Jones. Geomodification in query rewriting. In *Proc. of the 3th ACM workshop on Geographical information retrieval*, 2006.