Here and There: Goals, Activities, and Predictions about Location from Geotagged Queries

Robert West*
Stanford University
Stanford, California
west@cs.stanford.edu

Ryen W. White
Microsoft Research
Redmond, Washington
ryenw@microsoft.com

Eric Horvitz
Microsoft Research
Redmond, Washington
horvitz@microsoft.com

ABSTRACT

A significant portion of Web search is performed in mobile settings. We explore the links between users' queries on mobile devices and their locations and movement, with a focus on interpreting queries about addresses. We find that users tend to have a primary location, likely corresponding to home or workplace, and that a user's location relative to this primary location systematically influences the patterns of address searches. We apply our findings to construct a statistical model that can predict with high accuracy whether a user will be soon observed at an address that had been recently retrieved via search. Such an ability to predict that a user will transition to a location can be harnessed for multiple uses including provision of directions and traffic information, the rendering of competitive advertising, and guiding the opportunistic completion of pending tasks that can be accomplished en route to a target location.

Categories and Subject Descriptors: H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*search process, selection process*.
Keywords: mobile search, user mobility, log analysis.

1. INTRODUCTION AND BACKGROUND

Web search on mobile devices, such as smartphones, is becoming increasingly popular.[1] Research has demonstrated that search on mobile devices differs significantly from search in desktop settings. For example, Teevan et al. [7] show that queries on such devices are frequently issued while people are in transit and are often related to in-world destinations such as businesses. A user's current location influences search, as information goals may be driven by surrounding entities and resources. Conversely, search can provide information about a user's forthcoming locomotion and engagement with physical surroundings.

Geolocation information captured by GPS receivers within smartphones can be used to build predictive models that link online interactions to forthcoming transitions among locations in the physical world. Related prior work studied the relationship between queries about concerning health symptoms and about proximal medical facilities [9] and has demonstrated with mobile data when intentions expressed in search queries may translate into healthcare utilization, with evidence of near-term travel to medical facilities [10]. However, mobile search is still poorly understood. Understanding the interplay of queries and the physical world promises to help with improving the quality of mobile search, e.g., by considering geography, proximity, and predictions about future locations as critical aspects of the context of searches [2]. As a concrete example, consider a user who is querying for MEGAMART. If we can infer from the user's recent queries and her recent locations that she intends to go to the closest Megamart store, we can rank driving directions highest, whereas otherwise the Megamart website might be more appropriate. A service could also provide traffic updates and status information about the target location, such as the density of crowds, lengths of lines, and specials at Megamart. Knowledge that a user is heading to a location can also be used in competitive advertising campaigns—aimed at persuading the searcher to shift destinations to an alternate business. In another scenario, an opportunistic planning application tied to search could provide guidance on visiting locations and addressing tasks on a pending to-do list (e.g., deposit checks at a bank located next to Megamart) on the way to the location [5].

We present a study of user location and locomotion as captured in geotagged queries logged from a large set of consenting mobile users. We make two main contributions. First, we show that people tend to have a primary or dominant location where they use their device (perhaps home or workplace), and that the search behavior of users changes when they are not at the primary location and also as a function of distance from that location. For example, we found that address queries become more common as users move further away from their primary location, suggesting that address queries are used more in less familiar settings. Address queries (e.g., 1 UN PLAZA, NYC) are particularly interesting, as they have an obvious connection to the physical world. Second, we characterize empirically how address queries are used, and we construct and evaluate a statistical model capable of predicting if an address query will be followed by the user being observed (via their geotagged queries) visiting the searched location. Finally, we discuss, in the context of prior work, the potential applications enabled by these results.

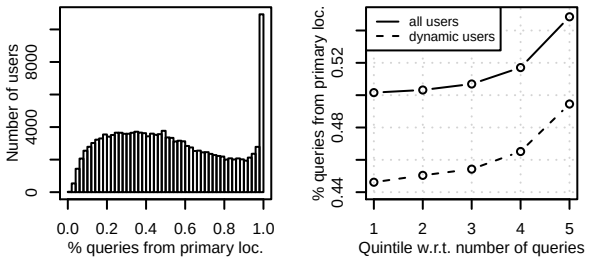
2. DATA

We collected datasets of location and search activities of users with consent via logs of a major mobile search provider. The data was provided via a widely available mobile search and navigation application installed on the iPhone and Android platforms. Users agreed to share their geotagged queries with the host service of the application in accordance with a published privacy statement. The log includes about a year of data logged from approximately 140K consenting users and consists of millions of search entries stored on

*Research done during an internship at Microsoft Research.

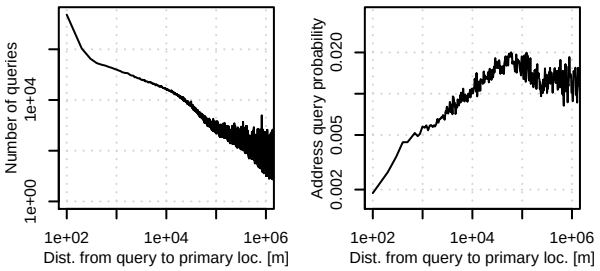
[1] http://www.comscore.com/Insights/Presentations_and_Whitepapers/2011/Mobile_Search_Techniques_and_Tactics_for_Marketers

SIGIR'13, July 28–August 1, 2013, Dublin, Ireland.



(a) (b)

Figure 1: (a) Histogram of percentage of queries from primary location. (b) Percentage of queries from primary location as a function of users' overall numbers of queries; 'dynamic' users are those with less than 95% of queries from primary location (standard error bars smaller than dot size).



(a) Number of queries (b) Address query probability

Figure 2: (a) Queries are more frequent closer to the primary location. (b) Probability of query being an address query increases with user's distance from primary location.

a secure server. Each entry corresponds to a search interaction with the mobile application issued by an anonymized user (represented by a unique numerical identifier), which contains the search query, time, and GPS coordinates (i.e., when and where the query was submitted). Location information was recorded only if and when a search query was issued.

3. PRIMARY LOCATIONS

We found that people have regular patterns of mobility, characterized by a small number of spatial anchors at several primary locations that are regularly visited—potentially representing such recurrent locations as home, work, gyms, restaurants, clubs, and bars. Even though users' lives may revolve around several frequent locations, we simplify our analysis by considering as a reference point in our analyses the location where most querying is performed.

We divide the Earth's surface into discrete cells and define the cell from which a user queries most as their *primary location.* We map GPS latitude/longitude coordinates to discrete map cells by rounding the GPS coordinates to three decimal places. Due to the Earth's spherical shape, the resulting cells vary in size across the planet; for reference, cells are 84 meters by 110 meters in New York City. We consider in our analyses only those users who have issued at least 100 queries during the course of the logging, such that reliable primary locations could be computed.

Referring to Fig. 1(a), we observe two classes of users: (1) those who query only from their primary location, and (2) those who also query from other locations. The latter group vary by the portion of queries made from the primary location. The distribution of this percentage is very broad, with most users spending a large percentage of their time at their primary location.

Fig. 1(b) shows the relationship between users' overall numbers of queries and the fraction of queries made from the primary location. We see that, the higher a user's absolute number of queries, the higher typically the fraction of queries from the primary location. There may be several reasons. One possibility is that heavy users may employ their mobile devices as substitutes for desktop computers at home. Another explanation could be that users have only limited information needs or cognitive capacity while mobile, whereas searching at home or at work is less constrained.

Per our definition, most queries are made from users' primary locations. However, it is unclear what happens as users stray ever further from their respective primary locations. As a thought experiment, if each user visited everywhere on Earth with uniform probability, we would expect the number of queries made at distance d from the primary location to increase in d, as the number of places at distances d grows with d. However, the uniform assumption is unrealistic, and Fig. 2(a) shows that, rather than increasing, we observe a smooth decay in the number of queries as d increases. At least two factors may play a role in this diminishment: (1) Users are more likely to stay in places close to than far away from their primary location. (2) Even when conditioned and renormalized for the likelihood of a user being at a given location, the probability of querying there might be smaller the further away that location is from the primary location. Unfortunately, as we only observe locations at query time, we have no ground truth about actual user motion, so we cannot characterize the influence of these factors.

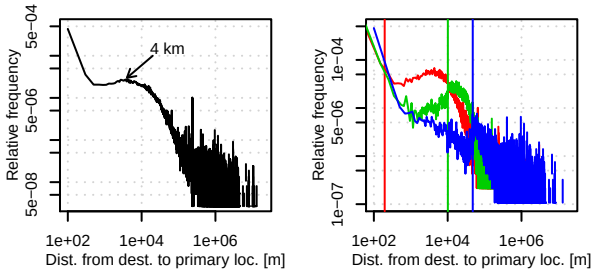
4. ADDRESS QUERIES

In the last section, we showed that query volume decreases as users travel further from their primary locations. We observe in the data that, in addition to quantity, the qualitative nature of queries also varies spatially. We classified the queries into 48 categories of query intent, using a proprietary classifier, where queries could be classified into multiple categories (e.g., WEATHER, HOWTO, MAPS). For example, we found that the fraction of queries tagged as adult content is higher at the primary location, while queries for street addresses, among others, are more likely at other locations. Fig. 2(b) shows that the probability of a query being an address query rises steadily as a function of the distance between the location where the query is issued and the primary location. We consider address queries particularly interesting as they may reveal an intention to move in the physical world, in contrast to most other queries, which often signal an intention to navigate among locations on the Web. We do not know from the logs whether address queries are input to access information about entities at locations or as a means of accessing mapping and routing services.

4.1 Characterization

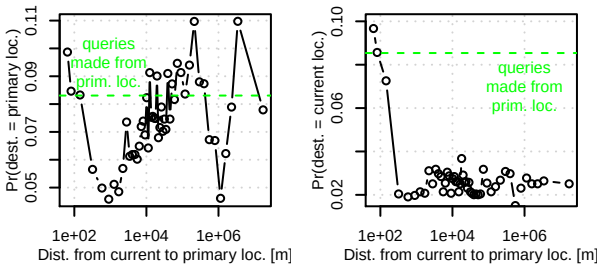
We now characterize the nature of address queries more fully. We define as an address query those queries which receive a high score for being an address by the query classifier (e.g., 47 MAIN ST, MUNICH, ND 58352) and that could also be resolved to GPS coordinates by a proprietary mapping API. Each address query provides us with three locations (cf. Fig. 5(a)): the user's primary location (P), the location at query time (Q), and the location of the destination address (D). Several ways of entering address queries are conceivable: e.g., users might enter the address manually, or they might copy–paste it from a Web page, email, etc. Our log data does not specify details about how users input queries.

Fig. 2(b) shows that queries are more likely to be address queries as the distance to the primary location increases, up to around 100 km. The increase is intuitive: the further you are from the place you stay at most, the less likely you are to be familiar with your surroundings, and the more likely you will need to travel to unfamiliar locations from a current location. The address query proba-



(a) All address queries (b) Stratified by QP

Figure 3: Histograms of distance DP from primary location to destination, (a) including all address queries, (b) stratified by distance QP from current to primary location; vertical lines mark lower bounds d on QP (cf. Section 4.1).



(a) (b)

Figure 4: (a) Probability of a destination being very close to the primary location (defined as $DP < 100$ m) as a function of the current distance from the primary location. (b) Probability of a destination being very close to the current location (defined as $QD < 100$ m). Queries made from the primary location are not considered in the black curves.

bility peaks around 100 km, then drops and remains constant. Interestingly, previous work [3, Fig. 2(b)] has found that the distribution of distances between homes resembles the inverse of Fig. 2(b). This might indicate that users are more likely to query for addresses when they are outside of residential neighborhoods.

Next, consider Fig. 3(a), which shows a histogram of DP, the distance from the primary location to the queried location. The finding that small values are most probable implies that users are most likely to query for locations in the immediate vicinity of their primary locations. Interestingly, there is a local maximum at around 4 km from the primary location. For distances larger than this we observe ever fewer address queries. Interpreting these findings is complicated by the fact that Fig. 3(a) is based on all address queries, regardless of where they are made. To further investigate, we group address queries by their distance QP from the user's primary location. In Fig. 3(b), we repeat the histogram of Fig. 3(a), but now only for address queries made at a distance of at least d from the user's primary location ($QP \geq d$), for different values of d. Qualitatively, these stratified histograms are similar to the overall histogram of Fig. 3(a). In particular, destinations very close to the primary location are still most likely, regardless of how far away a query is made from the primary location. However, the 4 km hump shifts with d: now the local maximum lies at around $d+4$ km.

Ignoring the uptick on the far left of Fig. 3(b), the distributions roughly resemble log normals (downward-sloped parabolas on log axes), i.e., they are long-tailed, with traveled distances normally being short but sometimes extremely long. The modes of $d+4$ km might result from 4 km being a conveniently traveled distance, taking just under an hour by foot or 5 min by car. We may then interpret the curves of Fig. 3 as the overlay of two dominant motifs on queried locations. We find that they are: (1) close to the user's primary location (the far-left uptick in the figures), and (2) that they are close to the current location (the long-tailed part to the right).

To conclude our characterization of address queries, we seek to understand how prominent these two motifs are as a function of distance from the primary location. Fig. 4(a) pertains to motif 1 and plots the probability that a queried address is essentially the user's primary location (which we define as lying within a 100-m radius from there), as a function of how far away the user is from the primary location. This probability is highest when the user queries from a location proximal to the primary location, reaches a minimum at around 1 km from there, and then increases with distance. The final increase may be expected as the farther away one is from one's 'home turf' the more likely help would be needed to find one's way home. Perhaps more surprising is the finding that users often query for their primary locations while they are at the location. This may be caused by users' exploring their neighborhood, rather than looking up a predetermined destination. Motif 2 is explored further in Fig. 4(b). The x-axis is again the distance between the user's current and primary locations; the y-axis shows the probability that the destination lies within 100 m from the current location. This probability is highest very close to the primary location (due to the same queries causing the high values on the far left of Fig. 4(a)) and roughly constant everywhere else.

4.2 Prediction

We now apply the characterizations from above to construct a statistical model with the ability to predict whether a user will follow up on an address query by going to the queried destination. Such a classifier would allow us to infer near-term intentions to *utilize* a real-world resource rather than access information about it (cf. Sec. 5). As we can only sense user locations when queries are issued, we cannot identify each instance in which the user moves to the destination. We can only note the subset of instances in which the user moves to the destination and then issues a query from there.

Setup. We consider for predictions address queries that are issued at least 100 m away from the destination. This ensures that the user can really move towards the destination, rather than query for it while already there, as we showed was fairly common in the previous section. Next, we divide these queries into two groups of address queries: (1) Queries for which we see another query from the same user on the same day, issued from within 50 m from the destination address. A radius, such as the 50 m we chose, is necessary because the exact destination is a single point, with a vanishing probability of observing a query from exactly there. (2) Queries for which we see no such query from the same user. This distinction defines the positive and negative classes for our binary classification task. The resulting dataset contains around 60K address queries and is highly unbalanced, with only 4.2K (7%) positive examples. We compose features based on insights gleaned in the characterizations above, and build a binary classifier with multiple additive regression trees [4] to predict follow-up visitation.

Features. We now describe the features used in the classifier. We use A to denote the user issuing the query.

(U) User features

- Overall number of queries by A
- Percentage of queries A makes from their primary location
- Latitude/longitude standard deviations over all of A's queries
- Parameters of a log-normal fit to A's inter-query times (log-normals match inter-query times well empirically; these parameters capture the burstiness of A's querying behavior)

(Q) Query features (including features of the destination)

- Search type (regular Web search, image search, maps search, driving/walking directions, news search, etc.)
- Time of day and day of week
- Distances *QD*, *QP*, and *DP* (cf. Fig. 5(a))
- Is query issued from A's primary location?
- Is A's primary location the destination?
- Number of Yellow-Page entities at destination (if any)
- Prior probability (based on log-normal fit to A's inter-query times) of seeing any query by A before end of the day

(C) Context features (the first seven features are present for each of the three queries preceding the input address query)

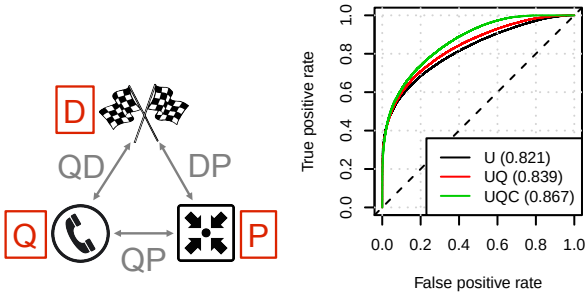
- Search type of previous query
- Edit distance between previous and address query
- Jaccard distance between the previous and the address query
- Is previous query a substring of the Yellow-Page categories of the businesses at the destination?
- Speed estimate based on the previous and the current location
- Is A currently confirmed mobile? (Definition: previous query is less than 10 min ago, the distance is at least 10 m, and the average speed has been at least 4 km/h)
- Distance to destination reduced (absolute and relative) between the previous and the current queries
- Number of queries within the last $\{1,2,4,8,16,32,64\}$ min
- Number of address queries within the last $\{1,2,4,\dots\}$ min

Results. We evaluate our method via 10-fold cross validation. To avoid overfitting, we include all queries by the same user in the same cross-validation fold. As the dataset is unbalanced, trivially predicting a 'no' on every input would result in an accuracy of 93%. To better assess the classifier's power, we create a balanced dataset by subsampling negative examples randomly without replacement. We assess prediction quality in terms of area under the ROC curve (AUC). A random classification achieves an AUC of 0.5.

We evaluate different combinations of the feature groups. The most basic classifier uses no information about the input query but only characteristics of the user's overall behavior (i.e., the U features). Then, we incrementally add features from the two remaining groups (Q and C). The observed boost in AUC provides insights on the utility of the additional features. The resulting ROC curves and AUC values are displayed in Fig. 5(b). We note that user features alone perform significantly better than random (AUC 0.821). Query and especially context features further improve performance, resulting in a maximum AUC of 0.867 when using all features. The steep initial increase of all ROC curves implies that the top-ranked predictions are particularly reliable. Also, the corresponding examples appear particularly obvious to predict, as even the classifier using only user features gets them right. The additional feature groups are most effective farther down in the ranking, which follows from the ROC curves diverging more when true positive rate (i.e., recall) is high. We also evaluated our classifiers on the unbalanced dataset, obtaining similar ROC curves.

5. DISCUSSION AND CONCLUSION

We described a study of search queries issued by people on the move, focusing on queries about addresses. We are not the first to use geotagged log data; cf. [1] for a recent survey. Venetis et al. [8] use historical address queries for ranking potential places of interest in the vicinity of the user's current location. However, our characterizations and analysis of address queries are distinct from prior research. For example, we assume a single place of interest (the queried location) as given and try to predict whether the user will transition to that place.



(a) (b)

Figure 5: (a) Three locations associated with each address query: the user's current location Q, user's primary location P, and destination location D. (b) ROC curves using different feature sets, averaged over the 10 cross-validation folds (AUC in parentheses). For feature-class abbreviations, cf. Sec. 4.2. AUCs differ significantly at $p < 10^{-6}$ for all feature set pairs.

Short- and long-term patterns of human movement are investigated by Sadilek and Krumm [6] and Cho et al. [3]. However, in these efforts, future locations are predicted based only on the user's previous locations, independent of search queries. We anticipate that combining our query-centric approach with ideas from these works could further increase accuracy by leveraging information present in historical long-term patterns of users' movements.

Given the potential sensitivity of location data, applications need to be aware of users' privacy preferences and to seek their consent for different uses of location. Beyond sharing location information with service providers, privacy-sensitive applications can feasibly provide such operations as location-sensitive ranking and filtering of results while limiting access of location data to local devices.

We see rich possibilities ahead for harnessing inferences about users' interests and forthcoming engagements with in-world resources. Inferences about near- and longer-term Web-to-world transitions can be leveraged to provide a range of services in search and advertising, including provision of information about directions and traffic information, wait times, special offers, and tasks that might be accomplished along the way.

6. REFERENCES

[1] R. Becker, R. Cáceres, K. Hanson, S. Isaacman, J. M. Loh, M. Martonosi, J. Rowland, S. Urbanek, A. Varshavsky, and C. Volinsky. Human mobility characterization from cellular network data. *CACM*, 56(1), 2013.

[2] P. Bennett, F. Radlinski, R. White, and E. Yilmaz. Inferring and using location metadata to personalize Web search. In *SIGIR*, 2011.

[3] E. Cho, S. Myers, and J. Leskovec. Friendship and mobility: User movement in location-based social networks. In *KDD*, 2011.

[4] J. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: A statistical view of boosting. *Annals of Statistics*, 28(2), 2000.

[5] E. Horvitz and J. Krumm. Some help on the way: Opportunistic routing under uncertainty. In *UbiComp*, 2012.

[6] A. Sadilek and J. Krumm. Far out: Predicting long-term human mobility. In *AAAI*, 2012.

[7] J. Teevan, A. Karlson, S. Amini, A. Brush, and J. Krumm. Understanding the importance of location, time, and people in mobile local search behavior. In *MobileHCI*, 2011.

[8] P. Venetis, H. Gonzalez, C. Jensen, and A. Halevy. Hyper-local, directions-based ranking of places. In *VLDB*, 2011.

[9] R. White and E. Horvitz. Web to World: Predicting transitions from self-diagnosis to the pursuit of local medical assistance in web search. In *AMIA Annual Symposium*, 2010.

[10] R. White and E. Horvitz. From Web search to healthcare utilization: Privacy-sensitive studies from mobile data. *JAMIA*, 2013.