Our House, In The Middle Of Our Tweets

Dan Tasse, Alex Sciuto, and Jason I. Hong
Human-Computer Interaction Institute
Carnegie Mellon University
dantasse@cmu.edu, sciutoalex@gmail.com, jasonh@cs.cmu.edu

Abstract

Twitter, Flickr, Instagram, and other public social media sites have inspired lots of analysis of public geotagged posts. In order to understand these posts, it is important to know where their authors live. Based on a study of 195 prolific Twitter users in the Pittsburgh area, and their ground truth home locations, we show that simple algorithms can find about 80% of people's home addresses within 1 kilometer. We show why this is near the upper bound of feasibility, show that studying as few as 10 tweets can achieve almost the same results, and discuss implications for future social media analyses.

Introduction

"Where do you live?" It's a simple question, but its answer shapes and defines many aspects of our lives. We spend the majority of our lives working, living, and socializing in the city and neighborhood we call home.

When users post content on social media sites, they are often given the opportunity to include their current geographic location. These geotags publicly indicate a user was at a specific location at a specific time. As users go throughout their lives posting information to social media sites, they can accrue a long trail of geotagged posts. But without knowing users' home locations, these posts are decontextualized events: "someone said something here at this moment."

If we can figure out users' home locations, though, we can start to say something about people who live in a certain neighborhood: what they talk about; where they eat, drink, shop, and travel; who they interact with. We can separate the views of tourists from the views of locals in order to learn what restaurants more knowledgeable people recommend. We can understand who works in a neighborhood and who lives in a neighborhood to see how their views differ on new infrastructure projects. This information, in turn, can help tourists, new residents, businesses, and even city planners to know more about a given location.

In order to explore how well we can infer users' home locations based on their geotags, we conducted a study of 195 Twitter users in the Pittsburgh area. We gathered ground truth data about where these people live, and then gathered the previous year's tweets as the body of geotags.

On the surface, it might seem that using these kinds of geotags to find users' homes is trivial because geotags provide specific location data. One might imagine looking at a set of geotags from a user and declaring that the most common place is that user's home. However, our evaluation shows that this simplistic approach misses a large portion of users' homes. In addition, it remains unclear how accurate or inaccurate this approach actually is: what percent of people it works for, and how close it finds their houses. This lack of clarity inspired our work.

In this paper, we present the results of testing different home location prediction methods. After investigating multiple algorithms, we found one that worked significantly better than our baseline approach. Our contributions are:

- This algorithm, a variant of Grid Search, which finds a location within 1km of almost 80% of users' homes.
- Evidence, based on in-depth error analysis, that the upper bound for any possible Twitter home-finding algorithm is likely around 85%.

Related Work

Related work falls into two categories: motivations for finding users' homes and previous attempts to do so.

Motivations and Applications for Finding Homes

Finding a person's home could be important in many ways, to learn about places and the people who live there. Censuses and other government efforts gather data about demographics of an area, but they are expensive and limited. There is often a need to understand people's behaviors in a city which change quickly and are not reflected by censuses anyway. For example, Livehoods (Cranshaw et al. 2012) showed how people's social media posts can describe people's actions in places better than official boundaries.

One could figure out land zoning, for example, as in (Toole et al. 2012), by classifying places as residential if a lot of people's home locations are there. Similarly, (Smith-Clarke, Mashhadi, and Capra 2014) were able to understand where deprivation is spread throughout a country based on call data records. Drawing similar conclusions from social media would be difficult, due to the scarcity of the data, but adding home locations would help.

On the other hand, social media offers additional benefits, thanks to the content that is associated with each post. For example, (Eichstaedt et al. 2015) used tweet language to predict heart disease mortality. However, they had to rely on the unreliable user-provided location to determine where people live. Similarly, news organizations often mine tweets to report on breaking news, but it's difficult to tell whether a source is trustworthy unless they know where that person lives (Abrol and Khan 2010).

Previous Attempts to Find People's Homes

Given the importance of the task, it is not surprising that many researchers have attempted to find users' homes from a wide variety of sources. Early work has been done with cell phone call data records (Isaacman et al. 2011; de Montjoye et al. 2013) or GPS trackers carried by people (Ashbrook and Starner 2003) or on cars (Krumm 2007), but our work focuses on social media because it is more widely available on a much larger scale.

Researchers have found the locations of tweets based on their contents (Cheng, Caverlee, and Lee 2010; Mahmud, Nichols, and Drews 2012), the locations of users based on their tweet contents (Chandra, Khan, and Muhaya 2011), the locations of users based on the contents of their photographs (Zheng et al. 2015), Foursquare users' locations based on their mayorships (Pontes et al. 2012b), and the location of users based on their friends' locations (Backstrom, Sun, and Marlow 2010).

All of these sources provided some useful information and accurate results, but we approached this study with three distinguishing goals. First, we aimed to achieve higher accuracy (by using a richer data source); second, we wanted to know the limits of accuracy given this richer data source; and third, we wanted to truly validate our approach by using ground truth data.

Pontes *et al* (2012a) expanded the scope in a comparative study of finding users' home addresses on Twitter, Foursquare, and Google+. However, they relied on each user's location field instead of getting ground truth. While expedient, this approach leads to inaccuracy because location fields are largely an unreliable source of data. Users often list an incorrect location and rarely list a location finer than city-level (Hecht et al. 2011). We hope to expand upon this work by finding a ground truth data set, which helps us to better evaluate how accurate various methods are.

Data Collection

To build a ground truth data set, we began by collecting 3.3 million geotagged tweets via Twitter's public streaming API. This API allows a developer to listen for new tweets that match a geographic parameter in near real time, so we chose to stream all tweets geotagged within 0.2 degrees latitude and longitude from the center of Pittsburgh. The rectangle we selected had corners at (40.241667, -80.2) and (40.641667, -79.8), and we collected tweets from January 2014 to January 2015. Following other work (Morstatter et al. 2013), we can assume that if our sample is less than about 1% of all tweets, we collected the vast majority of geotagged tweets in the region.

Near the end of that year, we used our data set of streamed geotagged tweets to compile a list of the 4119 most prolific tweeters for analysis, in order to ensure that our participants had enough geotagged tweets to analyze. We recruited these prolific tweeters to take a survey by tweeting a link to them. Our survey asked seven questions: their age, gender, home address, length of time they had lived there, work address, standard commute mode, and any other places they spend a lot of time. Respondents were paid with a $5 Amazon.com gift card via email. We received 195 responses.

For each of our 195 users, we used the non-streaming Twitter API to gather that user's previous 3,200 tweets (the maximum number allowed by Twitter). We added any geotagged tweets that occurred outside of Pittsburgh to our data set. The data collection and survey process were approved by our university's Institutional Review Board.

Data Set Descriptive Statistics

Our final data set consisted of 146,852 geotagged tweets from 195 users, who had a median of 533 geotagged tweets (mean=753, min=15, 1st quartile=271, 3rd quartile=1050, max=3639). These represented a subset of all of their tweets; the median percent geotagged was 41.1% (mean=46.2%, min=2.3%, 1st quartile=25.1%, 3rd quartile=61.6%, max=100%).

One notable surprise in our data set was that we had many young participants (mean=26.9, median=22). This may be because Twitter is most popular with younger users (Duggan et al. 2015) or because younger users felt more comfortable revealing their personal information on our survey. Many of these young 18-22 year old participants were students who had multiple "homes": they lived at their family home (often outside of Pittsburgh) during the summer and at their campus home (in Pittsburgh) during the school year. Because the school year lasts 8 months or more, we asked them on the survey for their campus home, but many of them still put their family home. As a result, we manually edited 19 students' "home" addresses to be their campus addresses, based on inspection of their tweets showing places where they talked about being "home" near a university.

Methods for Finding Home

In this section, we present a systematic evaluation of several algorithms for finding users' homes. In this paper, by "finding users' homes", we mean predicting a latitude-longitude point that is as close as possible to the geocoded address that they provided. We do not do reverse geocoding to find a street address.

Baseline (Mode of Geotagged Tweets)

As a trivial baseline, we binned tweets by rounding each tweet to the nearest 0.01 degree of latitude and longitude, then predicted that the bin with the most tweets (i.e. the mode) was the user's home location.

Last Destination, Weighted Median, Largest Cluster

Krumm (2007) found people's homes based on GPS traces of their cars. We re-implemented three of his methods:

- Last Destination, where we take the median of the latitude and longitude of all points that are the last coordinate pair of the day (where a day ends at 3:00 AM)
- Weighted Median, where each point is weighted by the time until the next point.
- Largest Cluster, using the scikit-learn (Pedregosa et al. 2011) implementation of agglomerative clustering on all tweet locations.

Grid Search

We binned tweets as in the Mode algorithm, but did so recursively, as in (Cheng et al. 2011). First we rounded tweets to the nearest whole number degree and discarded all tweets outside the most common bin. We repeated this rounding to the nearest 0.1 degree, the nearest 0.01 degree, the nearest 0.001 degree, and the nearest 0.0001, predicting the latter as their home.

Multi-level DBSCAN

To cluster points in a more principled way, we used the DBSCAN algorithm (Ester et al. 1996), as implemented in the scikit-learn library (Pedregosa et al. 2011), to cluster tweets into clusters of different sizes. We set the Eps parameter (maximum distance between two samples in the same neighborhood) to be 0.2 degrees (latitude/longitude) for "city-level" clusters, 0.005 degrees for "neighborhood-level" clusters, and 0.0005 degrees for "building-level" clusters [1].

For each user, we chose the city-level cluster with the most tweets, then chose the neighborhood-level cluster with the most tweets, then the building-level cluster with the most tweets. We guessed that the centroid of the building-level cluster was the user's home location.

Grid Search Without Cross-posts

Given the similar accuracy of grid search and DBSCAN, we returned to grid search with a revised data set. We realized that 10.4% of our Twitter data set (15,261 of 146,852 tweets) were cross-posts from social apps. These apps include (in descending order of frequency) Foursquare/Swarm, Instagram, Untappd, Path, Camera on iOS, Spotify, MLB.com At the Ballpark, Frontback, Wordpress.com, Klout, LivingSocial, Sportacular, and MySpace. In each of these social apps, tweeting was a byproduct of another action (as opposed to Twitter clients such as Tweetdeck and Tweetcaster). Furthermore, most of these are intended to be used outside the home. Therefore, they cannot help (and indeed would hurt) any home-finding algorithm. We removed them from the data set and performed grid search and DBSCAN again.

[1] Of course, "distance" does not make sense in terms of degrees longitude, because the length of a degree of longitude varies based on the latitude. However, because most of the points we considered were at a similar latitude, we accepted this inaccuracy in order to test the method.

We then reasoned that nighttime tweets (from 8:00PM to 6:00AM) would be more predictive of home location than daytime tweets, so we removed daytime tweets and ran our algorithms again. This removed 77,122 of our tweets, leaving us with 54,469 tweets. We found the highest accuracy removing both of these data sets.

Finding Home Results

For each algorithm, we computed the distance from the user's true home to the algorithm's prediction of home. We computed median prediction error across all users, as well as the percent of users with prediction error less than 100 meters, 1 km, and 5km. Results are in Table 1.

By all metrics, grid search after removing cross-posts and daytime posts was the most successful. We can predict almost 80% of people's homes within 1km, localizing them at about the neighborhood level.

Locating people at the building level is a bit more difficult; just over half of people's exact home location within 100m could be predicted. This may be because of GPS inaccuracy or because people tweet in their neighborhoods but not at their homes. In our data set, over 10% of people had zero tweets within 100m of their home, while only 1% had zero tweets within 1km, so we chose 1km as a reasonable threshold. Geotagged tweets are currently not a promising means to find users' homes at the building level.

How many tweets are necessary?

Grid search after removing daytime posts and crossposts was the most effective method, but it led to another question: if an application wants to find someone's home location, how many tweets does it need? To answer this, we re-ran grid search on the last N non-daytime non-crosspost tweets per user, for various values of N. Our results are in Table 2; they show that, as expected, more tweets allows for a more accurate prediction, but even as few as 10 tweets allows for remarkably high accuracy.

Error Analysis and Discussion

We have shown that it is not hard, but not trivial, to find most geotagging Twitter users' home neighborhoods based on their geotagged tweets. We found multiple common reasons that we could not find more users' exact homes.

People Moving Residences Forty-nine people in our data set (25.1%) had moved within the last 6 months, and we were unable to accurately locate 19 of them, mostly because they had not yet tweeted very much at their new house. In the United States overall, 11.6% of all people moved within the last year. Assuming that this distribution is roughly uniform, then about half that many, or 5.8%, moved within the last six months. Therefore, our sample has over four times as many recent movers as the average in the United States. Taking into account our users' younger age does not solve this problem: only 23.1% of 20-24-year-olds have moved in the last year (U.S. Census Bureau 2015), so we would expect 11.5% of young people to have moved in the last 6 months, not 25.1%.

Algorithm	Cross-posts removed	Night only	Median error	% of users within 100m	% of users within 1km	% of users within 5km
Mode			553m	1.5	63.1	79.0
Grid Search			57m	54.4	73.3	86.7
Grid Search	✓		54m	56.2	76.8	88.1
Grid Search		✓	51m	56.2	77.3	87.6
Grid Search	✔	✔	**49m**	**56.9**	**79.0**	**88.2**
Multi-level DBSCAN			75m	52.8	72.3	87.2
Multi-level DBSCAN	✓	✓	75m	52.3	74.4	87.2
Last Destination			350m	40.5	66.7	85.6
Last Destination	✓	✓	520m	33.3	64.1	82.6
Weighted Median	✓	✓	400m	40.5	65.6	79.0
Largest Cluster	✓	✓	362m	33.8	69.7	87.1

Table 1: Results for each algorithm trying to predict each user's home. Best results are in bold. Results for Weighted Median and Largest Cluster without cross-posts and daytime posts removed were significantly worse, so we do not present them here.

Last N Tweets	Median error	% users w/in 100m	% users w/in 1km	% users w/in 5km
1	245m	44.6	61.7	74.1
5	84m	51.3	66.3	76.2
10	62m	58.0	75.1	81.9
100	65m	56.0	74.6	86.0
1000	51m	57.0	79.3	88.6

Table 2: Results using grid search on the most recent N non-crosspost non-daytime tweets for each user, for various values of N. More tweets allows better prediction, but prediction is remarkably good with as few as 10 tweets.

It may seem natural to propose that a solution to recent movers is to build a model that uses only the most recent tweets. However, in our sample, about half of all users had not posted a geotagged tweet in the month before the survey, so recency-based approaches would exclude too many users.

Twitter Account Lifespan Since we began collecting data, 4 of our 195 participants (2.1%) closed their accounts or protected their tweets. As a result, we were not able to search for any of their newer tweets; we could only rely on the tweets we had collected already via the streaming API.

Frequent Travelers and Students Three people in our data set are frequent cross-country travelers. They had multiple homes or constantly moved between cities. Because of their frequent movement, they had roughly equal tweets in a variety of cities. For them, "home" itself was ill-defined.

In addition, the problem we mentioned before about students affected our accuracy here as well. We manually corrected 19 students who incorrectly listed their family's home as their home, so that we accurately predicted their campus home. However, for four students, we predicted their family's home instead of their campus home. Students are a special population who often move frequently between multiple homes, so finding their "real" home will always be difficult.

Best Case Scenario Removing these 30 people (19 recent movers, 4 closed accounts, 3 frequent travelers, and 4 students) leaves a "best case" scenario where we should be able to predict 165 users' home locations. This means our estimated upper limit of correct prediction is 84.6%.

Difficulty of defining a home location

Comparing the results of our survey with our estimated home locations has revealed a discrepancy between what people consider home and what they write when asked to list their home address. When trying to locate people in the future, one might try to categorize people as having one home, having multiple homes, or having no particular home. More deeply, though, the students and frequent movers in our data set call into question the usefulness of our goal of defining a "home" location at all. These cases suggest an alternative goal of defining a number of locations that are important to individuals, or of defining how "homely" a tweet cluster is. A number of other researchers have attempted this using different sets of data (Krumm 2007; Qu and Zhang 2013); this may be a more fruitful approach.

Generalizability of Home-Finding Results

The simplicity of our algorithm (grid search after removing cross posts and daytime posts) is beneficial in two ways: it is easy to re-implement, and it avoids concerns of overfitting. Overfitting is another reason we chose not to pursue a more complex machine learning-based algorithm. If we had proposed a more complex solution, it may only work on our limited set of tweets around Pittsburgh, but we feel confident that our algorithm should perform well on other Twitter data sets (and may even generalize to other social media).

Conclusion

Finding a user's home is an important first step in order to accurately make sense of their geotagged tweets. We have shown a simple but effective way to find about 80% of users' tweets within 1km of their homes, by using grid search after removing daytime posts and cross-posts. We have also shown, through a detailed error analysis, that a reasonable upper bound would be about 84%. We hope that this helps developers to better make use of this rich source of data.

Acknowledgements

This research was funded in part by the HCII Foundation.

References

Abrol, S., and Khan, L. 2010. TweetHood: Agglomerative clustering on fuzzy k-closest friends with variable depth for location mining. *Proceedings - SocialCom 2010: 2nd IEEE International Conference on Social Computing* 153–160.

Ashbrook, D., and Starner, T. 2003. Using gps to learn significant locations and predict movement across multiple users. *Personal and Ubiquitous Computing* 7(5):275–286.

Backstrom, L.; Sun, E.; and Marlow, C. 2010. Find me if you can: improving geographical prediction with social and spatial proximity. *Proceedings of the 19th international conference on World wide web* 61–70.

Chandra, S.; Khan, L.; and Muhaya, F. B. 2011. Estimating Twitter User Location Using Social Interactions–A Content Based Approach. *2011 IEEE Third Int'l Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third Int'l Conference on Social Computing* 838–843.

Cheng, Z.; Caverlee, J.; Lee, K.; and Sui, D. 2011. Exploring Millions of Footprints in Location Sharing Services. *Proceedings of ICWSM*.

Cheng, Z.; Caverlee, J.; and Lee, K. 2010. You Are Where You Tweet: A Content-Based Approach to Geo-locating Twitter Users. *Proc. of the 19th ACM International Conference on Information and Knowledge Management* 759–768.

Cranshaw, J.; Schwartz, R.; Hong, J. I.; and Sadeh, N. 2012. The Livehoods Project: Utilizing Social Media to Understand the Dynamics of a City. *ICWSM*.

de Montjoye, Y.-A.; Hidalgo, C. a.; Verleysen, M.; and Blondel, V. D. 2013. Unique in the Crowd: The privacy bounds of human mobility. *Nature Scientific reports* 3.

Duggan, M.; Ellison, N. B.; Lampe, C.; Lenhart, A.; and Madden, M. 2015. Social media update 2014. *Pew Research Center* (January):18.

Eichstaedt, J. C.; Schwartz, H. A.; Kern, M. L.; Park, G.; Labarthe, D. R.; Merchant, R. M.; Jha, S.; Agrawal, M.; Dziurzynski, L. a.; Sap, M.; Weeg, C.; Larson, E. E.; Ungar, L. H.; and Seligman, M. E. P. 2015. Psychological Language on Twitter Predicts County-Level Heart Disease Mortality.

Ester, M.; Kriegel, H.-P.; Sander, J.; and Xu, X. 1996. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *KDD*, volume 2, 635–654.

Hecht, B.; Hong, L.; Suh, B.; and Chi, E. H. 2011. Tweets from Justin Biebers Heart: The Dynamics of the Location Field in User Profiles. In *CHI Conference on Human Factors in Computing Systems*, 237–246.

Isaacman, S.; Becker, R.; Cáceres, R.; Kobourov, S.; Martonosi, M.; Rowland, J.; and Varshavsky, A. 2011. Identifying important places in peoples lives from cellular network data. In *Pervasive computing*. Springer. 133–151.

Krumm, J. 2007. Inference Attacks on Location Tracks. *Pervasive Computing* 10(Pervasive):127–143.

Mahmud, J.; Nichols, J.; and Drews, C. 2012. Where Is This Tweet From? Inferring Home Locations of Twitter Users. In *ICWSM*, 511–514.

Morstatter, F.; Pfeffer, J.; Liu, H.; and Carley, K. 2013. Is the sample good enough? Comparing data from Twitter's streaming API with Twitter's firehose. *Proceedings of ICWSM* 400–408.

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; and Duchesnay, E. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12:2825–2830.

Pontes, T.; Magno, G.; Vasconcelos, M.; Gupta, A.; Almeida, J.; Kumaraguru, P.; and Almeida, V. 2012a. Beware of what you share: Inferring home location in social networks. In *12th IEEE International Conference on Data Mining Workshops, ICDMW 2012*, 571–578.

Pontes, T.; Vasconcelos, M.; Almeida, J.; Kumaraguru, P.; and Almeida, V. 2012b. We Know Where You Live: Privacy Characterization of Foursquare Behavior. *Proceedings of the ACM Conference on Ubiquitous Computing* 898.

Qu, Y., and Zhang, J. 2013. Regularly visited patches in human mobility. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, 395–398. New York, NY, USA: ACM.

Smith-Clarke, C.; Mashhadi, A.; and Capra, L. 2014. Poverty on the Cheap: Estimating Poverty Maps Using Aggregated Mobile Communication Networks. In *Proceedings of the Conference on Human Factors in Computing Systems - CHI*.

Toole, J.; Ulm, M.; González, M.; and Bauer, D. 2012. Inferring land use from mobile phone activity. *Proceedings of the ACM SIGKDD International Workshop on Urban Computing - UrbComp '12* 1.

U.S. Census Bureau. 2015. Current population survey data on migration/geographic mobility. `http://www.census.gov/hhes/migration/data/cps.html`.

Zheng, D.; Hu, T.; You, Q.; Kautz, H.; and Luo, J. 2015. Towards lifestyle understanding: Predicting home and vacation locations from user's online photo collections. In *AAAI International Conference on Web and Social Media*.