Tweets from Justin Bieber's Heart: The Dynamics of the "Location" Field in User Profiles

Brent Hecht[*], Lichan Hong[†], Bongwon Suh[†], Ed H. Chi[†]

*Northwestern University
Electrical Engineering and Computer Science
brent@u.northwestern.edu

[†]Palo Alto Research Center
Augmented Social Cognition Group
3333 Coyote Hill Road, Palo Alto, CA
{hong,suh,echi}@parc.com

ABSTRACT

Little research exists on one of the most common, oldest, and most utilized forms of online social geographic information: the "location" field found in most virtual community user profiles. We performed the first in-depth study of user behavior with regard to the location field in Twitter user profiles. We found that 34% of users did not provide real location information, frequently incorporating fake locations or sarcastic comments that can fool traditional geographic information tools. When users did input their location, they almost never specified it at a scale any more detailed than their city. In order to determine whether or not natural user behaviors have a real effect on the "locatability" of users, we performed a simple machine learning experiment to determine whether we can identify a user's location by only looking at what that user tweets. We found that a user's country and state can in fact be determined easily with decent accuracy, indicating that users implicitly reveal location information, with or without realizing it. Implications for location-based services and privacy are discussed.

Author Keywords

Location, location-based services, Twitter, privacy, geography, location prediction, social networks

ACM Classification Keywords

H5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

INTRODUCTION

Interest in geographic information within the HCI community has intensified in the past few years. Academic HCI research has seen an increase in the number of papers on geographic information (e.g. [8, 18, 20, 21, 24, 26]). Industry has experienced an even greater spike in activity. Geotagged photo map interfaces have become commonplace (e.g. in Flickr and iPhoto), Google's Buzz has integrated a geographic component since its inception, and companies like Yelp have embraced the geographic nature of their user-generated content wholeheartedly.

Despite this increased interest in "geo", one of the oldest, most common forms of geographic information in the Web 2.0 world has escaped detailed study. This is the information that exists in the "location" field of user profiles on dozens of immensely popular websites. Facebook has had "Current City" and "Hometown" fields for years. Flickr allows users to enter their hometown and current location in their user profile, and the recently-launched music social network Ping by Apple has "Where I Live" as one of its profile fields.

This gap in understanding has not stopped researchers and practitioners from making ample use of the data entered into location fields. In general, it has been assumed that this data is strongly typed geographic information with little noise and good precision – an assumption that has never been validated. Backstrom et al. [1], for instance, wrote that "there is little incentive to enter false information, as leaving the field blank is an easier option". Similarly, Twitter reported that many location-based projects "are built using the simple, account-level location field folks can fill out as part of their profile". [25] This includes the "Nearby" feature of Twitter's official iPhone app, which is designed to show tweets that are close to the user's present location.

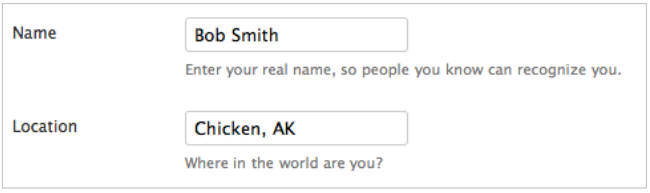

Figure 1. A screenshot from the webpage on which Twitter users enter location information. Location entries are entirely freeform, but limited to 30 characters.

In this paper, we conduct an in-depth study of user profile location data on Twitter, which provides a freeform location field without additional user interface elements that encourage any form of structured input (Figure 1). The prompt is simply "Where in the world are you?" This environment allows us to observe users' natural, "organic"

CHI 2011, May 7–12, 2011, Vancouver, BC, Canada.

behavior as best as possible, thus illuminating actual user practices.

In the first part of this paper, we report the results derived from an extensive investigation of thousands of users' location entries on Twitter. We demonstrate that users' behavior with respect to the location field is richly varied, contrary to what has been assumed. We also show that the information they enter into the field is both highly diverse and noisy. Finally, our results suggest that most users organically specify their location at the city scale when they do specify their location.

For practitioners and researchers, it may be important to discover the rough location of the large percentage of users who did not disclose their true location. How can location-based services (LBS) ranging from information retrieval to targeted advertising leverage location field information given its noisy nature? Do users reveal location information through other behaviors on Twitter that can be used to effectively "fill in" the location field?

To answer both these questions, we considered users' *implicit* location sharing behavior. Since there are many forms of this implicit behavior, we decided to evaluate the most basic: the act of tweeting itself. In other words, how much information about her or his location does the average Twitter user disclose implicitly *simply by tweeting*? The second part of this paper presents a machine learning experiment that attempts to answer this question. We found that by observing only a user's tweets and leveraging simple machine learning techniques, we were reasonably able to infer a user's home country and home state. While we might never be able to predict location to GPS-level accuracy reliably using tweet content only, knowing even the country or the state of a user would be helpful in many areas such as answering search queries and targeted advertisement. In other words, users' most basic behavior on Twitter somewhat implicitly "fills out" the location field for them, better enabling LBS but also raising privacy concerns.

In summary, our contributions are fourfold:

- To the best of our knowledge, we provide the first in-depth study of user behavior in relation to one of the oldest and most common forms of online social geographic information: the location field in user profiles.
- We find that users' natural location field behavior is more varied and the information they submit is more complex than previously assumed.
- We show that the traditional tools for processing location field information are not properly equipped to handle this varied and noisy dataset.
- Using simple machine learning techniques to guess at users' locations, we demonstrate that the average user reveals location information simply by tweeting.

Following this introduction and a related work section, we describe how we collected our data from Twitter, as this is central to both of our studies. Next, we detail our characterization study and its implications. Following that, we describe the machine learning study. Finally, we close with a conclusion and discussion of future work.

Finally, before moving on, it is important to note that this work is descriptive in nature and does not focus on causal explanations for users' natural behavior. For instance, some users may decide not to enter their location for privacy reasons, while others may do so due to lack of interest or the belief that interested people already know their location. While some clues as to users' motivations can be gleaned from our first study, we leave in-depth causal analysis to future work.

RELATED WORK

Work related to this paper primarily arises from four areas: (1) research on microblogging sites like Twitter, (2) work on location disclosure behavior, (3) the location detection of users who contribute content to Web 2.0 sites, and (4) prediction of private information.

Various researchers have studied Twitter usage in depth. For instance, Honeycutt and Herring [10] examined the usage of the "@" symbol in English tweets. boyd et al. [3] studied how retweets are used to spread information. By manually coding 3,379 tweets, Naaman et al. [17] found that 20% of users posted tweets that are informational in nature, while the other 80% posted tweets about themselves or their thoughts.

With regard to the Twitter location field, Java et al. [11] found that in their dataset of 76K users, 39K of them provided information in their "location" field. They applied the Yahoo! Geocoding API[1] to the location field of these 39K users to show the geographical distribution of users across continents. Using the self-reported "utc_offset" field in user profiles, Krishnamurthy et al. [12] examined the growth of users in each continent over time. In the area of machine learning, Sakaki et al. [22] used the location field as input to their spatiotemporal event detection algorithms.

Location disclosure behavior has been investigated both in the research community and in the popular press. For instance, Barkhuus et al. [2] concluded that this behavior must be understood in its social context. In our case, this context is the entire "Twittersphere", as all data examined was in public profiles. Ludford et al. [15] identified several heuristics for how people decide which locations to share, such as "I will not share residences [or] private workplaces." In the popular press, the New York Times recently featured an article [4] reporting that just 4% of U.S. residents had tried location-based services.

[1]http://developer.yahoo.com/maps/rest/V1/geocode.html

In the third area – location detection – the most relevant works include Lieberman and Lin [13], Popescu and Grefenstette [19], and Backstrom et al. [1]. The recentness of these papers, all published in the past two years, demonstrates that this is an active area of research. Lieberman and Lin sought to determine the location of Wikipedia users, but did so using very specific properties of the Wikipedia dataset that do not generalize to the rest of the Web 2.0 world. In addition, they did not examine the natural behavior of Wikipedia users on their "user pages", which are the Wikipedia equivalent of user profiles.

Popescu and Grefenstette [19] attempted to predict the home country of Flickr users through the analysis of their place name photo tags and latitude and longitude geotags. In contrast to both this paper and the Lieberman and Lin work, once our model has been trained, our location prediction algorithms do not depend on a user submitting any geographic information. Popescu and Grefenstette also did no qualitative examination.

Backstrom et al. [1] used the social network structure of Facebook to predict location. As noted below, our work focuses on the content submitted by users, not the social network, although both approaches could be combined in future work.

In terms of prediction of profile fields or other withheld information, our work stands out from other recent research (e.g. [1, 14]) in two ways: (1) first we examine the user practices surrounding the information that we are trying to predict, and (2) we make predictions solely from content innate to its medium and do not leverage any portion of the social graph.

DATA COLLECTION

From April 18 to May 28, 2010, we collected over 62 million tweets from the Spritzer sample feed, using the Twitter streaming API[2]. The Spritzer sample represents a random selection of all public messages. Based on a recent report that Twitter produced 65 million tweets daily as of June 2010 [23], we estimate that our dataset represents about 3-4% of public messages.

From these 62 million tweets, we further identified the tweets that were in English using a two-step combination of LingPipe's text classifier[3] and Google's Language Detection API[4]. All together, we identified 31,952,964 English tweets from our 62 million tweets, representing 51% of our dataset.

This research purposely does not consider the recent change to the Twitter API that allows location information to be embedded in each individual tweet [25]. We made this choice for two reasons. First, our focus is on the geographic information revealed in the "location" field of user profiles, a type of geographic information that is prevalent across the Web 2.0 world. Second, we found that only 0.77% of our 62 million tweets contained this embedded location information. With such a small penetration rate, we were concerned about sampling biases.

STUDY 1: UNDERSTANDING EXPLICIT USER BEHAVIOR

Study 1: Methods

Our 32 million English tweets were created by 5,282,657 unique users. Out of these users, we randomly selected 10,000 "active" users for our first study. We defined "active" as having more than five tweets in our dataset, which reduced our sampling frame to 1,136,952 users (or 22% of all users). We then extracted the contents of these 10,000 users' location fields and placed them in a coding spreadsheet. Two coders examined the 10,000 location field entries using a coding scheme described below. Coders were asked to use any information at their disposal, from their cultural knowledge and human intuition to search engines and online mapping sites. Both coders agreed initially on 89.2% of the entries, and spent one day discussing and coming to an agreement on the remaining 10.8%.

The coding scheme was designed to determine the *quality* of the geographic information entered by users as well as the *scale* of any real geographic information. In other words, we were interested in examining the 10,000 location entries for their properties along two dimensions: quality and geographic scale. We measured quality by whether or not geographic information was imaginary or whether it was so ambiguous as to refer to no specific geographic footprint (e.g. "in jail" instead of "in Folsom Prison"). In the case of location field entries with even the most rudimentary real geographic information, we examined at what scale this information specified the user's location. In other words, did users disclose their country? Their state? Their city? Their address?

Since both coders are residents of the United States, only data that was determined to be within the United States was examined for scale. This choice was made due to the highly vernacular nature of many of the entries, thus requiring a great deal of cultural knowledge for interpretation.

Study 1: Results

Information Quality

As shown in Figure 2, only 66% of users manually entered any sort of valid geographic information into the location field. This means that although the location field is usually assumed by practitioners [25] and researchers (e.g. in [11] and [22]) to be a field that is as associated with geographic information as a date field is with temporal information, this is definitely not the case in our sample. The remaining one-third of users were roughly split between those that did not enter any information and those that entered either non-

[2] http://dev.twitter.com/pages/streaming_api

[3] http://alias-i.com/lingpipe/demos/tutorial/langid/read-me/html

[4] http://code.google.com/apis/ajaxlanguage/documentation/

real locations, obviously non-geographic information, or locations that did not have specific geographic footprints.

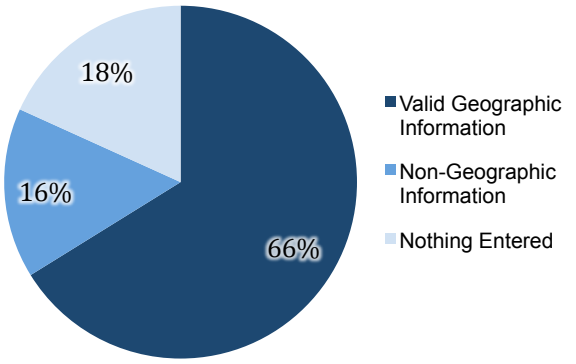

Figure 2: The distribution of manually entered location field data. Roughly one-third of users did not enter valid geographic information into the location field. 16% entered non-geographic information, while 18% entered nothing at all.

An analysis of the non-geographic information entered into the location field (the 16% in Figure 2) revealed it to be highly unpredictable in nature (see Table 1). A striking trend was the theme of Justin Bieber, who is a teenage singer. A surprising 61 users (more than 1 in 200 users) co-opted the location field to express their appreciation of the pop star. For instance, a user wrote that s/he is located in "Justin Biebers heart" (inspiring the title of this paper) and another user indicated s/he is from "Bieberacademy". Justin Bieber was not the only pop star that received plaudits from within the location field; United Kingdom "singing" duo Jedward, Britney Spears, and the Jonas Brothers were also turned into popular "locations".

Another common theme involved users co-opting the location field to express their desire to keep their location private. One user wrote "not telling you" in the location field and another populated the field with "NON YA BISNESS!!" Sexual content was also quite frequent, as were "locations" that were insulting or threatening to the reader (e.g. "looking down on u people"). Additionally, there was a prevalent trend of users entering non-Earth locations such as "OUTTA SPACE" and "Jupiter".

A relatively large number of users leveraged the location field to express their displeasure about their current location. For instance, one user wrote "preferably anywhere but here" and another entered "redneck hell".

Entering non-real geographic information into the location field was so prevalent that it even inspired some users in our sample to make jokes about the practice. For instance, one user populated the location field with "(insert clever phrase here)".

Frequency counts for these types of non-geographic information are reported in Table 1. To generate this table, non-geographic entries were coded by two human coders and the lists were merged. Categories were determined using a grounded approach, and each "location" was allowed to have zero or more categories. Because of the highly vernacular nature of this data, coders were instructed to only categorize when highly confident in their choice. As such, the numbers in Table 1 must be considered lower bounds.

Information Type	# of Users
Popular Culture Reference	195 (12.9%)
Privacy-Oriented	18 (1.2%)
Insulting or Threatening to Reader	69 (4.6%)
Non-Earth Location	75 (5.0%)
Negative Emotion Towards Current Location	48 (3.2%)
Sexual in Nature	49 (3.2%)

Table 1: A selection of the types of non-geographic information entered into the location field. Many of these categories exhibited large co-occurrence, such as an overlap between "locations" that were sexual in nature and those that were references to popular culture (particularly pop and movie stars). Percentages refer to the population of non-geographic information location field entries.

Note that, in the 66% of users who did enter real geographic information, we included all users who wrote *any* inkling of real geographic information. This includes those who merely entered their continent and, more commonly, those who entered geographic information in highly vernacular forms. For example, one user wrote that s/he is from "kcmo--call da po po". Our coders were able to determine this user meant "Kansas City, Missouri", and thus this entry was rated as valid geographic information (indicating a location at a city scale). Similarly, a user who entered "Bieberville, California" as her/his location was rated as having included geographic information at the state scale, even though the city is not real.

Information Scale

Out of the 66% of users with any valid geographic information, those that were judged to be outside of the United States were excluded from our study of scale. Users who indicated multiple locations (see below) were also filtered out. This left us with 3,149 users who were determined by both coders to have entered valid geographic information that indicated they were located in the United States.

When examining the scale of the location entered by these 3,149 users, an obvious city-oriented trend emerges (Figure 3). Left to their own devices, users by and large choose to disclose their location at exactly the city scale, no more and no less. As shown in Figure 3, approximately 64% of users specified their location down to the city scale. The next most popular scale was state-level (20%).

When users specified intrastate regions or neighborhoods, they tended to be regions or neighborhoods that engendered significant place-based identity. For example, "Orange County" and the "San Francisco Bay Area" were common entries, as were "Harlem" and "Hollywood". Interestingly, studying the location field behavior of users located within

a region could be a good way to measure the extent to which people identify with these places.

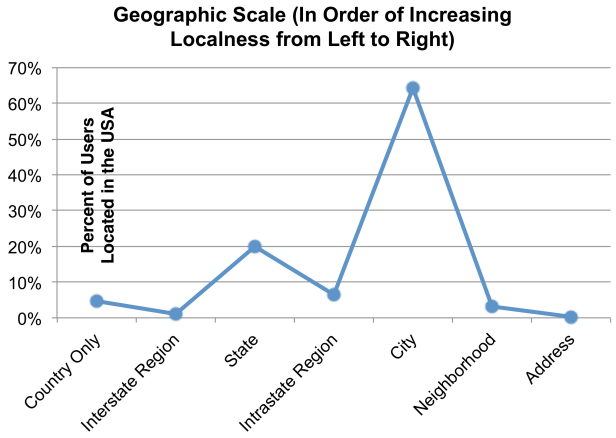

Figure 3: The scale of the geographic information entered by 3,149 users who indicated that they lived in the United States.

Multiple Locations

2.6% of the users (4% of the users who entered any valid geographic information) entered multiple locations. Most of these users entered two locations, but 16.4% of them entered three or more locations. Qualitatively, it appears many of these users either spent a great deal of time in all locations mentioned, or called one location home and another their current residence. An example of the former is the user who wrote "Columbia, SC. [atl on weekends]" (referring to Columbia, South Carolina and Atlanta, Georgia). An example of the latter is the user who entered that he is a "CALi b0Y $TuCC iN V3Ga$" (A male from California "stuck" in Las Vegas).

Automatically-entered Information

The most categorically distinct entries we encountered were the automatically populated latitude and longitude tags that were seen in many users' location fields. After much investigation, we discovered that Twitter clients such as ÜberTwitter for Blackberry smartphones entered this information. Approximately 11.5% of the 10,000 users we examined had these latitude and longitude tags in their location field. We did not include these users in Figure 2 or 3, as they did not manually enter their location data.

Study 1: Implications for Design

Failure of Traditional Geographic Information Tools

Our study on the information quality has vital implications for leveraging data in the location field on Twitter (and likely other websites). Namely, many researchers have assumed that location fields contain strongly typed geographic information, but our findings show this is demonstrably false. To determine the effect of treating Twitter's location field as strongly-typed geographic information, we took each of the location field entries that were coded as not having any valid geographic information (the 16% slice of the pie chart in Figure 2) and entered them into Yahoo! Geocoder. This is the same process used by Java et al. in [11]. A geocoder is a traditional geographic information tool that converts place names and addresses into a machine-readable spatial representation, usually latitude and longitude coordinates [7].

Of the 1,380 non-geographic location field entries, Yahoo! Geocoder determined 82.1% to have a latitude and longitude coordinate. As our coders judged none of these entries to contain any geographic information or highly ambiguous geographic information, this number should be zero (assuming no coding error). Some examples of these errors are quite dramatic. "Middle Earth" returned (34.232945, -102.410204), which is north of Lubbock, Texas. Similarly, "BieberTown" was identified as being in Missouri and "somewhere ova the rainbow", in northern Maine. Even "Wherever yo mama at" received an actual spatial footprint: in southwest Siberia.

Since Yahoo! Geocoder assumes that all input information is geographic in nature, the above results are not entirely unexpected. The findings here suggest that geocoders alone are not sufficient for the processing of data in location fields. Instead, data should be preprocessed with a geoparser, which disambiguates geographic information from non-geographic information [7]. However, geoparsers tend to require a lot of context to perform accurately. Adapting geoparsers to work with location field entries is an area of future work.

Attention to Scale in Automated Systems

Another important implication comes from the mismatch in revealed scale between the latitude and longitude generated automatically by certain Twitter clients and that revealed naturally by Twitter users. The vast majority of the machine-entered latitude and longitude coordinates had six significant digits after the decimal point, which is well beyond the precision of current geolocation technologies such as GPS. While it depends somewhat on the latitude, six significant digits results in geographic precision at well under a meter. This precision is in marked contrast with the city-level organic disclosure behavior of users. In our dataset, we found a total of only nine users (0.09% of the entire dataset) who had manually entered their location at the precision of an address, which is still less precise than a latitude and longitude coordinate expressed to six significant digits. However, this number could have been affected somewhat by the 30-character limit on the Twitter location field.

This mismatch leads us to a fairly obvious but important implication for design. Any system automatically populating a location field should do so, not with the exact latitude and longitude, but with an administrative district or vernacular region that contains the latitude and longitude coordinate. Fortunately, these administrative districts are easy to calculate with a reverse geocoding tool. Users should also be given a choice of the scale of this district or region (i.e. city, state, country), as users seem to have different preferences. This implication may apply to the

"location" field on other sites as well as the location metadata associated with user-contributed content such as tweets and photos.

Other Implications

Another design implication is that users often want to have the ability to express sarcasm, humor, or elements of their personality through their location field. In many ways, this is not a surprise; people's geographic past and present have always been a part of their identity. We are particularly interested in the large number of users who expressed real geographic information in highly vernacular and personalized forms. Designers may want to invite users to choose a location via a typical map interface and then allow them to customize the place name that is displayed on their profile. This would allow users who enter their location in the form of "KC N IT GETS NO BETTA!!" (a real location field entry in our study) to both express their passion for their city and receive the benefits of having a machine-readable location, if they so desire.

Our findings also suggest that Web 2.0 system designers who wish to engender higher rates of machine-readable geographic information in users' location fields may want to force users to select from a precompiled list of places.

People who entered multiple locations motivate an additional important implication for design. This gives credence to the approach of Facebook and Flickr, which allow users to enter both a "current" location and a "hometown" location. However, the behavior of these users also suggests that this approach should be expanded. We envision a flexible system that would allow users to enter both an arbitrary number of locations and describe each of those locations (e.g. "home", "favorite place", etc.)

STUDY 2: UNDERSTANDING IMPLICIT USER BEHAVIOR THROUGH MACHINE LEARNING

In the first study, we used human judges to look closely at the explicit information included in the location field. However, in domains such as location-based services it may be important to discover the rough location of the large percentage of users who did not disclose their true location. Privacy advocates would likely also be interested in understanding whether or not this can be done. Given the results of prior research on location detection [1, 13, 19], we wanted to determine how much *implicit* location information users disclose simply by their day-to-day tweeting behavior. To do so, we used the data gathered above to conduct a set of machine learning experiments.

The goal of these experiments was to determine users' locations simply by examining the text content of their tweets. Specifically, we sought to predict a user's country and state solely from the user's tweets. We did not have enough data to work at a city level. As noted above, the contribution here is to demonstrate the implicit location sharing behavior of users in the context of their explicit behavior (with an eye towards location-based services, as well as privacy).

Study 2: Methods

In this subsection, we describe the general methodology behind our machine learning experiments, in which we use a classifier and a user's tweets to predict the country and state of that user. First, we discuss how we modeled each Twitter user for the classifier and how we shrank these models into a computationally tractable form. Next, we highlight the methodology behind the building of our training sets for the classifier and explain how we split off a subset of this data for validation purposes. Finally, we describe our classification algorithm and sampling strategies, as well as the results of our machine learning experiments.

Model Construction and Reduction

To classify user locations, we developed a Multinomial Naïve Bayes (MNB) model [16]. The model accepts input in the form of a term vector with each dimension in the vector representing a term and the value of the dimension representing the term count in a user's tweets. We also tried advanced topic models including Explicit Semantic Analysis [6]. However, a pilot study revealed that the simple term frequency (TF) MNB model greatly outperformed the more complex models. Thus, we only report the TF results.

For computational efficiency, we settled on using a fixed-length 10,000-term vector to represent each user in all cases. We tried two different methods for picking which 10,000 terms to use. The first was the standard frequency-based selection model in which we picked the 10,000 most common terms in our corpus. We called this algorithm "COUNT", for its reliance on term counting.

We also developed a more advanced algorithm designed to select terms that would discriminate between users from different locations. This simple heuristic algorithm, which we call the "CALGARI" algorithm, is based on the intuition that a classification model would perform better if the model includes terms that are more likely to be employed by users from a particular region than users from the general population. It is our assumption that these terms will help our classifier more than the words selected by the COUNT algorithm, which includes many terms that are common in all countries or states considered (e.g. "lol").

The CALGARI algorithm calculates a score for each term present in the corpus according to the following formula:

$$CALGARI(t) = \begin{cases} 0 & if \quad users(t) < MinU \\ \dfrac{\max\left(P(t \mid c = C)\right)}{P(t)} & if \quad users(t) \geq MinU \end{cases}$$

where t is the input term, *users* is a function that calculates the number of users who have used t at least once, $MinU$ is an input parameter to filter out individual idiosyncrasies and spam (set to either 2 or 5 in our experiments), and C is a geographic class (i.e. a state or country). The max function simply selects the maximum conditional

probability of the term given each of the classes being examined. Terms are then sorted in descending order according to their scores and the top 10,000 terms are selected for the model. After picking the 10,000 terms, each user's Twitter feed was represented as a term vector using this list of 10,000 terms as dimensions, populated by the feed's term frequencies for each dimension.

A good example of the differences between CALGARI and COUNT was found in the average word vector for each algorithm for users in Canada. Among the terms with the highest weights for the CALGARI algorithm were "Canada", "Calgari", "Toronto" and "Hab". On the other hand, the top ten for COUNT included "im", "lol", "love", and "don't". Note that the CALGARI algorithm picked terms that are much more "Canadian" than those generated by the COUNT algorithm. This includes the #2 word "Calgari" (stemmed "Calgary"), which is the algorithm's namesake.

Developing Ground Truth Data

In order to build a successful classifier, we first needed to generate high-precision ground truth data. The main challenge here was to match a large group of users with their *correct* country and/or state. Through this group of users, the classifier could then learn about the tweeting patterns of each country and state population, and use these patterns to make predictions about *any* user.

Our starting point in developing the ground truth data was the 32 million English tweets created by over 5 million users. We first applied an extremely high-precision, very low-recall geocoder similar to that used in Hecht and Gergle [8]. The geocoder examines the text of the location field of each user and attempts to match it against all English Wikipedia article titles. If the location field matches (case-insensitive) a title exactly, latitude and longitude coordinates are searched for on the corresponding Wikipedia page[5]. If coordinates are found, the user is assigned that latitude and longitude as her location. If not, the user is excluded. We validated the precision of this method by testing it against the same non-geographic data that was input into the Yahoo! Geocoder in Study 1. Our Wikipedia-based geocoder correctly determined that none of the input entries was an actual location.

The Wikipedia-based geocoder and the automatically entered latitude and longitude points allowed us to identify the coordinates for 588,258 users. Next, we used spatial data available from ESRI and the United States Census to calculate the country and state (if in the United States) of the users. This process is known as reverse geocoding.

In order to avoid problems associated with having a small number of tweets for a given user, we further restricted our ground truth data to those users who had contributed ten or more tweets to our dataset. In doing so, we removed 484,449 users from consideration.

We also required that all users in our dataset have a consistent country and state throughout the sample period. A tiny minority of users manually changed their location information during the sample period. In addition, a larger minority of users had their location changed automatically by Twitter clients. This temporal consistency filter pruned an additional 4,513 users from consideration.

In the end, our ground truth data consisted of 99,296 users for whom we had valid country and state information and 10 or more tweets. As noted earlier, this ground truth data was the sampling frame for deriving our training and validation sets for all machine learning experiments.

Training and Validation Sets

In each experiment, we used a specific subset (described below) of the ground truth data as training data. Since the CALGARI algorithm and the COUNT algorithm both involve "peeking" at the ground truth data to make decisions about which dimensions to include in the term vectors, the use of independent validation sets is vital. In all experiments, we split off 33% of the training data into validation sets. These validation sets were used *only* to evaluate the final performance of each model. In other words, the system is totally unaware of the data in the validation sets until it is asked to make predictions about that data. The validation sets thus provide an accurate view of how the machine learner would perform "in the wild." We used two sampling strategies for generating training and validation sets.

Sampling Strategies

In both our country-scale and state-scale experiments, we implemented two different sampling strategies to create the training data from the ground truth data. The first, which we call "UNIFORM", generated training and validation sets that exhibited a uniform distribution across classes, or countries and states in this context. This is the sampling strategy employed by Popescu and Grefenstette [19]. The experiments based on the UNIFORM data demonstrate the ability of our machine learning methods to tease out location information in the absence of the current demographic trends on Twitter.

The second sampling strategy, which we call "RANDOM", randomly chose users for our training and validation datasets. When using "RANDOM" data, the classifier considers the information that, for example, a user is much more likely to be from the United States than from Australia given population statistics and Twitter adoption rates. In other words, prior probabilities of each class (country or state) are considered. The results from experiments on the "RANDOM" data represent the amount of location information our classifier was able to extract *given* the demographics of Twitter.

[5] Hundreds of thousands of Wikipedia articles have latitude and longitude points embedded in them by users.

Sampling Strategy	Model Selection	Accuracy	Baseline Accuracy	% of Baseline Accuracy
Country-Uniform-2500	**Calgari**	**72.71%**	**25.00%**	**291%**
Country-Uniform-2500	Count	68.44%	25.00%	274%
Country-Random-20K	**Calgari**	**88.86%**	**82.08%**	**108%**
Country-Random-20K	Count	72.78%	82.08%	89%
State-Uniform-500	**Calgari**	**30.28%**	**5.56%**	**545%**
State-Uniform-500	Count	20.15%	5.56%	363%
State-Random-20K	Calgari	24.83%	15.06%	165%
State-Random-20K	**Count**	**27.31%**	**15.06%**	**181%**

Table 2: A summary of results from the country-scale and state-scale experiments. The better performing model selection algorithm is bolded for each experiment. The CALGARI result reported is the best generated by *MinU* = 2 or *MinU* = 5.

Evaluation of the Classifier

In the end, we conducted a total of four experiments, each on a differently sampled training and validation set (Table 2). In each experiment, we tested both the CALGARI and COUNT algorithms, reporting the accuracy for both. The machine learning algorithm and training/validation set split were identical across all four experiments.

For the country-prediction experiments, we first focused on the UNIFORM sampling strategy. From our ground truth data, 2,500 users located in the United States, the United Kingdom, Canada, and Australia were randomly selected, resulting in 10,000 users total. These four countries were considered because there are less than 2,500 users in each of the other English-speaking countries represented among the 99,296 ground truth users. As noted above, 33% of these users were then randomly chosen for our validation set and removed from the training set. The remainder of the training set was passed to one of two model selection algorithms: CALGARI and COUNT. We then trained our Multinomial Naïve Bayes classifier with the models and evaluated on the validation set removed earlier.

Next, we performed the same exercise, replacing the UNIFORM with the RANDOM sampling strategy, which selected 20,000 different users from our ground truth data, all of whom lived in one of the four countries listed above.

Our state-prediction experiments were roughly the same as our country experiments, with the only major difference in the development of the UNIFORM datasets. Since the U.S. states range in population from California's 36+ million people to Wyoming's 0.5+ million people, our dataset was skewed in a similar fashion. We only had very limited data for small-population states like Wyoming. In fact, out of all our 99,296 ground truth users, we only had 31 from Wyoming. As such, we only included the 18 states with 500 or more users in our UNIFORM dataset.

Study 2: Results

Country-prediction Experiments

For the UNIFORM sampling strategy, the best performing algorithm was CALGARI, which was able to predict the country of a user correctly 72.7% of the time, simply by examining that user's tweets. Since we considered four different countries in this case, one could achieve 25% accuracy by simply randomly guessing. Therefore, we also report the accuracy of our classifier relative to the random baselines, which in the best case here was 291% (or 2.91x).

With the RANDOM sampling strategy, we needed to use a different baseline. Since 82.08% of sampled users were from the U.S., one could achieve 82.08% accuracy simply by guessing "United States" for every user. However, even with these relatively decisive prior probabilities, the CALGARI algorithm was capable of bringing the accuracy level approximately 1/3 of the way to perfection (88.9%). This represents a roughly 8.1% improvement.

State-prediction Experiments

The results of our state-prediction experiments were quite similar to those above but better. As can be seen in Table 2, the classifier's best UNIFORM performance relative to the random baseline was a great deal better than in the country experiment. The same is true for the RANDOM dataset, which included users from all 50 states (even if there were only a dozen or so users from some states).

The baselines were lower in each of these experiments because we considered more states than we did countries. The UNIFORM dataset included 18 states (or classes). The RANDOM dataset included all 50 plus the District of Columbia, with New York having the maximum representation at 15.06% of users. A baseline classifier could thus achieve 15.06% accuracy simply by selecting New York in every case.

Study 2: Discussion

Table 2 shows that in every single instance, the classifier was able to predict a user's country and/or state from the user's tweets at accuracies better than random. In most cases, the accuracy was several *times* better than random, indicating a strong location signal in tweets. As such, there is no doubt that *users implicitly include location information in their tweets*. This is true even if a user has not entered any explicit location information into the location field, or has entered a purposely misleading or humorous location (assuming that these users do not have significantly different tweeting behavior).

We did not attempt to find the optimal machine learning technique for location prediction from tweet content. As such, we believe that the accuracy of location prediction can be enhanced significantly by improving along four fronts: (1) better data collection, (2) more sophisticated

machine learning techniques, (3) better modeling of implicit behaviors, especially those involving social contexts on Twitter, and (4) inclusion of more user metadata.

Study 2: Implications

An interesting implication of our work can be derived from the conditional probabilities tables of the classifier. By studying these tables, we developed a list of terms that could be used to both assist location-based services (LBS) and launch location "inference attacks" [14]. A selection of terms that have strong predictive power at the country and state scales is shown in Table 3.

Stemmed Word	Country	"Predictiveness"
"calgari"	Canada	419.42
"brisban"	Australia	137.29
"coolcanuck"	Canada	78.28
"afl"	Australia	56.24
"clegg"	UK	35.49
"cbc"	Canada	29.40
"yelp"	United States	19.08
Stemmed Word	**State**	**"Predictiveness"**
"colorado"	Colorado	90.74
"elk"	Colorado	41.18
"redsox"	Massachusetts	39.24
"biggbi"	Michigan	24.26
"gamecock"	South Carolina	16.00
"crawfish"	Louisiana	14.87
"mccain"	Arizona	10.51

Table 3: Some of the most predictive words from the (top) Country-Uniform-Calgari and (bottom) State-Uniform-Calgari experiments. Predictiveness is calculated as a probability ratio of the max. conditional probability divided by the average of the non-maximum conditional probabilities. This can be interpreted as the number of times more likely a word is to occur given that a person is from a specific region than from the average of the other regions in the dataset. In other words, an Arizonan is 10.51 times more likely to use the term "mccain" than the average person from the other states.

There appear to be four general categories of words that are particularly indicative of one's location. As has been known in the social sciences for centuries (e.g. the gravity model [5]) and seen elsewhere with user-generated content (UGC) [9,13], people tend to interact with nearby places. While in some cases this has been shown to be not entirely true [8], mentioning place names that are close to one's location is very predictive of one's location. In other words, tweeting about what you did in "Boston" narrows down your location significantly on average.

Tweeting about sports assists in location inference significantly, as can be seen in Table 3. Similarly, our classifier found that a user from Canada was six times more likely to tweet the word "hockey" than a user from any other country in our study.

A third major category of predictive terms involves current events with specific geographic footprint, emphasizing the spatio*temporal* nature of location field data. During the period of our data collection, several major events were occurring whose footprints corresponded almost exactly with the scales of our analyses. The classifier easily identified that terms like "Cameron", "Brown", and "Clegg" were highly predictive of users who were in the United Kingdom. Similarly, using terms related to the 2010 NBA playoffs was highly indicative of a user from the United States. More generally speaking, a machine learner could theoretically utilize any regionalized phenomenon. For example, a tweet about a flood at a certain time [24, 26] could be used to locate a user to a very local scale.

Finally, regional vernacular such as "hella" (California) and "xx" (U.K.) were highly predictive of certain locations. It is our hypothesis that this category of predictive words helped our term frequency models perform better than the more complex topic models. It seems that the more abstract the topic model, the more it smoothes out the differences in spelling or slang. Such syntactic features can be powerful predictors of location, however.

Given some Twitter users' inclination towards privacy, users might value the inclusion of this predictive word list into the user interface through warnings. Moreover, given some users' inclination towards location field impishness, users may enjoy the ability to easily use this type of information to fool predictive systems. In other words, through aversion or purposeful deception, users could avoid location inference attacks by leveraging these terms.

FUTURE WORK

Much future work has arisen from this study of explicit and implicit location field behavior. The most immediate is to examine the causal reasons for the organic location disclosure behavior patterns revealed by this work. This could be explored through surveys, for example.

With regard to the classifier, we are looking into including social network information into our machine learners. This would allow us to explore the combination of content-based and network-based [1] location prediction.

We also are working to extend our predictive experiments to other cultural memberships. For instance, there is nothing about our models that could not be adapted to predict gender, age group, profession, or even ethnicity.

Other directions of future work include examining per-tweet location disclosure, as well as evaluating location disclosure on social network sites such as Facebook. Of course, accessing a large and representative sample of location field data on Facebook will be a major challenge. We have also done research investigating the ability to use the surprisingly noisy yet very prevalent "time zone" field in user profiles to assist in location prediction.

CONCLUSION

In this work, we have made several contributions. We are the first to closely examine the information embedded in user profile location fields. Through this exploration, we have shown that many users opt to enter no information or non-real location information that can easily fool geographic information tools. When users do enter their

real locations, they tend to be no more precise than city-scale.

We have also demonstrated that the explicit location-sharing behaviors should be examined in the context of implicit behaviors. Despite the fact that over one-third of Twitter users have chosen not to enter their location, we have shown that a simple classifier can be used to make predictions about users' locations. Moreover, these techniques only leverage the most basic activity in Twitter – the act of tweeting – and, as such, likely form something of a lower bound on location prediction ability.

Given the interest in LBS and privacy, we hope the research here will inspire investigations into other natural location-based user behaviors and their implicit equivalents.

ACKNOWLEDGEMENTS

We thank our reviewers, Bjoern Hartmann, and our colleagues at PARC and the Northwestern CollabLab for their helpful suggestions.

REFERENCES

1. Backstrom, L., Sun, E. and Marlow, C. Find Me If You Can: Improving Geographical Prediction with Social and Spatial Proximity. *WWW '10*, Raleigh, NC.
2. Barkhuus, L., Brown, B., Bell, M., Hall, M., Sherwood, S. and Chalmers, M. From Awareness to Repartee: Sharing Location within Social Groups. *CHI '08*, Florence, Italy, 497-506.
3. boyd, d., Golder, S. and Lotan, G. Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter. *HICSS '10*, Kauai, HI.
4. Miller, C. and Wortham, J. Technology Aside, Most People Still Decline to Be Located. *The New York Times.* August 30, 2010.
5. Fellman, J.D., Getis, A. and Getis, J. *Human Geography: Landscapes of Human Activities*. Mcgraw-Hill Higher Education, 2007.
6. Gabrilovich, E. and Markovitch, S. Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis. *IJCAI '07*, Hyberabad, 1606-1611.
7. Hecht, B. and Gergle, D. A Beginner's Guide to Geographic Virtual Communities Research. *Handbook of Research on Methods and Techniques for Studying Virtual Communities*, IGI, 2010.
8. Hecht, B. and Gergle, D. On The "Localness" Of User-Generated Content. *CSCW '10*, Savannah, Georgia.
9. Hecht, B. and Moxley, E. Terabytes of Tobler: Evaluating the First Law in a Massive, Domain-Neutral Representation of World Knowledge. *COSIT '09*, L'Aber Wrac'h, France, 88-105.
10. Honeycutt, C. and Herring, S.C. Beyond Microblogging: Conversation and Collaboration via Twitter. *HICCS '09*.
11. Java, A., Song, X., Finin, T. and Tseng, B. Why We Twitter: Understanding Microblogging Usage and Communities. *Joint 9th WEBKDD and 1st SNA-KDD Workshop '07*, San Jose, CA, 56-65.
12. Krishnamurthy, B., Gill, P. and Arlitt, M. A Few Chirps about Twitter. *First Workshop on Online Social Networks*, Seattle, WA, 19-24, 2008.
13. Lieberman, M.D. and Lin, J. You Are Where You Edit: Locating Wikipedia Users Through Edit Histories. *ICWSM '09,* San Jose, CA.
14. Lindamood, J., Heatherly, R., Kantarcioglu, M. and Thuraisingham, B. Inferring Private Information Using Social Network Data. *WWW '09*, Madrid, Spain.
15. Ludford, P.J., Priedhorsky, R., Reily, K. and Terveen, L.G. Capturing, Sharing, and Using Local Place Information. *CHI '07*, San Jose, CA, 1235-1244.
16. McCallum, A. and Nigam, K. A Comparison of Event Models for Naive Bayes Text Classification. *AAAI-98 Workshop on Learning for Text Categorization*, 41-48.
17. Naaman, M., Boase, J. and Lai, C.-H. Is it Really About Me? Message Content in Social Awareness Streams. *CSCW '10*, Savannah, GA.
18. Panciera, K., Priedhorsky, R., Erickson, T. and Terveen, L.G. Lurking? Cyclopaths? A Quantitative Lifecycle Analysis of User Behavior in a Geowiki. *CHI '10*, Atlanta, GA, 1917-1926.
19. Popescu, A. and Grefenstette, G. Mining User Home Location and Gender from Flickr Tags. *ICWSM '10*.
20. Priedhorsky, R. and Terveen, L.G. The Computational Geowiki: What, Why, and How. *CSCW '08*, San Diego.
21. Reddy, S., Shilton, K., Denisov, G., Cenizal, C., Estrin, D. and Srivastava, M. Biketastic: Sensing and Mapping for Better Biking. *CHI '10*, Atlanta, GA, 1817-1820.
22. Sakaki, T., Okazaki, M. and Matsuo, Y. Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors. *WWW '10*, Raleigh, NC.
23. Schonfeld, E. Costolo: Twitter Now Has 190 Million Users Tweeting 65 Million Times A Day. *TechCrunch.* http://techcrunch.com/2010/06/08/twitter-190-million-users/, 2010.
24. Starbird, K., Palen, L., Hughes, A.L. and Vieweg, S. Chatter on The Red: What Hazards Thread Reveals about the Social Life of Microblogged Information. *CSCW '10*, Savannah, GA.
25. Twitter. Location, Location, Location. *twitterblog.* http://blog.twitter.com/2009/08/location-location-location.html, 2009.
26. Vieweg, S., Hughes, A.L., Starbird, K. and Palen, L. Microblogging During Two Natural Hazards Events: What Twitter May Contribute to Situational Awareness. *CHI '10*, Atlanta, GA, 1079-1088.