2014 47th Hawaii International Conference on System Science

Using Large Scale Aggregated Knowledge for Social Media Location Discovery

Dennis Thom, Harald Bosch, Robert Krüger, Thomas Ertl
Institute for Visualization and Interactive Systems
University of Stuttgart
Stuttgart, Germany
Email: <firstname.lastname>@vis.uni-stuttgart.de

***Abstract*—Geospatial analysis of location-enabled social media networks can be utilized to generate vital insights in areas where situational awareness is important, such as disaster prevention and crisis response. However, several recent approaches struggle under the challenge that only a small fraction of the data is actually provided with precise geo-tags or even GPS information of their origin. In this work we introduce two strategies that are suitable to assign probable locations of origin to social media messages of unknown locations. They are based on aggregated knowledge about the author and/or the textual content of the message. Using our prototype implementation and a collected dataset comprising more than one year of geolocated Twitter data, we evaluate the effectiveness of our strategies. Our results show that we can locate up to 74% of all messages that were written in specific cities and about 20% of messages written in specific districts.**

***Keywords*-Data mining; Text mining; Predictive models; Decision support systems**

I. Introduction

In its seventh year of existence Twitter has gathered 200 million users composing more than 400 million Tweets per day[1]. Sometimes the Tweets are provided together with information about their location of origin either as geo-tag (e.g. *San Francisco*) or as precise geolocation determined by Global Positioning System (GPS)-enabled mobile devices. Since the location-enabled messages can often be interpreted as situation reports of ongoing local events, the data offers important new possibilities for situation awareness applications such as location related market analysis or city planning, but also in areas where human lifes are at stake such as disaster prevention and emergency management. Because of the global distribution of Twitter users, recent research has shown that such data can have great impact on the information gathering and decision making process during major incidents like wildfires, floodings, hurricanes or epidemics [1], [2], [3], [4], [5], [6], [7]. Based on these observations, researchers have designed several geovisualization systems. They help experts to harvest and filter the data, to explore the spatiotemporal relationships between ongoing events and eyewitness reports, and to generate insights based on aggregated content analysis [8], [9], [10]. Although more than 7 million location-enabled messages are produced each day and although their amount is steadily growing (see Figure 1), they still only comprise about 1.7% of the complete data volumes. This poses a severe challenge to researchers and analysts as the location-enabled data in particular offers high chances to identify event-related and first-hand information.

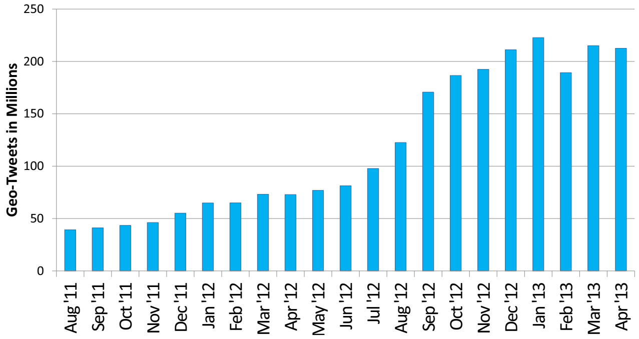

Figure 1. Global number of Tweets with location information between August 2011 and April 2013

In this work we present and compare two methods suitable to identify probable locations of origin for messages where such metadata is not provided by the user. The first method is based on extracting spatial language patterns from a very large dataset of collected geolocated messages. Large scale data aggregation is used to generate a priori probabilities that a given combination of terms might have been composed at a given coordinate to find the highest density peak. Based on a very large dictionary of terms frequently used in Twitter, we create and persistently store a spatial density map for each term, such that it can later be used for fast location estimation. The details of this strategy as well as the design of our reference implementation are explained in Section III. Our second strategy is based on the analysis of historic data about the users movement behavior and will be discussed in Section IV. The method uses a simple cluster analysis approach in order to identify often frequented places, facilities or home locations and determine the most probable whereabouts of a user writing a message with disabled location-features. In contrast to existing approaches our location estimation is based on large scale data and our algorithms adhere to a scalable design in order to handle

[1] http://articles.washingtonpost.com/2013-03-21/business/37889387_1_tweetsjackdorseytwitter

978-1-4799-2504-9/14 $31.00 © 2014 IEEE
DOI 10.1109/HICSS.2014.189

such data. Furthermore, our estimation strategies support uncertainty metrics and are thus specifically suited to be used in Visual Analytics [11] environments.

We evaluated the precision and effectiveness of our methods based on a held-out set of geolocated messages and present the results in Section V, where we also compare the strengths and weaknesses of the methods and present an application scenario. Section VI will briefly discuss the implications of our approach and and conclude with a summary of the results. Final remarks and an outlook on future work are given in Section VII.

II. Related Work

Finding the most probable location of origin of a document has been a research issue long before social media became popular. In this section we discuss some of the most important approaches and studies that exist in the field. Most solutions have tried to predict the location based on geographic references, cultural specifics or differences in language. Wing & Baldridge [12] overlay the world with a regular grid and assign term usage probabilities to each cell based on existing document sets with known geolocations. They use a range of supervised methods like Kullback-Leibler and Naive Bayes to predict the most probable cell for a document with unknown location. Roller et al. [13] follow a similar approach but use a grid with adaptive resolution based on a k-d-splitting. Documents with known location of origin are assigned to grid coordinates and locations are predicted by finding the document set with smallest similarity distance. However, as documents are discretely distributed to the grid cells, the quality of the measure is decreasing when resolution is increased. In contrast, our Kernel Density based method approximates a continuous distribution and the quality can thus be increased with increasing resolution. A more traditional approach was pursued by Eisenstein et al. [14] that use a generative language model to find regional specifics in topics and language. They train their model based on 380k Twitter messages and can predict the correct US state of origin of an unknown message with 24% accuracy.

Cheng et al. [15] proposed a framework to predict the location of Twitter users. They apply maximum likelihood methods for term usage distributions on a per-city level based on 5 million recorded tweets. They achieve an accuracy of 51% in predicting a users location within a 100 mile radius. In contrast to our method they use location information from the users profile instead of the individual tweets. Also, they concentrate on words that are *local* in terms of low spatial variation, while we normalize by term usage volumes. We believe that terms with strong spatial focus will automatically dominate the words with weak spatial correlation when the estimation is based on a sufficiently large data set. Kinsella et al. [16] pursue a similar approach by establishing term-distributions on a per-district, -city, -state and -country level, but in contrast to Cheng they also use geo-tags directly from the messages. They primarily propose Kullback-Leibler Divergence and Query Likelihood methods for the prediction. As they are sampling the data from the Spritzer and Firehose stream of the Twitter API, just a very small portion (less than 1%) of the tweets they evaluated contains precise latitude/longitude coordinates. By contrast, our data is taken from the filter stream with geospatial query parameters, such that we receive more than 7 million messages with location information each day. Therefore, and in contrast to other works, we were able to build a continuous large scale probability distribution using Kernel Density Estimation (KDE) based on a high resolution adaptive grid. Furthermore, we present algorithms that are specifically designed to process such data volumes either in one batch or in real-time directly based on social media streams.

Instead of using the textual content of the message several approaches also try to infer the most probable user location based on user profile attributes, links within the social network, or the available location information within the user history. Pontes et al. [17] compared the suitability of these publicly available information types to infer the home-location of a user within multiple social networks, i.e., by using Google+ profile attributes and friends' locations, Foursquare *check-ins* and friends' location, as well as previously published geolocated Twitter messages. While the Twitter dataset provided the highest accuracy, it also had the least amount of eligible users which shared the needed data. Foursquare, being centered around location based interaction, had the most eligible users but reduced accuracies.

Hecht et al. [18] examined the quality of the users' self provided location information in their profile and have shown that while $2/3$ of the places were valid, almost half of the remaining users supplied invalid information instead of none at all. As an additional problem, available geocoding APIs supplied geolocations for over 80% of the actually invalid places, such as *In my dreams*.

III. Term Density Maps

The content of a message might already tell us a lot about its possible origins. If a tweet is written in French language there is a certain probability that it was written in France. If it additionally contains the word *Toulouse* we can assume an added probability that it was written in the city of Toulouse. But we should not be too confident about it. However, if the message contains the words *just visited* and *Pont Neuf* the probability that it originated from Toulouse can be considered quite high. In our approach we want to drive this idea further to determine, whether one can assign a numeric probability that a given term was used at a given geographic location. Furthermore, we combine the a priori probabilities to find out whether one can find probable

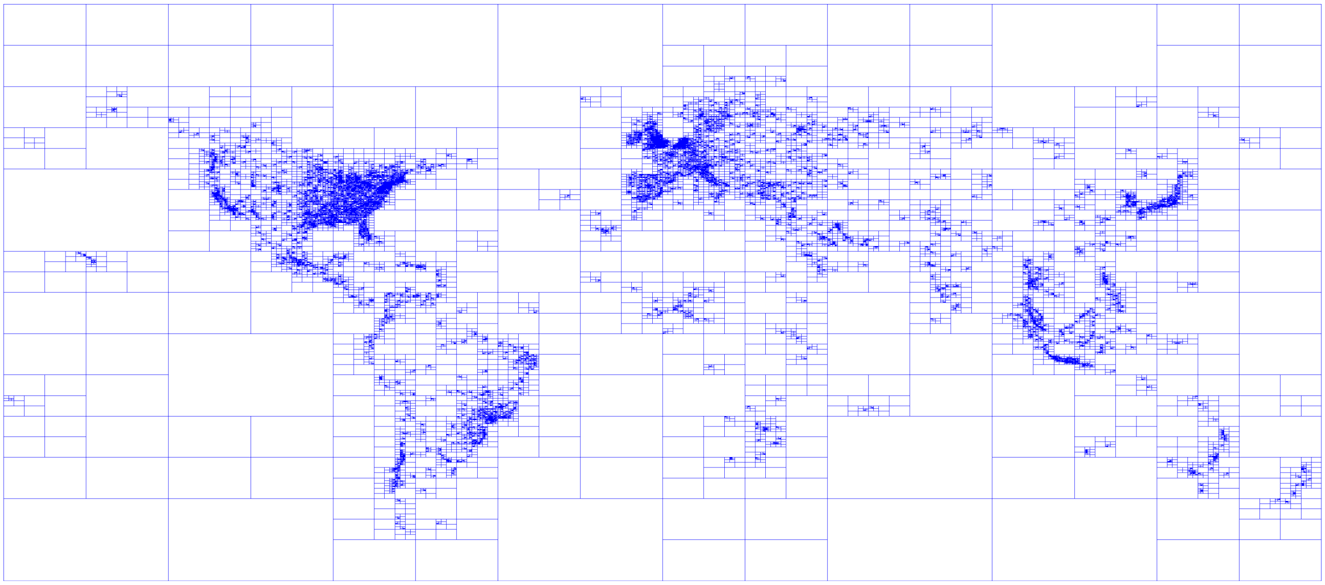

Figure 2. Result of adaptive grid creation based on 10 days of Twitter usage. Areas more densely populated by active Twitter users receive higher resolutions.

locations of origin for a given combination of words. Based on one year of collected geolocated Twitter data we present a solution that calculates such probabilities in a high resolution on a global scale.

The prerequisite for this approach is the availability of an average usage characteristic for each term at each geolocation. For this, our solution builds upon the smooth measure for location-dependent term irregularities defined in our previous work [19]. The measure was developed to allow the fast retrieval of continuous and localized inverse document frequencies for computing location-dependent *term frequency inverse document frequency* (tf-idf) values of messages. We adopt the smooth measure approach to compute term density estimations, combine the densities of individual terms of a message, and finally identify the most probable locations based on this combined density.

A. Term Density Estimation

If computing power would not be an issue, one could apply KDE [20], [21] in order to assign term usage probabilities to arbitrary geographic locations. Let $M_t = \{m_1, \ldots, m_n\}$ be a set of known messages that contain term t. The term usage probability at coordinate $x \in \mathbb{R}^n$ can be estimated by the density function

$$td_t(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{d(x, m_i)}{h}\right)$$

Here $d(x, m)$ is a distance function, e.g. Euclidean distance[2], and K is a kernel function, which is used in combination with a fixed bandwidth h to compose a smooth function. A common choice for K is the standard Gaussian

$$K(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right)$$

which will lead to rapidly decreasing weights with greater distance of the sample. Thus, if we are looking at an area where many messages contained t, td_t will assign a high probability that another message with t could have been written in that neighborhood. If there are just few messages in the vicinity, the estimated probability will be rather low.

If we want to find the most probable location of origin for some message m containing terms $t_1, \ldots, t_k$, a straightforward solution would be to sample term densities on a high resolution grid in order to find the location $\hat{x}$ that maximizes the independent combined probability under a naïve Bayes assumption:

$$\hat{x} = \max_{x} \prod_{i=1}^{k} td_{t_i}(x)$$

However, in order to yield accurate results the density estimation should be based on a very large dataset, such that actual term usage distributions of the microblogging service are properly reflected. For our evaluation (Section V) we applied the density estimation to a nearly complete set of geolocated Twitter messages written between August 2011 and August 2012, comprising more than 700 million messages. If we sample the term density on a global uniform grid that has at least a 0.5 kilometer precision, a resolution of 80 000 × 40 000 is needed and we would thus have to compute and accumulate more than 2.24×10^{18} weighted distances.

[2] In our case, the haversine formula is appropriate because of the spherical coordinates.

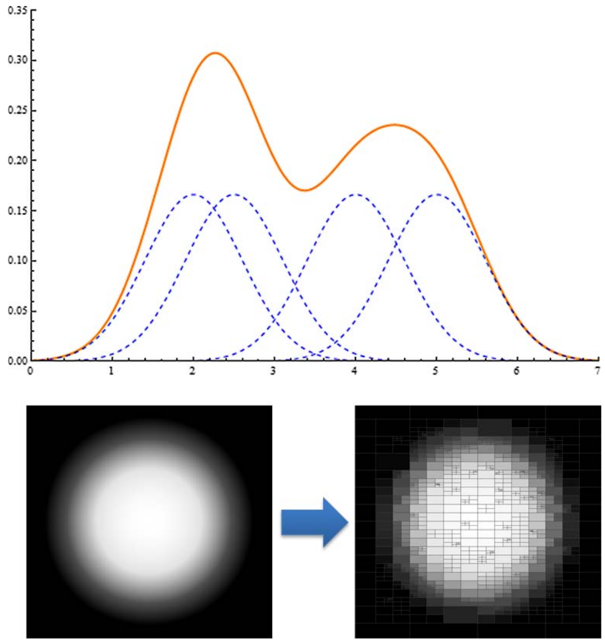

Figure 3. Top: Gaussian splats are placed and accumulated to approximate KDE. Bottom: Based on the adaptive grid, the splats are discretized and stored as density maps for each term.

This would not be feasible in a reasonable amount of time. Therefore, we apply three computation steps to dramatically speed up the process and allow fast location estimation for individual messages: constructing an adaptive grid, splatting, and peak detection.

B. Adaptive Grid

Usually we can expect messages to originate from densely populated areas. Assuming that the Twitter population is adequately represented in our collected dataset, we can apply a Quadtree algorithm to create an adaptive grid with higher resolutions where many Twitter users live and lower resolutions where the occurrence of messages is rather improbable. Starting with a single cell that comprises the complete world map, we split the cell into four subcells and evaluate for each cell whether the total amount of tweets written in that cell exceeds some fixed threshold. If the threshold is exceeded and the desired maximum resolution is not yet reached, we recursively apply the process for each cell again. Otherwise the cell becomes part of the final grid. In our case a good threshold was empirically determined to be 50 messages. A smaller threshold would lead to many grid cells and thus only a small improvement over a fixed grid, while a larger threshold would generate more cell artifacts during splatting (see rectangles in Figure 4). The result of this technique can be observed in Figure 2 and has roughly 236k cells, which reduces the effort by 4 orders of magnitude.

C. Splatting

Although the adaptive grid dramatically reduces the number of required sampling points, we would still need to iterate several times through the complete message set in order to compute $td_t(x)$ for just one term t. By accepting a slight error, the KDE can be approximated using a splatting scheme as described in [22]. Since our Gaussian kernel $K(u)$ almost descends to zero for $|u > 3 * h|$, we can invert the sampling process for a given term t. Instead of iterating through the complete message set for any given sample point x, we iterate through the message set just once and apply a Gaussian splat to the grid for every message that contains t. This is illustrated in Figure 3.

The splats for all messages containing t are then accumulated to form a density landscape, that can be permanently stored for later usage during the location estimation. Therefore, we can create and store a fixed density map $TD_t(id)$, mapping leaf-node identifiers (id) of the Quadtree structure to density values, for every term t in a large dictionary. For our approach, we created density maps for all terms that were used in Twitter at least 1000 times in the course of one year, totaling to about 130k terms. We store each density map in a single file that maps Quadtree leaf-node *id*s to density values. Two visualized examples of such density maps for the terms *love* and *amore* can be observed in Figure 4.

D. Peak Detection

Based on the precomputed term density maps, we detect areas where a message most likely originated from. For a given message m, the combined density map is built by accumulating the available density maps of all terms contained in m such that areas with high probability for specific term combinations receive an increased density value (see Figure 5). Based on the application scenario, and if it requires a higher recall or a higher precision, it might be valid to find the j most distinguished peaks $p_{1...j} \in X$ in the combined density map. Therefore, the map is scanned for the highest density values. In order to identify separated peaks and avoid the large values in close vicinity of a peak, which might refer to the same local data burst, we introduce an inhibition zone. In this zone around a peak, other values are ignored. This is realized through sorting the map cells by their density value and descending the list until j peaks have been found outside of inhibition zones of larger peaks.

As each of the combined density map constitutes a confidence landscape, the values of the resulting combined density map can be used to assign a confidence score for each peak in the landscape. This can be utilized to sort the peaks and compare the confidence of the location estimations against each other.

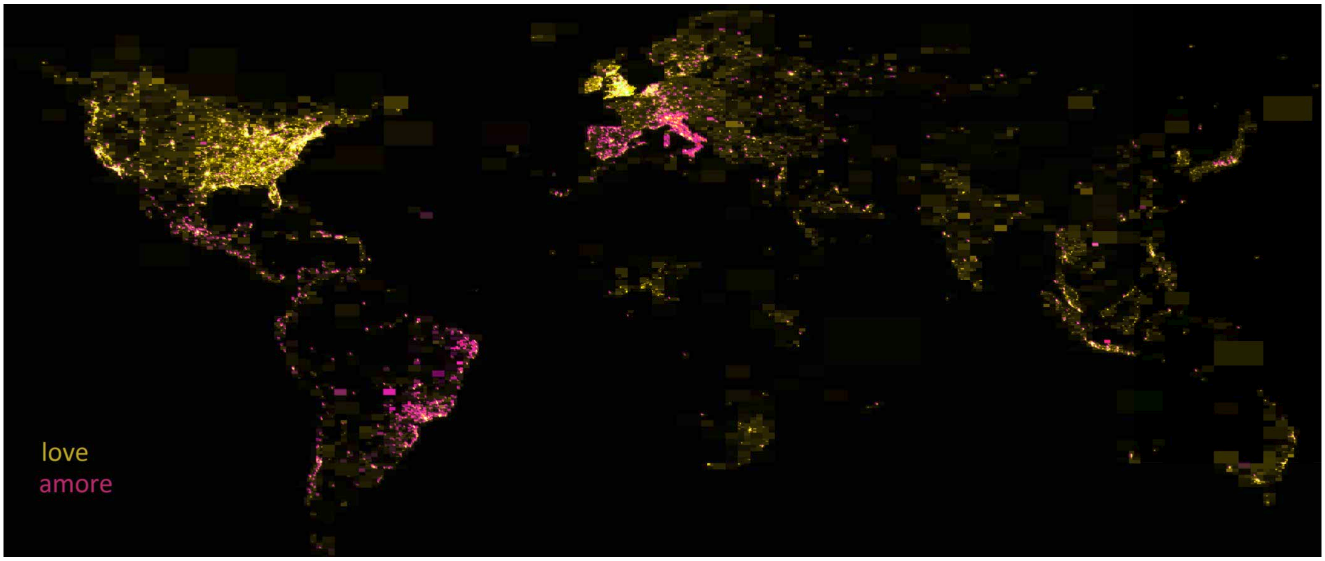

Figure 4. Term density maps for the terms *love* (yellow) and *amore* (magenta) created from one year of Twitter usage. The values are normalized per term to illustrate the difference.

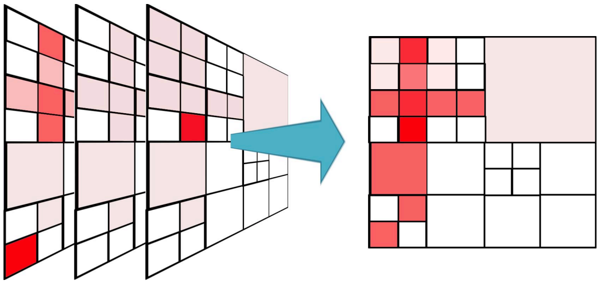

Figure 5. Accumulation of term density maps to allow peak detection. As the same grid structure is used for every term, values can be calculated separately for each cell.

IV. User-History based Estimation

Sometimes, the textual content of a message will tell us little to nothing about its possible origins. In such situations one can still use provided metadata in order to estimate a probable location. In most cases a user identifier or user screenname will be provided and one can thus record a message history for known Twitter users. Recent research has shown that most Twitter users will rarely write messages outside their hometown or most frequently visited locations. Furthermore there are a lot of users that have location features disabled most of the time but still have written at least a few messages with the feature enabled. In an analysis situation the analyst would often be interested in a particular map area and get as much information related to some event that happened in that area. Therefore, in addition to messages that already have a geo-tag or GPS position, he would also be interested in messages that were written recently by users living in that area.

A. Location Clusters

In our second approach to location estimation we create a large database of known users from the geolocated dataset. For each user we try to find his home location as well as frequently visited other locations based on his history of recorded geocoordinates. Usually a user would not constantly move around and write tweets all the time but rather stay at some place for a period of time, write some tweets, and then move to the next place. This will result in clearly distinguishable clusters of messages with small pairwise distances due to minimal movements or GPS inaccuracy. Thus, to find such locations we can simply collect the set of all coordinates for a user and then apply a clustering scheme. If the user writes a new message with unknown location, we can assign it to any of his known locations by placing it at the centroid of the clusters. Furthermore, we can assign a confidence score to each of the clusters by relating their message count to the amount of all geolocated messages written by the user. Therefore our algorithm is specifically suited for systems that support the visualization of uncertainty and allow to focus of high recall.

B. Iterative Implementation

Our experiments showed that existing clustering solutions, as they can be found in data analysis systems like Weka[3] and Apache Mahout[4] are not suitable for the problem. As we would have to apply cluster analysis to several million individual sample sets, sometimes comprising thousands of items each, most hardware setups would need a very long

[3] http://www.cs.waikato.ac.nz/ml/weka/

[4] http://mahout.apache.org

time – i.e. several days or weeks – to process the data. However, in contrast to other cluster analysis challenges, our datasets usually have a very simple overall structure as a result of the aforementioned typical usage behavior of Twitter users. Most often there will be hardly any noise between clusters and the clusters will be easy to identify and distinguish from one another. We can therefore apply the algorithm shown in Listing 1 iteratively to each message m in a given message set to identify the clusters.

```
procedure add_message(m) is
begin
  $user \leftarrow m.getUserID()$
  $p \leftarrow m.getPosition()$
  $\hat{c} \leftarrow \arg\min_{c \in clu[user]} \|p - c.getCentroid()\|$
  if $\hat{c} = null$ OR $\|p - \hat{c}\| > k_1$ then
    $\hat{n} \leftarrow newCluster()$
    $\hat{n}.setCentroid(p)$
    $\hat{n}.setXSum(p.x)$
    $\hat{n}.setYSum(p.y)$
    $\hat{n}.setSize(1)$
    $clu[user] \leftarrow clu[user] \cup \{\hat{n}\}$
  else
    $\hat{c}.setCentroid((\frac{c.getXSum()+p.x}{c.getSize()+1}, \frac{c.getYSum()+p.y}{c.getSize()+1}))$
    $\hat{c}.setXSum(c.getXSum() + p.x)$
    $\hat{c}.setYSum(c.getYSum() + p.y)$
    $\hat{c}.setSize(c.getSize() + 1)$
  end if
  $count{+}{+}$
  if $count \ \% \ k_2 = 0$ then
    for each $c \in clu[user]$ do
      if $c.getSize() < 2$ then
        $clu[user] \leftarrow clu[user] - \{c\}$
      end if
    end for
  end if
end
```

Listing 1. Cluster Detection

In this algorithm $clu[user]$ represents an initially empty hashmap, mapping each user identifier to a set of location clusters identified for the user. Furthermore we use two fixed thresholds: k_1 is a distance limit that defines whether m can be assigned to its nearest cluster or whether it should establish a new cluster. In our case we set k_1 to about the size of a city district, thus everything within that radius is assigned to the same cluster. The second threshold k_2 defines the frequency of a clean-up step which eliminates clusters with very low member counts. This step avoids filling the global space with noise clusters that would lead to reduced performance and increased memory consumption. If memory constraints are not an issue one can choose a very high k_2, i.e., low clean-up frequency. In our case we used $k_2 = 800$. A sample result of detected location clusters for a user living in England can be seen in Figure 6. The detected clusters and their respective message counts are shown as yellow squares and green labels. The white dots resemble individual recorded locations for the user. The cluster analysis indicates that the main location for the observed user seems to be in the northern part of Southend-on-Sea, Essex, where about 87 messages were aggregated within a cluster. Several smaller clusters indicate visits to London and other towns in the area.

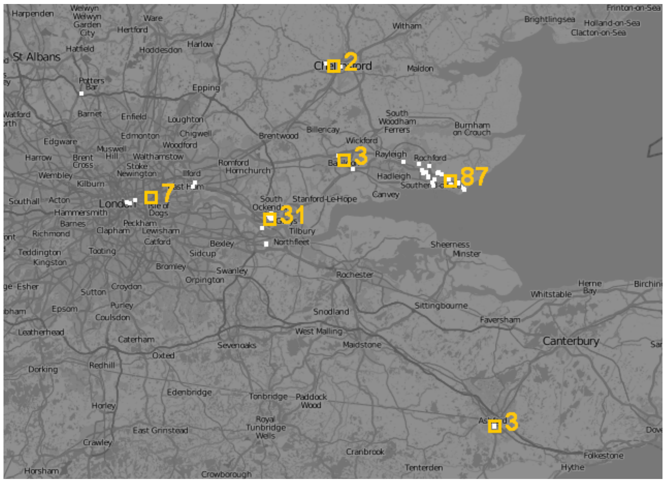

Figure 6. Detected location clusters for a user supposedly living near London. One can observe that the largest detected cluster (containing 87 messages) is located in Southend-on-Sea, Essex.

V. Evaluation

To train and evaluate our models we collected data from the Twitter Streaming API using the statuses/filter stream[5]. Although rate limitations apply for the number of Tweets that can be collected by the stream, our query strategy allows us to capture almost all worldwide messages having either textual geo-tags or precise geocoordinates[6]. We collected approximately 1.3 billion geolocated tweets between 8 August 2011 and 12 August 2012. From this data, we took the subset of 0.7 billion messages that had actual latitude/longitude coordinates to train both of our models. On October 2012, about two months after the time frame of the training data, we took another sample of 10 000 geolocated messages randomly distributed all over the world and about 190 000 messages from selected geographic areas collected within a 24 hour timeframe. In an actual application case the term density and user history data would presumably be generated only once and then used for a very long time. We therefore chose this large temporal gap between training and sample data in order to prevent positive bias from ignoring the time-dependent changes in the data (e.g. changing user base, prevalent topic variation, etc.). The geographic distribution of the training set and the randomly chosen test set can be seen in Table I. The dataset was cleaned from automatically generated tweets from services like Foursquare. These messages always contain precise coordinates and as such would not have to be processed

[5] https://dev.twitter.com/docs/api/1.1/post/statuses/filter

[6] The Twitter API automatically sends notifications when rate limitation applies. Thus we can estimate our miss-rate to be less than 5%

	Train	Test
total messages	735838811	10000
United States	36.7%	35.0%
Indonesia	7.9%	5.9%
United Kingdom	7.4%	7.2%
Brazil	6.9%	7.2%
Malaysia	3.3%	3.4%
Mexico	3.1%	2.8%
Japan	3.0%	2.4%
Spain	2.9%	4.3%
Turkey	2.7%	3.9%
Netherlands	2.3%	1.8%
Russia	1.7%	1.4%
France	1.7%	2.4%
Philippines	1.6%	1.5%
Chile	1.5%	1.1%
Canada	1.4%	1.3%

Table I
GEOSPATIAL DISTRIBUTION OF THE TRAINING AND TEST SET AS A LIST OF THE FIFTEEN MOST FREQUENT COUNTRIES.

in actual application scenarios. Additionally, they usually contain toponyms that would allow for an easy localization which would distort our results presented in the following sections.

A. *Accuracy: Term Density*

For our first evaluation we identified only the highest overall density for the merged term density maps. Often there is a single term in the message that dominates all others, as it has just one extremely large peak. This is most often the case for location names, like *Bombay* or *Rushmore*, and such messages have, of course, the highest chance to be correctly assigned to their actual location. For each sample we calculated the distance between the actual location and the estimated location. The solid red curve in Figure 7 shows the result of this first test with the 10k random sample set and error distances within 100 kilometers. The diagram shows the error distance from the actual location versus the fraction of estimations that had this or a smaller error. The method achieves an accuracy of about 6% within the first 5 kilometers from the target, which is approximately the size of two larger Zip-Code areas. The average radius of large cities like London and Berlin is about 20 kilometers, where the accuracy increases to 13.4% In an analysis scenario where recall is more important than precision one could consider multiple possible locations for each message by increasing the number of peaks that should be considered for an observed area. Therefore, we introduced an inhibition zone during peak detection in order to generate a selection of possible locations for the target that are at least 100 kilometers apart. The dotted and dashed curves in Figure 7 show accuracy rates when the true location is among the, e.g., 5, 9 or 13 most probable locations for the same target. This way the accuracy is increased to about 6% for the 2 kilometers range and to almost 30% for the 20 kilometers range.

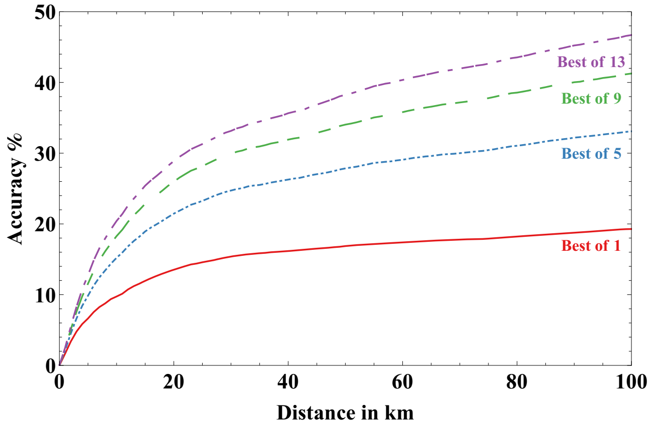

Figure 7. Term Densities: Error distance versus accuracy based on the 10k sample messages. The curves show the performance for using the best of 1, 5, 9 and 13 guesses (=peaks) for a given message.

B. *Accuracy: User Histories*

Based on the same training and sample data set as in the term density evaluation we also evaluated location guessing based on the users' location clusters. From the training data set we extracted profiles for 9 134 562 individual users. On average 80.94 messages were written and 2.16 clusters were detected per user, which is a surprisingly low number and manual investigation confirmed that there is indeed a large amount of users writing from less than 3 locations (e.g. home and office).

The results for accuracy within the first 100 kilometers from the target are depicted as the lower solid red curve in Figure 8. Here we can see an accuracy of already 30.1% for the 2 kilometer range and 48.5% for the 20 kilometer range.

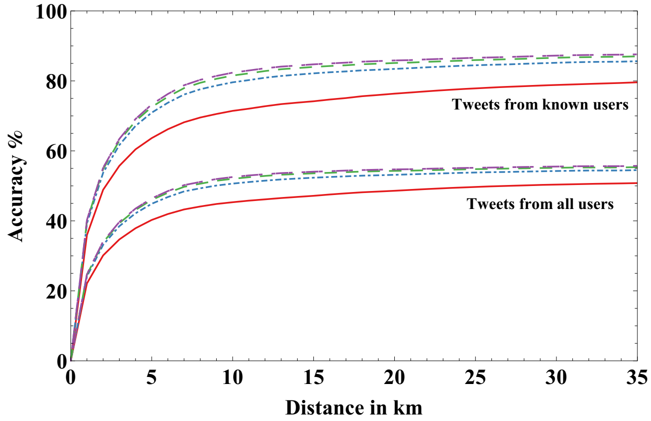

Figure 8. User History: Error distance versus accuracy based on the 10k sample messages. The lower curves show the performance for best of 1, 2, 3 and 4 guesses (=clusters) if all sample messages are counted. The upper curves show the performance if only messages from previously known users are counted.

The dashed curves directly above the lower solid red curve illustrate the accuracy for counting the best of 2, 3 and 4 estimations. With the average of 2.16 clusters per user, this, of course, yields no significantly better results.

It came as an interesting observation that from the 10k sample set, only about 60% were known users from the one year training set—i.e. about 4000 messages came from users that used Twitter and/or the location features for the first time since at least 14 months. Thus, if we count only the location guesses for known users, the accuracy can even be raised to about 80% for the 20 kilometer range, which is depicted by the curves in the upper part of Figure 8.

C. Comparison

From an examination of successfully placed evaluation samples, we learned that only in few cases specific combinations of terms helped to indicate the right location in the term density approach. For example, the term occurrences for *mirage*, *poker* and *room* are distributed all over the world, but the peak of the combined maps is in Las Vegas, which was the correct guess for the message *Im playing at mirage poker room w drunken tourists ftw!*. More common are tweets that already contain one specific term that is frequently used at just one location like toponyms or names of returning events. However, in such cases the true location could also be found using geocoding services like Geonames or Yahoo Placemaker.

Our evaluation indicated that user history based estimation by far outperforms the language based approach (see Table II), which does not come as a big surprise. But the method has several drawbacks including the expectable decrease in accuracy when confronted with users that never use the location feature or when users report important events apart from their usual locations. These are the cases where the term density approach can lead to better results, as it is independent from the behavior of individual users.

	TD	All User	Known User
5 km	5.9	37.9	60.4
10 km	9.4	44.8	70.5
50 km	16.8	51.8	81.1
100 km	19.3	53.6	84.1
1000 km	44.8	59.9	94.3
8000 km	82.0	62.6	98.9

Table II
COMPARISON OF RESULTS (BEST OF 1). COLUMNS SHOW THE ACCURACY VS. ERROR DISTANCE VALUES FOR THE TERM DENSITY (TD) AND USER HISTORY BASED APPROACHES.

Based on these results, a combination of both approaches can easily be introduced. The combination applies the user cluster based estimation if the user is already known by the training data and the term density approach otherwise.

D. Application Scenario

In an actual application scenario, the analyst is interested in a certain geographic location and retrieves a set of messages about the location using the location estimation approaches. Here, it is also important to know how often messages are falsely located to an area of interest, thus reducing the precision of the approach and leading to larger datasets containing more irrelevant data. However, for crisis management and similar applications, where an analyst has to retrieve as much eyewitness information as possible, the analyst requires a high recall approach. We evaluate the suitability of our approach for this scenario by calculating precision and recall scores for a range of geographic areas. After collecting all relevant messages of one day for these locations, based on the true location of the message, we tested what percentage of messages the analyst would have retrieve by our combined approach (recall). Based on a sample from all messages, we evaluated how many messages would be falsely retrieved for the given area to estimate the percentage of relevant messages in the whole retrieval set (precision).

	Test	Recall	Prec.
Total	4 035 857		
Florida	80432	54%	71%
London	46696	61%	25%
Paris	27519	73%	38%
Moscow	17683	74%	35%
Manhattan	12348	37%	27%
Berlin	2668	66%	29%
Mumbai	2374	55%	85%
Arr. l'Hôtel de Ville	355	17%	12%
Sunset Dist.	234	23%	20%

Table III
TABLE SHOWS NUMBER OF ACTUAL TWEETS IN THE AREA, RECALL AND PRECISION.

Our results (Table III) indicate good recall rates above 50% for most areas on the city level, reaching more than 70% for cities where English is not a dominating language. However, the recall on the district level, e.g. Arrondissement de l'Hôtel de Ville (Paris) and Sunset District (San Francisco) is relatively low. Overall the precision is often below 50% and thus depending on the area the analyst would sill have to cope with many falsely located messages alongside the correct ones. Nevertheless, the overall amount of messages he has to investigate to receive probable eyewitness accounts is still dramatically reduced to an amount that could be handled in a reasonable amount of time.

VI. Discussion and Conclusion

The main purpose of this paper is to give recommendations to researchers and practitioners challenged by the same problem that we had when analyzing social media in order to gain situational awareness. In past events people provided essential first-hand information that already helped to coordinate relief efforts during major disasters like the 2010 Haiti earthquake or hurricane Sandy. In such events provided location information can play a critical role in separating trustworthy eyewitness accounts from general social media chatter, rumors and uncertain second hand information. Although location data in Twitter can be faked, such acts require a certain malicious effort and by checking tweets from other users in the area, one can often identify false accounts. Unfortunately only a small fraction of messages are provided with the needed metadata and the vast majority remains unexploited.

In this paper we therefore presented and investigated two approaches for social media location discovery using large scale aggregated knowledge from a massive historical tweet dataset to enlarge the set of geolocated messages. With the strategies presented in this paper, analysis systems can increase the recall rate when an area of interest is already defined. Additionally, the approach provides uncertainty scores to support the analyst's decisions making.

We are, however, well aware of the severe privacy concerns that can be raised when government agencies, companies, or other institutions use such technology to infer data about individual users' whereabouts that they explicitly did not share. Recent history has shown that pariah states have exploited social media data to turn it against civil rights movements and that even intelligence agencies of developed countries have collected and analyzed such data to spy on citizens. We hope that the presentation of our approach will also increase public awareness and help people to understand the dangers of sharing too much information.

The large communities of location-enabled social media networks have generated a unique dataset mapping language and content to geographic coordinates, thereby forming a digital sociocultural landscape of unprecedented richness and extent. Because of the high value of the provided data, social media mining has become an important topic in many different research areas ranging from practical applications of traditional NLP methods over computational sociology to stock market analysis and we are sure to see a lot of interesting developments in the near future.

VII. Future Work

Mobile devices and the capabilities of service providers bring their users to develop a very specific language and different means of communication. While our data-driven approach already addresses some of these aspect, future work will take on the temporal dimension of this fast-paced domain by considering seasonality and trends. Additionally, interactive visual means to assess the uncertainty of the approach's results are needed for a better understanding of which types of messages are most promising for term-usage based geocoding. Finally, we are confident that our research is also applicable to other social media as well as text data coming from completely different areas and will further investigate its usage scenarios.

Acknowledgment

This work was supported in part by the German Federal Ministry of Education and Research (BMBF) in context of the VASA project and by the cooperative graduate program 'Digital Media' of the University of Stuttgart, the University of Tübingen, and the Stuttgart Media University (HdM).

References

[1] C. Chew and G. Eysenbach, "Pandemics in the age of Twitter: Content analysis of tweets during the 2009 H1N1 outbreak," *PLoS One*, vol. 5, no. 11, 2010.

[2] T. Heverin and L. Zach, "Microblogging for crisis communication: Examination of Twitter use in response to a 2009 violent crisis in seattle-tacoma, washington area," in *Proc. 7th Intl. ISCRAM Conf.*, 2010.

[3] A. Hughes and L. Palen, "Twitter adoption and use in mass convergence and emergency events," *Intl. J. Emergency Management*, vol. 6, no. 3, pp. 248–260, 2009.

[4] M. Mendoza, B. Poblete, and C. Castillo, "Twitter Under Crisis: Can we trust what we RT?" in *Proc. 1st WS Social Media Analytics*. ACM, 2010, pp. 71–79.

[5] L. Palen, S. Vieweg, S. Liu, and A. Hughes, "Crisis in a networked world: features of computer-mediated communication in the april 16, 2007, virginia tech event," *Social Science Computer Review*, 2009.

[6] T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake shakes Twitter users: real-time event detection by social sensors," in *Proc. 19th Intl. Conf. World Wide Web*. ACM, 2010, pp. 851–860.

[7] K. Starbird and L. Palen, "Pass it on?: Retweeting in mass emergency," in *Proc. 7th Intl. ISCRAM Conf.*, 2010.

[8] A. MacEachren, A. Jaiswal, A. C. Robinson, S. Pezanowski, A. Savelyev, P. Mitra, X. Zhang, and J. Blanford, "SensePlace2: GeoTwitter analytics support for situational awareness," in *IEEE Conf. Visual Analytics Science and Technology*, Providence, RI, 2011. [Online]. Available: http://www.geovista.psu.edu/publications/2011/MacEachren_VAST_2011_reducedsize.pdf

[9] A. Marcus, M. Bernstein, O. Badar, D. Karger, S. Madden, and R. Miller, "TwitInfo: Aggregating and visualizing microblogs for event exploration," in *Proc. Conf. Human Factors in Computing Systems*. ACM, 2011, pp. 227–236.

[10] D. Thom, H. Bosch, S. Koch, M. Wörner, and T. Ertl, "Spatiotemporal anomaly detection through visual analysis of geolocated Twitter messages," in *IEEE Pacific Visualization Symp.*, 2012.

[11] J. J. Thomas and K. A. Cook, *Illuminating the Path: The Research and Development Agenda for Visual Analytics*. National Visualization and Analytics Ctr, 2005. [Online]. Available: http://www.worldcat.org/isbn/0769523234

[12] B. P. Wing and J. Baldridge, "Simple supervised document geolocation with geodesic grids," in *Proc. 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, ser. HLT '11. Stroudsburg, PA, USA: Association for Computational Linguistics, 2011, pp. 955–964. [Online]. Available: http://dl.acm.org/citation.cfm?id=2002472.2002593

[13] S. Roller, M. Speriosu, S. Rallapalli, B. Wing, and J. Baldridge, "Supervised text-based geolocation using language models on an adaptive grid," in *Proc. 2012 Joint Conf. Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 2012, pp. 1500–1510.

[14] J. Eisenstein, B. O'Connor, N. A. Smith, and E. P. Xing, "A latent variable model for geographic lexical variation," in *Proc. 2010 Conf. Empirical Methods in Natural Language Processing*, ser. EMNLP '10. Stroudsburg, PA, USA: Association for Computational Linguistics, 2010, pp. 1277–1287. [Online]. Available: http://dl.acm.org/citation.cfm?id=1870658.1870782

[15] Z. Cheng, J. Caverlee, and K. Lee, "You are where you tweet: a content-based approach to geo-locating twitter users," in *Proc. 19th ACM Intl. Conf. Information and Knowledge Management*. ACM, 2010, pp. 759–768.

[16] S. Kinsella, V. Murdock, and N. O'Hare, "I'm eating a sandwich in glasgow: modeling locations with tweets," in *Proc. 3rd Intl. WS on Search and Mining User-generated Contents*. ACM, 2011, pp. 61–68.

[17] T. Pontes, G. Magno, M. Vasconcelos, A. Gupta, J. Almeida, P. Kumaraguru, and V. Almeida, "Beware of What You Share: Inferring Home Location in Social Networks," in *IEEE 12th Intl. Conf. Data Mining Workshops (ICDMW)*, 2012, pp. 571–578.

[18] B. Hecht, L. Hong, B. Suh, and E. H. Chi, "Tweets from justin bieber's heart: the dynamics of the location field in user profiles," in *Proc. SIGCHI Conf. Human Factors in Computing Systems*, ser. CHI '11. New York, NY, USA: ACM, 2011, pp. 237–246.

[19] D. Thom, H. Bosch, and T. Ertl, "Inverse document density: A smooth measure for location-dependent term irregularities," in *Proc. COLING 2012*, T. C. . O. Committee, Ed., 2012, pp. 2603–2618.

[20] M. Rosenblatt, "Remarks on some nonparametric estimates of a density function," *The Annals of Mathematical Statistics*, vol. 27, no. 3, pp. 832–837, Sep. 1956.

[21] E. Parzen, "On Estimation of a Probability Density Function and Mode," *The Annals of Mathematical Statistics*, vol. 33, no. 3, pp. 1065–1076, 1962.

[22] L. A. Westover, "Splatting: a parallel, feed-forward volume rendering algorithm," Ph.D. dissertation, University of North Carolina at Chapel Hill, NC, USA, 1991, uMI Order No. GAX92-08005.