**LITERATURE SURVEY**

**Articles:**

## 1) Geolocation Prediction in Social Media Data by Finding Location Indicative Words

**AUTHORS: Han, Bo & Cook, Paul & Baldwin, Timothy**

# Geolocation prediction is vital to geospatial applications like localised search and local event detection. Predominately, social media geolocation models are based on full text data, including common words with no geospatial dimension (e.g. today) and noisy strings (tmrw), potentially hampering prediction and leading to slower/more memory-intensive models. In this paper, we focus on finding location indicative words (LIWs) via feature selection, and establishing whether the reduced feature set boosts geolocation accuracy. Our results show that an information gain ratiobased approach surpasses other methods at LIW selection, outperforming state-of-the-art geolocation prediction methods by 10.6% in accuracy and reducing the mean and median of prediction error distance by 45km and 209km, respectively, on a public dataset. We further formulate notions of prediction confidence, and demonstrate that performance is even higher in cases where our model is more confident, striking a trade-off between accuracy and coverage. Finally, the identified LIWs reveal regional language differences, which could be potentially useful for lexicographers.

# 2) Where Are You Settling Down: Geo-locating Twitter Users Based on Tweets and Social Networks

**AUTHORS:**  **Shaowu ZhangHongfei Lin**

In this paper, we investigate the advantages of taking two dimensions of tweet content and social relationships to construct models for predicting where people settle down as their profiles reveal city- and town-level data. Based on the users who voluntarily reveal their locations in their profiles, we propose two local word filters - Inverse Location Frequency (ILF) and Remote Words (RW) filter - to identify local words in tweets content. We also extract separately the place name mentioned in tweets using the Named Entity Recognition application and then filter them by computing the city distance. We consider users’ friends and 2-hop of followings. In our experiment, we finally combine these two dimensions to estimate user location and achieve an Accuracy of 56.6% within 100 miles in city-level and 45.2% within 25 miles in town-level of their actual location which outperforms the single dimension prediction and the baseline. Twitter’s open and succinct service allows it to gather vast amounts of data and updates by users who come from different places. The user always inadvertently leaks some dialect words and place names of his/her residence in the process of adding updates. Understanding the geographic features of those update statuses enables the system to push better local advertising, highlight points of interest, show local news, create recommendations for friends living in the vicinity, and even help search engines understand users’ search intentions better. In this paper we build textual models of local words and place names based on pure tweets to estimate a user’s place of residence, even when the user does not explicitly reveal the place name, or his/her geographic coordinates in the profile.

# 3) Multiple Location Profiling for Users and Relationships from Social Network and Content

**AUTHORS** **: Li, Rui & Wang, Shengjie & Chen-Chuan Chang, Kevin**

Users’ locations are important for many applications such as personalized search and localized content delivery. In this paper, we study the problem of profiling Twitter users’ locations with their following network and tweets. We propose a multiple location profiling model (MLP), which has three key features: 1) it formally models how likely a user ollows another user given their locations and how likely a user tweets a venue given his location, 2) it fundamentally captures that a user has multiple locations and his following relationships and tweeted venues can be related to any of his locations, and some of them are even noisy, and 3) it novelly utilizes the home locations of some users as partial supervision. As a result, MLP not only discovers users’ locations accurately and completely, but also “explains” each following relationship by revealing users’ true locations in the relationship. Experiments on a large-scale data set demonstrate those advantages. Particularly, 1) for predicting users’ home locations, MLP successfully places 62% users and outperforms two state-of-the-art methods by 10% in accuracy, 2) for discovering users’ multiple locations, MLP improves the baseline methods by 14% in recall, and 3) for explaining following relationships, MLP achieves 57% accuracy. In the literature, many methods [8, 5, 11] have been proposed to profile users’ locations in the context of social network. Specifically, they focus on profiling a user’s home location, which is the single ”permanent” resident location of the user, by exploring her social network (e.g., friendships) and content (e.g., tweets). Intuitively, both types of data provide valuable signals for profiling users’ locations, as a user is likely to 1) connect to others living close to her, and 2) tweet her nearby “venues”.

# 4) A multi-indicator approach for geolocalization of tweets

**AUTHORS** : **A. Schulz, A. Hadjakos, H. Paulheim, J. Nachtwey, and M. M¨**

**uhlh**

Real-time information from microblogs like Twitter is useful for different applications such as market research, opinion mining, and crisis management. For many of those messages, location information is required to derive useful insights. Today, however, only around 1% of all tweets are explicitly geotagged. We propose the first multi-indicator method for determining (1) the location where a tweet was created as well as (2) the location of the user’s residence. Our method is based on various weighted indicators, including the names of places that appear in the text message, dedicated location entries, and additional information from the user profile. An evaluation shows that our method is capable of locating 92% of all tweets with a median accuracy of below 30km, as well as predicting the user’s residence with a median accuracy of below 5.1km. With that level of accuracy, our approach significantly outperforms existing work. Twitter has become a very popular microblogging platform during the last years with more than 400 million tweets created per day.1 Research has shown that tweets provide valuable real-time information, e.g., for opinion analysis preceding political elections (Tumasjan et al. 2010), for regional health monitoring (Aramaki 2011), or local emergency detection (Starbird et al. 2010). However, according to recent analyses (Hale and Gaffney 2012), only around 1% of all tweets are explicitly geotagged. Thus, without a possibility to predict the location of tweets, 99% of all tweets cannot be used for the above mentioned purposes.

**5. Collect Data From Twitter: A step by step implementation using tweepy**

**Author: Zoumana Keita**

Getting data comes as the second step in any data science/machine learning project lifecycle, right after framing the problem you want to solve, which would make this step be the backbone of the rest of the phases. Also, social media are great places to collect data, especially for competitor analysis, topic research, sentiment analysis, etc. This article aims to perform a step-by-step implementation on how to get credentials and the implementation on a simple use case.

**Books**

1. **Python for Geospatial Data Analysis**

**Author: Bonny P. McClain**

This book helps you:

* Understand the importance of applying spatial relationships in data science
* Select and apply data layering of both raster and vector graphics
* Apply location data to leverage spatial analytics
* Design informative and accurate maps
* Automate geographic data with Python scripts
* Explore Python packages for additional functionality
* Work with atypical data types such as polygons, shape files, and projections
* Understand the graphical syntax of spatial data science to stimulate curiosity

## You are Here: From the Compass to GPS, the History and Future of How We Find Ourselves

**Author: Hiawatha Bray**

Written by Hiawatha Bray in 2014, *You are Here* is a compelling read perfect for anyone interested in learning about the history of location technology. This book examines how having constant access to location data has come to be a part of modern life, and highlights the key players in the development of location and mapping technologies. Informative and interesting, You are Here discusses how solving one problem can sometimes usher in a new set of concerns.

**Research Paper**

1. **Uncovering the location of Twitter User**

**Authors** Renato Assunc¸ao, Gisele L. Pappa, Renato Miranda, Wagner Meira Jr.

Social networks, like Twitter and Facebook, are valuable sources to monitor real-time events, such as earthquakes and epidemics. For this type of surveillance the user’s location is an essential piece of information, but a substantial number of users choose not to disclose their geographical information. However, characteristics of the users’ behavior, such as the friends they associate with and the types of messages published can hint on their spatial location. In this paper, we present a method to infer the spatial location of Twitter users. Unlike the approaches presented so far, we incorporate two sources of information to learn the geographical position: the text posted by users and their friendship network. We propose a probabilistic approach that jointly models the geographical labels and the Twitter texts of the users organized in the form of a graph representing the friendship network. We use Markov random field probability model to represent the network and learning is carried out through a Markov chain Monte Carlo simulation technique to approximate the posterior probability distribution of the missing geographical labels. We demonstrate the utility of this model in a large dataset of Twitter users, where the ground truth is the location given by GPS dispositives. The method is evaluated and compared to two basline algorithms that user these two types of information separatdly, The accuracy rates obtained are significantly better than those of the baseline methods. Index Terms—Network Learning; Geographic Targeting; Geolocation Estimation; Spatial Data Mining.

**2. Decision Tree, Naïve Bayes and Support Vector Machine Applying on Social Media Usage in NYC / Comparative Analysis**

**Author: Ahmed Burhan Mohammed**

Data mining and classification are most research idea that used in many topics by researchers. This study presents the comparison of three algorithms for classifications such as (Decision Tree, Naïve Bayes and Support Vector Machine), applying for social media usage dataset by NYC, to get the best result of the classification algorithm that can classify the instances according to the platforms. The final result of this research refer to the Support Vector Machine returned the best result among these techniques

**3. Machine-Learning-Based User Position Prediction and Behavior Analysis for Location Services**

**Authors: Haiyang Jiang 1 , Mingshu He 2,\* , Yuanyuan Xi 3 and Jianqiu Zeng 1**

Machine learning (ML)-based methods are increasingly used in different fields of business to improve the quality and efficiency of services. The increasing amount of data and the development of artificial intelligence algorithms have improved the services provided to customers in shopping malls. Most new services are based on customers’ precise positioning in shopping malls, especially customer positioning within shops. We propose a novel method to accurately predict the specific shops in which customers are located in shopping malls. We use global positioning system (GPS) information provided by customers’ mobile terminals and WiFi information that completely covers the shopping mall. According to the prediction results, we learn some of the behavior preferences of users. We use these predicted customer locations to provide customers with more accurate services. Our training dataset is built using feature extraction and screening from some real customers’ transaction records in shopping malls. In order to prove the validity of the model, we also crosscheck our algorithm with a variety of machine learning algorithms. Our method achieves the best speed–accuracy trade-off and can accurately locate the shops in which customers are located in shopping malls in real time. Compared to other algorithms, the proposed model is more accurate. User preference behaviors can be used in applications to efficiently provide more tailored services.