Multi-stage Collaborative filtering for Tweet Geolocation

Keerti Banweer
University of Oklahoma
Norman, Oklahoma
keerti.banweer@ou.edu

Austin Graham
University of Oklahoma
Norman, Oklahoma
austin.graham@ou.edu

Joe Ripberger
University of Oklahoma
Norman, Oklahoma
jtr@ou.edu

Nina Cesare
University of Washington
Seattle, Washington
ninac2@uw.edu

Elaine Nsoesie
University of Washington
Seattle, Washington
en22@uw.edu

Christan Grant
University of Oklahoma
Norman, Oklahoma
cgrant@ou.edu

ABSTRACT

Data from social media platforms such as Twitter can be used to analyze severe weather reports and foodborne illness outbreaks. Government officials use online reports for early estimation of the impact of catastrophes and to aid resource distribution. For online reports to be useful they must be geotagged, but location is often not available. Less then one percent of users share their location information and/or acquisition of significant sample of geolocation messages is prohibitively expensive. In this paper, we propose a multi-stage iterative model based on the popular matrix factorization technique. This algorithm uses the partial information and exploits the relationship of messages, location, and keywords to recommend locations for non-geotagged messages. We present this model for geotagging messages using recommender systems and discussion the potential applications and next steps in this work.

CCS CONCEPTS

• **Information systems** → *Location based services*; *Collaborative filtering*;

KEYWORDS

Geolocation, Collaborative Filtering, Twitter

ACM Reference Format:
Keerti Banweer, Austin Graham, Joe Ripberger, Nina Cesare, Elaine Nsoesie, and Christan Grant. 2018. Multi-stage Collaborative filtering for Tweet Geolocation. In *2nd ACM SIGSPATIAL Workshop on Recommendations for Location-based Services and Social Networks (LocalRec'18), November 6, 2018, Seattle, WA, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3282825.3282831

1 INTRODUCTION

Location is an important asset in analyzing and acting on data generated during disasters. Researchers and government officials use geolocated messages to monitor the earth for all types of disasters [2, 6]. Targeted advertising, content recommendation, or trend analysis can all be linked to a target user's location. Unfortunately for such platforms, location information is expensive to purchase or not available. In 2010, less than 25 percent of Twitter users' location was known in a random sample of over 1 million users [3]. Since then, this number has dwindled down to less than one percent [9].

A given location may be tied to particular key words or phrases. For example, "Big Easy" can be associated with "New Orleans" with a high probability because this is the city's nickname. It is also common to travel to new locations and encounter unique sayings, such as "y'all" (meaning "you all") in southern portions of the United States [7]. Users linked by a social network within a given location use similar language and phrases, and thus, a single common location can be inferred [5, 12, 14]. Such observations can be collected from social media and used to connect users to a location with some probability, which has been found to decrease with increasing distance [1].

In this paper, we describe a model to exploit the connections between keywords used by people in different locations to *recommend* locations for messages missing geolocation tags. Our proposed approach casts the geolocation problem as a recommendation one. Relationships between users and their locations can be represented as a series of matrices: a user-location matrix mapping users to tagged locations, a user-observation matrix mapping users to *important* keywords, and a location-observation matrix mapping locations to known related keywords. We use a geospecific tf-idf score to rank key phrases and extract location specific terms — observations — from the known tweet text. We propose SpinRec, a set of matrix factorization strategies applied to these matrices to *fill in the blanks*, or leverage the known user and location values to predict the unknown. These strategies have proven to show higher accuracy and provide more meaningful results in the recommendation space [10].

The contributions made by this paper are:

- SpinRec, a cyclical model for collaborative filtering.
- A novel method of determining location by the analysis of tweet content.
- A gradient-descent based algorithm for matrix factorization.

The remainder of the paper is organized as follows: we present past works on social media geolocation (§2), give a brief background on collaborative filtering (§3), outline the proposed cyclical matrix factorization approach (§4), and give experiments that will determine the efficacy of this approach (§5).

LocalRec'18, November 6, 2018, Seattle, WA, USA

ACM ISBN 978-1-4503-6040-1/18/11…$15.00
https://doi.org/10.1145/3282825.3282831

2 RELATED WORK

Most recent work on geolocation inference uses factor-graph models to infer location. Qian et al. [15] introduces deep learning techniques to support supervised and semi-supervised learning for this task. Most of early research work in location inference focuses on analyzing the contents of user's posted messages, such as Cheng et al. [3] who proposes and evaluates a probabilistic framework that infers city level location using only user's posted messages. Wing and Baldridge [17] infers user's location using supervised information retrieval techniques. Eisenstein et al. [4] proposes a model that predicts geographic location from raw text data with better performance than previously proposed supervised topic models. Ikawa et al. [8] estimated the user's location by learning the relation between specific locations and relevant keywords from previous microblog messages.

Other work focuses mainly on network structure to infer location without considering the contents of messages. The use of network structure in location inference modeling is based on the observation that there is higher probability of having a large network of friends from the same time-zone than probability of having friends with a distance of three time-zones [5]. McGee et al. [13] propose a network-based approach using social strength between users to infer location. Li et al. [11] introduces an approach combining the analysis of the contents of messages with social network of the users using only discrete names of locations for estimation.

Unlike the existing techniques in the literature, our approach utilizes matrix factorization techniques to learn relationships between a set of users and locations with respect to the set of observations obtained from contents of the tweets. Analysis of message content is used to create a set of observations which will include the local words, unique features, and phrases specific to locations.

3 BACKGROUND

Recommender systems are typically proposed as finding relations amongst a set of users U and items I such that a model may find items in I with which a user in U will interact. This is performed using either a *content based* method or *collaborative filtering* method. Content based recommender systems leverage item metadata to recommend items to a target user by finding similar items within I to some item with which the target user has already interacted. In the context of geolocation, this would translate to using location metadata.

Alternately, collaborative filtering techniques leverage similarity among users in U to recommend. A matrix M can be formed where each row is a user and each column an item. Thus, each entry is a value representing the interaction between a particular user and a particular item. This matrix is very sparse; it is unlikely a user has interacted with all items, or even most of them. Matrix factorization (or more specifically, Singular Value Decomposition) is commonly used to fill these blanks. Where M is the previous user-item matrix:

$$M = WSV^T \tag{1}$$

The matrix S is a diagonal matrix containing the unique *singular values* of the matrix M. These singular values can be used to measure how much a concept learned either applies to a user (matrix W) or an item (matrix V^T). To get an estimated interaction value r for a user j and item k, take the dot product of the corresponding vectors: $r = W[j] \cdot I[k]$.

Tf-idf (term frequency-Inverse document frequency) model is a popular techniques used for information retrieval and text mining in a collection of documents to estimate the importance of any given word in the collection.

Tf-idf is a two step model. The first step, term frequency, is defined as the frequency of occurrence of a word in a document. The second step, inverse document frequency, is defined as the frequency of occurrence of a word in the number of documents across the collection. The combination of these two steps normalize the weight of importance for a word; if a word has higher frequency in a document it will give higher weight to term frequency. High frequency in all documents implies decrease in the value of the idf. This feature will avoid giving higher weight to the frequently used common terms such as any articles or pronouns and provide more accurate weights to significant words and phrases. The formal formulation is as follows:

$$R_{tf-idf} = tf_{i,j} x \log(N/tf_j) \tag{2}$$

where $tf_{i,j}$ is the number of occurrences of word in the j^{th} document. N is the total number of documents in the collection. tf_j is the number of documents in which the word appeared.

4 APPROACH

In the proposed model, we formulate the geolocation inference problem as a recommendation problem where a target user's location can be estimated using other user's known locations tied with the contents of posts from the target user. For each user, our goal is to predict their location as set of latitude and longitude coordinates. We assume the implicit data relates users with known locations, posts with associated keywords, and locations tied with specific locations.

The use of latitude and longitude allows us to estimate the location of the user to an adjustable precision. The granularity of the geo-location is 'city, state'. The set of observations is a collection of key features such as meaningful phrases, local words, name of local events, etc. that can be mapped to a user's location. To compute the key location we cluster know posts by their locations, combine the text content as documents, and produce a ranked list of terms for each location using tf-idf scoring. This ranked list represents terms that are most important for each location, a type of geospecific tf-idf. Articles and other such meaningless phrases were filtered to achieve better accuracy.

We now define SpinRec, a cyclical matrix factorization technique to perform geolocation. The input to the algorithm is a set of three matrices, UL, UO and LO:

$$UL_{mxn} \rightarrow UserLocation\ Matrix$$

$$UO_{mxk} \rightarrow UserObservation\ Matrix$$

$$LO_{nxk} \rightarrow LocationObservation\ Matrix$$

Where m is the number of users, n is the number of locations and k is the number of observations. The general algorithm is detailed in Algorithm 1.

The algorithm attempts to optimize the three matrices in series of recommendation rounds, hence the name *SpinRec*. Each round

Algorithm 1 SpinRec: Multi-stage collaborative filtering

INPUT: UL, UO, LO
procedure MatrixUpdate(M_1, M_2, M_3)
  $\grave{M_1} \leftarrow MatrixFactorization(M_1)$
  $\tilde{M_2} \leftarrow \frac{1}{\left\|\grave{M_1}M_3\right\|_F}\grave{M_1}M_3$
  $M_{2_{diff}} \leftarrow M_2 - \tilde{M_2}$
  $\gamma \leftarrow M_{2_{diff}}$
  $M_{1_{new}} \leftarrow \grave{M_1}(\gamma)$
  return $M_{1_{new}}$
end procedure
while *no convergence* **do**
  $UL_{update} \leftarrow MatrixUpdate(UL, UO, LO)$
  $UO_{update} \leftarrow MatrixUpdate(UO, LO, UL)$
  $LO_{update} \leftarrow MatrixUpdate(LO, UL, UO)$
  $UL \leftarrow UL_{update}$, $UO \leftarrow UO_{update}$, $LO \leftarrow LO_{update}$
end while
return UL, UO, LO

consists of updating all three matrices using the *MatrixUpdate* function. *MatrixUpdate* takes as input M_1, the matrix to be updated, along with matrices M_2 and M_3 with which M_1 can be reconstructed. The steps of *MatrixUpdate* are as follows:

(1) Perform matrix factorization on M_1 to predict all missing values. Call this $\grave{M_1}$.
(2) Multiply $\grave{M_1}$ and M_3 and normalize to reconstruct M_2. Call this $\tilde{M_2}$.
(3) Subtract M_2 and $\tilde{M_2}$ to obtain $M_{2_{\text{diff}}}$, a matrix defining the amount by which $\tilde{M_2}$ has incorrectly predicted M_2.
(4) Compute γ, the average difference between values defined in $M_{2_{\text{diff}}}$.
(5) Adjust M_1 using γ to obtain the new M_1 matrix.

This function is applied to all three matrices in each round of SpinRec, such that updated versions of UL, UO, and LO are all obtained. The updated matrices replace the originals, and the process continues until convergence.

At each execution of *MatrixUpdate*, the input matrix M_1 is the matrix with which we want to obtain a new predictor, based on the M_2 and M_3 matrices. It is expected these matrices will contain majority gaps, thus the matrix factorization step on the current predictor matrix is a preliminary collaborative filtering method to attempt to predict the missing data. The predicted matrix is normalized to produce confidences in each prediction from 0 to 1.

It is necessary to identify a method by which error can be gauged. The matrix M_2 is treated as the "ground truth" matrix; and thus is the matrix for which we would like to obtain predicted values. The produced $\tilde{M_2}$ and original M_2 are subtracted and the result averaged to obtain the average direction in which the target M_1 should be shifted to decrease error, similar to back propagation in a neural network. As input data is initially given for each of the UL, UO, and LO matrices, each matrix is treated as ground truth in the update step of another.

5 FUTURE EXPERIMENTS

Our proposed model will be evaluated using two case studies which will explain the benefits of geolocation inference for the domain of disaster management. The first case study is to analyze and predict the location of a tweet implying a high probability of there being a hail storm. Predicting hail at different locations using data from social media is a complex problem; if using contents to analyze the key features of a tweet, the keyword "hail" may be included but with a different context. Also, the time of the original tweet will be an important factor to predict meaningful information. Our future work includes introducing deep-leaning in our model to address the problem when users are new and doesn't have much tweets to relate the observations with.

The second case study is to analyze and predict the location of users with the contents of tweets about foodborne illness, which may further aid in analyzing the probability of an outbreak at that location. The known location of the user can potentially allow us to estimate numbers of users who are at risk of illness at that particular location. With the increase in use of social media for day to day activities, a lot of meaningful information can be obtained from user's posted messages and social networks. Such observation may help us in analyzing the outbreaks and submit an early warning to the necessary audience.

To conduct these experiments we will need sets of "ground truth". For each of the applications above we have collected sets of tweets with known locations, thus we plan to use a tweet dropout strategy to measure the effectiveness of SpinRec as is typical with recommender systems [16]. In this strategy, a subset of the known tweet and location vectors are held out during training to be used as a testing set. The recommender is then run on these tweets and error is measured based on the strength of the recommendation for the corresponding location.

6 CONCLUSION AND FUTURE WORK

In this paper, we present our vision for a novel approach to predict a precise location of the user using data from social media. This is our first step towards utilizing different techniques of collaborative filtering to solve the problem of finding user's location using post data.

As with any newly developed learning method, it will be pertinent to prove the convergence of SpinRec in the general case. Presented in this paper is a sub problem of relating two attributes, users and locations, via a third: observations. This can be generalized to any of n attributes in a chain relation.

Alternately, it may be beneficial to explore deep learning options for the geolocation problem. The current approach does not exploit the continuous nature of latitude and longitude; we are currently treating each known location as a discrete value and thus expect to experience an instance of the common recommender cold start problem. When we wish to include new locations into the system, we will have to retrain the model.

We have high hopes for the efficacy of this model to predict locations of users based on social media data. The current work described above in predicting tweet locations has produced desirable results so far, however we have yet to test this model in a scalable environment.

7 ACKNOWLEDGEMENTS

This work was partially funded by the Robert Wood Johnson Foundation (Grant #73362). The FAA has partially sponsored this project through the Center of Excellence for Technical Training and Human Performance. The agency neither endorses or rejects the findings of this research.

REFERENCES

[1] Lars Backstrom, Eric Sun, and Cameron Marlow. 2010. Find me if you can: improving geographical prediction with social and spatial proximity. In *Proceedings of the 19th international conference on World wide web*. ACM, 61–70.

[2] Nina Cesare, Christan Grant, and Elaine O Nsoesie. 2017. Detection of user demographics on social media: A review of methods and recommendations for best practices. *arXiv preprint arXiv:1702.01807* (2017).

[3] Zhiyuan Cheng, James Caverlee, and Kyumin Lee. 2010. You are where you tweet: a content-based approach to geo-locating twitter users. In *Proceedings of the 19th ACM international conference on Information and knowledge management*. ACM, 759–768.

[4] Jacob Eisenstein, Brendan O'Connor, Noah A Smith, and Eric P Xing. 2010. A latent variable model for geographic lexical variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 1277–1287.

[5] Raymond Heatherly, Murat Kantarcioglu, and Bhavani Thuraisingham. 2009. Social network classification incorporating link type values. In *Intelligence and Security Informatics, 2009. ISI'09. IEEE International Conference on*. IEEE, 19–24.

[6] Erik Holbrook, Gupreet Kaur, Jared Bond, Josh Imbriani, Elaine Nsoesie, and Christan Grant. 2016. Tweet Geolocation Error Estimation. In *International Conference on GIScience Short Paper Proceedings*, Vol. 1.

[7] Yuan Huang, Diansheng Guo, Alice Kasakoff, and Jack Grieve. 2016. Understanding US regional linguistic variation with Twitter data analysis. *Computers, Environment and Urban Systems* 59 (2016), 244–255.

[8] Yohei Ikawa, Miki Enoki, and Michiaki Tatsubori. 2012. Location inference using microblog messages. In *Proceedings of the 21st International Conference on World Wide Web*. ACM, 687–690.

[9] David Jurgens, Tyler Finethy, James McCorriston, Yi Tian Xu, and Derek Ruths. 2015. Geolocation Prediction in Twitter Using Social Networks: A Critical Analysis and Review of Current Practice. *ICWSM* 15 (2015), 188–197.

[10] Yehuda Koren, Robert Bell, and Chris Volinsky. 2009. Matrix factorization techniques for recommender systems. *Computer* 8 (2009), 30–37.

[11] Rui Li, Shengjie Wang, Hongbo Deng, Rui Wang, and Kevin Chen-Chuan Chang. 2012. Towards social user profiling: unified and discriminative influence model for inferring home locations. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 1023–1031.

[12] Jack Lindamood, Raymond Heatherly, Murat Kantarcioglu, and Bhavani Thuraisingham. 2009. Inferring private information using social network data. In *Proceedings of the 18th international conference on World wide web*. ACM, 1145–1146.

[13] Jeffrey McGee, James Caverlee, and Zhiyuan Cheng. 2013. Location prediction in social media based on tie strength. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*. ACM, 459–468.

[14] Delia Mocanu, Andrea Baronchelli, Nicola Perra, Bruno Gonçalves, Qian Zhang, and Alessandro Vespignani. 2013. The twitter of babel: Mapping world languages through microblogging platforms. *PloS one* 8, 4 (2013), e61981.

[15] Yujie Qian, Jie Tang, Zhilin Yang, Binxuan Huang, Wei Wei, and Kathleen M Carley. 2017. A Probabilistic Framework for Location Inference from Social Media. *arXiv preprint arXiv:1702.07281* (2017).

[16] Guy Shani and Asela Gunawardana. 2011. Evaluating recommendation systems. In *Recommender systems handbook*. Springer, 257–297.

[17] Benjamin P Wing and Jason Baldridge. 2011. Simple supervised document geolocation with geodesic grids. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. Association for Computational Linguistics, 955–964.