User-movement Estimation in Social Media Sites Based on Seq2Seq Model

Masaharu Hirota
Department of Information Science, Okayama University of Science
1-1 Ridaicho, Kita-ku, Okayama-shi, 700-0005, Japan

E-mail: hirota@ous.ac.jp

Abstract

Many tourists upload content about tourist attractions to social media sites. The sequence of location information annotated on the content represents the user's movement. This movement information is an important source that can be used for recommendations and advertisements. However, users do not consistently post content about all the places they visit on social media. sites. This study aims to develop a method to estimate the location of users' movements. Our method uses a Seq2Seq model, which learns the reconstruction of users' movement trajectories.

Keywords: Sequence-to-sequence model, Human mobility, Trajectory, Twitter,

1. Introduction

In recent years, with the proliferation of social media sites, location-based services, and GPS-equipped mobile devices, many movement trajectories have been generated, such as people's movements, check-ins, vehicle trajectories, and surveillance camera data. These trajectories are essential information for applications such as arrival time prediction [1], traffic volume prediction [2], personalized location recommendation [3], and trajectory similarity measurement [4]. The movement trajectory dataset's quality significantly impacts these applications' performance. This is because low sample rate trajectories increase the uncertainty of the movement information as much of the user's movement is lost. Therefore, ensuring many movement trajectories with high sample rates is essential for these applications.

In addition, social media sites such as Twitter [5] have large amounts of location-annotated content posted by people. Therefore, the sequence of such content represents a user's movement and is used because of the applications mentioned above. However, to use a user's content sequence as a movement trajectory on Twitter, it is necessary to consider the fact that users do not post tweets at all destinations. In other words, there is a significant difference between the true sequence representing the actual movement of a user and the sequence represented by a tweet, which is the user's movement of a tweet. As a result, sequences obtained from social media, such as tweets, are missing information on many moves and have low sample rates. This paper addresses pseudo-estimation of user movement by performing trajectory recovery on sequences of user movements on Twitter.

Trajectory recovery increases the movement trajectory's sample rate by recovering the trajectory's lost points. For example, trajectory recovery methods have been proposed based on the Sequence-to-Sequence (Seq2Seq) [6] model and mainly focused on increasing the sample rate of vehicle and pedestrian movement trajectories. The movement trajectories in the datasets used in those studies often have a sample rate of a few seconds to a few minutes. On the other hand, tweet sequences' sample rate is from a few minutes to a few hours. In this study, we treat sequences of tweets with lower sample rates than those movement trajectories.

In this study, user movement is estimated by increasing the sample rate of tweet sequences using the attention-based Seq2Seq model. The user movement estimated in this study is the movement between sub-areas, where an area is divided into a grid. The model has an encoder that converts the sequence into a vector and a decoder that creates the sequence from the vector. The sequence of tweets is input to the encoder, and the output of the decoder is the estimated movement of the tweet.

The remainder of this paper is organized as follows. Section 2 describes work related to this topic. Section 3 presents our method for estimating user-movements using the Seq2Seq model. Section 4 describes the evaluation result of our method. Section 5 concludes the paper and future work.

2. Related Work

Studies have been conducted to recover low-sample-rate motion trajectories to generate high-sample-rate motion trajectories. Chen et al. propose an algorithm to discover the most popular route from a transfer network based on the popular indicators in a breadth-first manner [7]. Hoteit et al. evaluated cell phone data recovery for movement trajectory using various recovery methods considering mobility parameters [8]. The experiments showed that the effective methods of human movement trajectory recovery differ according to information such as the mode of movement and the user's activity zone.

Over the last few years, there has been an increasing number of studies on trajectory recovery using deep learning. In [9], Wang et al. proposed subseq2Seq, which uses Kalman filter calibration for motion trajectory recovery and considers both temporal and spatial attention. Xia et al. proposed a new model for trajectory recovery based on the neural network using attention [10]. Ren et al. proposed a Seq2Seq-based model for movement trajectory recovery because it uses information on road segments and trajectories [11]. The model implements map matching and recovery to a high sample rate in an end-to-end manner. Sun et al. proposed a model for trajectory recovery using a Graph Neural Network for graphs constructed from movement trajectories [12]. The model employs an attention layer which considers the multilevel periodicity and shifting periodicity of human mobility, respectively.

Most trajectory recovery studies deal with motion trajectory recovery with sample rates of several seconds to a few minutes, as in datasets such as Geolife dataset. For example, taxis usually report GPS locations every 2 ~ 6 minutes to reduce the energy consumption of communication [13]. Therefore, the above methods are helpful for movement trajectory recovery for such data sets. We address movement trajectory recovery for Twitter in this study, having an even lower sample rate.

3. Methodology

In this study, we use the Seq2Seq model because of the estimation of the user's movement. The problem of estimating user movement is similar to that of machine translation. For example, the Seq2Seq model used in machine translation can translate an English sentence into a Japanese sentence. This study translates the original sequence of tweets to produce a recovered sequence.

We denote the sequence of tweets annotated with latitude and longitude information by $S = \{(t_1, t_2, \cdots, t_n)\}$, where $t_i = (lat, lng, time)$, $\forall i$, $1 \leq i \leq n$, which denotes the latitude, longitude, and timestamp. Also, t_i is each tweet in S. Next, we sort the sequence S in order of post time by ascending order.

Next, for each tweet t_i in S, we obtain quadtree keys (quadkey) [14]. quadkey is the world map into tiles (i.e., grids) of the same size at different levels of size and uses quadkey for the grid. We get quadkey of a particular level from the latitude and longitude of t_i. Each tweet t_i is then transformed into $t_i = (quadkey, time)$. If two consecutive quadkey in the sequence are the same, we remove the backward value. The result of quadkey sequence is $K = \{k_1, k_2, \cdots, k_m\}$, where k_i is quadkey, $\forall\, i, 1 \leq i \leq m$.

In this study, we implement a model that learns to recover the original sequence K from this partially deleted sequence $\hat{K}$. We create a sequence $\hat{K} = \{\widehat{k_1}, \widehat{k_2}, \cdots, \widehat{k_l}\}$, $\forall\, i$, $1 \leq i \leq l$, by randomly deleting keys in key sequence K.

Our model consists of an encoder and a decoder. The embedding layer of the encoder transforms each key $\widehat{k_\iota}$ of a sequence $\hat{K}$ into a low-dimensional vector. We use a Gated Recurrent Unit (GRU) as the encoder. GRU is a variant of Long Short-term Memory networks, which is capable of learning long-term dependencies for sequential data without performance decay. The decoder consists of a dropout layer, an attention layer, and a GRU layer. We also use cross-entropy as the loss function of our model. The final model output is $\hat{K}$ transformed into $K' = \{k'_1, k'_2, \cdots, k'_o\}, \forall\, i, 1 \le i \le o$.

4. Experimental Evaluation

4.1. *Experimental Setting*

In this experiment, we use tweets, which include latitude and longitude information. The tweets posted in Kyoto, Japan and were obtained from January 1, 2015, to December 31, 2015. Next, we divided the sequence of user tweets into daily segments. We extracted only those sequences with more than five quadkeys. As a result, the number of tweet sequences is 11,050.

We randomly deleted 20% of each key sequence because we wanted to create sequences with low sample rates. We also split the set of key sequences into a training set, a validation set, and a test set in the ratio 0.90 : 0.05 : 0.05. We set the zoom level of the quadkey to 15. A tile at this level has a side of approximately 1,222 meters. The resulting number of quadkey tiles is 386.

4.2. *Evaluation Criterion*

In this experiment, we use four criteria. The first and second are the mean absolute error (MAE), and the root mean square error (RMSE) between predicted values K' and ground truth K. MAE and RMSE are formulated as:

$$\boldsymbol{MAE} = \frac{\mathbf{1}}{\boldsymbol{n}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \left| \boldsymbol{dis}(\boldsymbol{k_i}, \boldsymbol{k'_j}) \right|$$

$$\boldsymbol{RMSE} = \sqrt{\frac{\mathbf{1}}{\boldsymbol{n}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \boldsymbol{dis}(\boldsymbol{k_i}, \boldsymbol{k'_i})^2}$$

where $dis(k_i, k'_i)$ is the geodesic distance (meter) between k_i and k'_i. The smaller values of MAE and RMSE indicate the distance between the original sequence and the predicted sequence is close. The third and last is the recall and precision formulated as:

$$Recall = \frac{|K \cap K'|}{|K|}, Precision = \frac{|K \cap K'|}{|K'|}$$

Here, $|K \cap K'|$ is defined by the value of the longest common subsequence of K and K'. The larger recall and precision values indicate that methods predict the tiles more accurately.

4.3. *Evaluation Result*

Table 1. Evaluation result.

MAE	RMSE	Recall	Precision
1674.642	2447.872	0.482	0.686

Table 1 shows the evaluation result. As the MAE and RMSE values indicate, the difference in distance between the predicted results and the ground truth sequence is insignificant. Also, compared to Precision, Recall is small because it sometimes fails to recover tiles contained in the original sequence. However, each evaluation criterion indicates that the proposed method is somewhat effective. Improving these performances is a future challenge.

5. Conclusion

This paper deals with predicting Twitter users' moves using the Seq2Seq model. The model performed trajectory recovery by learning pairs of missing and true movement sequences.

Future work includes the following topics One is to improve the structure of the model. The model used here is a simple Seq2Seq model, which does not adequately represent the data. The second is to improve the information used by the model. The current model only considers the order of positions in learning Twitter sequences, but it could consider the text of tweets and time intervals.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number JP19K20418.

References

1. H. Zhang, H. Wu, W. Sun, and B. Zheng, “Deeptravel: a neural network based travel time estimation model with auxiliary supervision,” arXiv preprint arXiv:1802.02147, 2018.
2. M. Li, P. Tong, M. Li, Z. Jin, J. Huang, and X.-S. Hua, “Traffic flow prediction with vehicle trajectories,” in Proceedings of the AAAI Conference on Artificial Intelligence, 2021, vol. 35, pp. 294–302.
3. Y. Cui, H. Sun, Y. Zhao, H. Yin, and K. Zheng, “Sequential-knowledge-aware next POI recommendation: A meta-learning approach,” ACM Transactions on Information Systems (TOIS), vol. 40, Art. no. 2, 2021.
4. D. Yao, G. Cong, C. Zhang, and J. Bi, “Computing trajectory similarity in linear time: A generic seed-guided neural metric learning approach,” in 2019 IEEE 35th international conference on data engineering (ICDE), 2019, pp. 1358–1369.
5. Twitter, https://twitter.com

6. I. Sutskever, O. Vinyals, and Q. V. Le, “Sequence to sequence learning with neural networks,” Advances in neural information processing systems, vol. 27, 2014.
7. Z. Chen, H. T. Shen, and X. Zhou, “Discovering popular routes from trajectories,” in 2011 IEEE 27th International Conference on Data Engineering, 2011, pp. 900–911.
8. S. Hoteit, S. Secci, S. Sobolevsky, C. Ratti, and G. Pujolle, “Estimating human trajectories and hotspots through mobile phone data,” Computer Networks, vol. 64, pp. 296–307, 2014.
9. J. Wang, N. Wu, X. Lu, W. X. Zhao, and K. Feng, “Deep trajectory recovery with fine-grained calibration using kalman filter,” IEEE Transactions on Knowledge and Data Engineering, vol. 33, Art. no. 3, 2019.
10. T. Xia et al., “AttnMove: History Enhanced Trajectory Recovery via Attentional Network,” Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, Art. no. 5, May 2021.
11. H. Ren et al., “MTrajRec: Map-Constrained Trajectory Recovery via Seq2Seq Multi-task Learning,” in Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, 2021, pp. 1410–1419.
12. H. Sun, C. Yang, L. Deng, F. Zhou, F. Huang, and K. Zheng, “PeriodicMove: Shift-Aware Human Mobility Recovery with Graph Neural Network,” in Proceedings of the 30th ACM International Conference on Information & Knowledge Management, 2021, pp. 1734–1743.
13. J. Yuan, Y. Zheng, C. Zhang, X. Xie, and G.-Z. Sun, “An Interactive-Voting Based Map Matching Algorithm,” in 2010 Eleventh International Conference on Mobile Data Management, 2010, pp. 43–52.
14. quadkey, https://learn.microsoft.com/en-us/azure/azure-maps/zoom-levels-and-tile-grids

Authors Introduction

Dr. Masaharu Hirota

He received a Doctor of Informatics degree in 2014 from Shizuoka University. After working for the National Institute of Technology, Oita College, he has worked in the Faculty of Informatics, the Okayama University of Science, since April 2017. His research interests include data engineering, GIS, and social media. He is a member of ACM, DBSJ, and IPSJ.