Similarity measurement of mobile phone user trajectories - a modified edit distance method

Yihong Yuan[1,2], Martin Raubal[1]

[1]Institute of Cartography and Geoinformation, ETH Zurich, 8093 Zurich, Switzerland
Email: {yyuan, mraubal}@ethz.ch
[2]Department of Geography, University of California, Santa Barbara, CA, 93106, USA
Email: yuan@geog.ucsb.edu

1. Introduction

Due to the natural variability of human mobility and the data quality of recorded location information, analyzing human trajectories has been a challenging research topic in several fields such as Computer Science, Transportation, and Statistical Physics. Researchers have focused on different aspects, including *intro-trajectory* studies, i.e., understanding the internal regularity of human motions (Gonzalez et al. 2008), and *inter-trajectory* studies, i.e., measuring trajectory similarity between individuals (Xia et al. 2011). The latter has drawn more and more attention due to the increasing interest in understanding the social interaction among demographic groups (Eagle et al. 2009). Measuring trajectory similarity can help in many real-world applications, such as traffic analysis or crime prediction.

In addition, the rapid development of Information and Communication Technologies (ICTs) has introduced a wide range of novel spatio-temporal data sources (e.g., georeferenced mobile phone records) for researchers to explore the mobility patterns of its carriers. Although mobile phones are capable of recording location information through several ways such as assisted GPS, the collected data are normally generated as scattered points in call detail records (CDRs) which only include the location of cell towers; therefore, these data can be viewed as strings of cell IDs characterized by low accuracy and precision in both space and time dimensions. However, there has not been sufficient research on how to investigate the similarity between user trajectories based on these scattered and cell-constrained sample points. As argued in (Kang et al. 2009), a similarity measure in cellular space is different from that in Euclidean space because numerical information in cellular space is not necessarily continuous (Kolbe et al. 2008). Therefore most of the existing algorithms, such as the Synchronized Euclidean Distance (SED) (Zheng and Zhou 2011), are not easily applicable.

In this research we focus on measuring trajectory similarity based on a modified edit distance algorithm, which has been first proposed for string correction (Wagner and Fischer 1974). This method calculates the minimum number of operations required to switch one string to another and is therefore highly suitable for matching time series of cell IDs in CDRs. Another advantage of this method is that it can deal with sequences of different lengths, which is a typical situation in CDRs. However, traditional edit distance deals with purely qualitative information such as string matching; therefore, in order to preserve the spatial information in CDRs, we will modify the cost function of the algorithm to incorporate the spatial distribution of the cell towers. The modified

algorithm will be helpful for measuring inaccurate tracking information in CDRs, as well as facilitating the interpretation of user mobility patterns in the age of instant access.

2. Dataset

For this research we utilize a dataset from northeast China, which covers over one million people and includes CDRs for a time span of 9 days (5 weekdays, 4 weekend days). It includes the time, duration, and the location of the corresponding cell tower for each mobile phone connection (Yuan et al. 2012). Table 1 provides a sample record. The phone number, longitudes and latitudes are not shown for reasons of privacy.

Table 1. Sample record from the example data set.

Phone #	13601******
Longitude (cell tower)	126.*****
Latitude (cell tower)	45.*****
Time	16:10:31
Duration	11 mins

3. Methodology and preliminary results

As discussed in Section 1, the method in this research is based on the edit distance algorithm proposed by Wagner and Fischer (1974), which measures the distance between two strings by computing the edit operations when transforming one string to another. The pseudocode is shown as follows:

Given two strings $S(s_1 \ldots s_i)$ and $T(t_1 \ldots t_j)$, in the optimal solution, to transform S to T there are three solutions:

- s_i is deleted and the rest $s_1 \ldots s_{i-1}$ is transformed to $t_1 \ldots t_j$,
- $s_1 \ldots s_i$ is transformed into $t_1 \ldots t_{j-1}$ and we insert t_j at the end
- s_i is changed into t_j and the rest $s_1 \ldots s_{i-1}$ is transformed to $t_1 \ldots t_{j-1}$.

Thus the recursive algorithm can be defined as,

$$EditDis[i, j] = \min(EditDis[i - 1, j] + Cost[delete\ (s_i)], EditDis[i, j - 1] + Cost[insert\ (t_j)], EditDis[i - 1, j - 1] + Cost[replace(s_i, t_j)]).$$

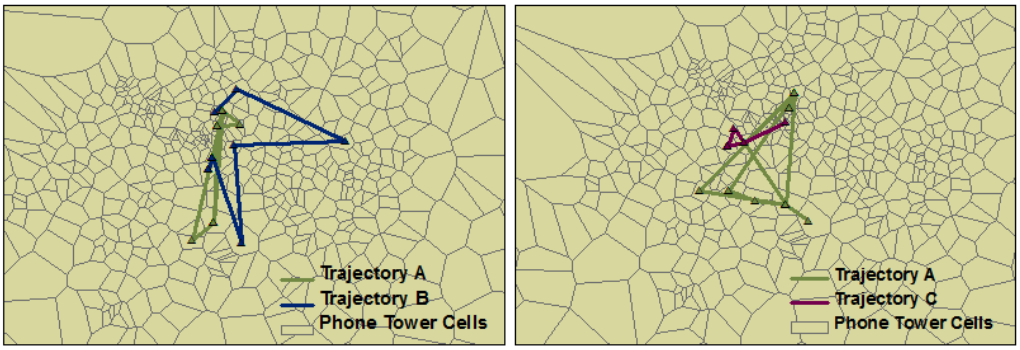
where the cost function is defined based on practical needs. In string matching the cost of each operation is usually set as constant 1.

In CDRs, the trajectory of a phone user can be represented by a sequence of cell IDs, for example, [Cell5, Cell6, Cell5, Cell4]. The distance between two trajectories can be measured by the cost of operations required to transform one sequence to the other. However, the tracking information in CDRs is not purely qualitative, indicating that each operation should be assigned a different cost value based on the locations of the deleted/inserted/replaced points. Given two trajectories $S(s_1 \ldots s_n)$ and $T(t_1 \ldots t_m)$ (where each s_i and t_j represents a record with the location of a corresponding tower), we assign the operation cost based on the "influence" of each operation on both S and T, for

example, deleting a far-away point will generate a higher cost than deleting a point which is close to the mass center of the original trajectories. Hence the cost function is defined as:

cost [delete (s_i)] = the average distance between s_i and the mass center of two trajectories *S* and *T;*
cost [insert (t_j)] = the average distance between t_j and the mass center of two trajectories *S* and *T;*
cost [replace(s_i, t_j)] = the distance between s_i and t_j.

To demonstrate the algorithm we randomly selected 1000 users whose CDRs include more than 10 records for both weekdays and weekends. Figure 1 shows the result of a comparison between an example user *A* and the other 999 users. For comparison we scale the result distances to the range [0, 1]. Trajectories *B* and *C* are the most similar trajectories to *A* on weekdays (Figure 1a, scaled distance = 0.017) and weekend days (Figure 1b, scaled distance = 0.035). As can be seen, the algorithm mainly focuses on matching trajectories which cover a similar activity area. Moreover, it also considers the movement direction of how the points are visited in a timely order; therefore, the shapes of the trajectories do not necessarily look very similar.



(a) (b)

Figure 1. Example analysis: the most similar trajectories for (a) weekdays and (b) weekend days.

As indicated in Figure 1, the modified edit distance method is effective in identifying similar trajectory patterns of phone users. The result of this analysis can be useful for phone companies to interpret user activities, as well as improving algorithms in social network applications, such as “friend recommendation” based on movement patterns.

For a better overview of the sample set, we also calculated the average distance between each user and the other 999 users. This measurement can be considered as an indicator of how “different” a user behaves in terms of mobility pattern compared to the others. As shown in the weekday histogram (Figure 2, mean value = 0.20), the distance follows a skewed normal distribution, and the positive tail indicates that there are a certain number of users with large average distance (>0.5).

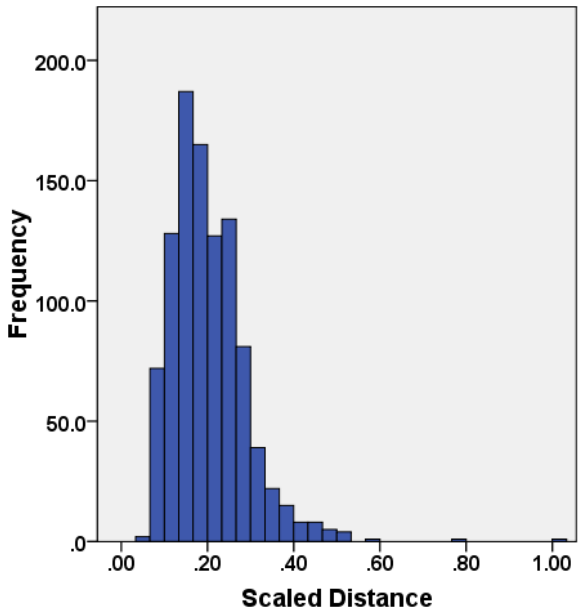


Figure 2. Histogram of average distance in weekdays.

For illustration purposes we also plot the spatial points of the user who has the largest average distance in our sample. As can be seen in Figure 3, this user's activity area is quite large (some points are even located outside of the target area). Such analysis and visualization can be helpful for detecting abnormal patterns in mobile networks.

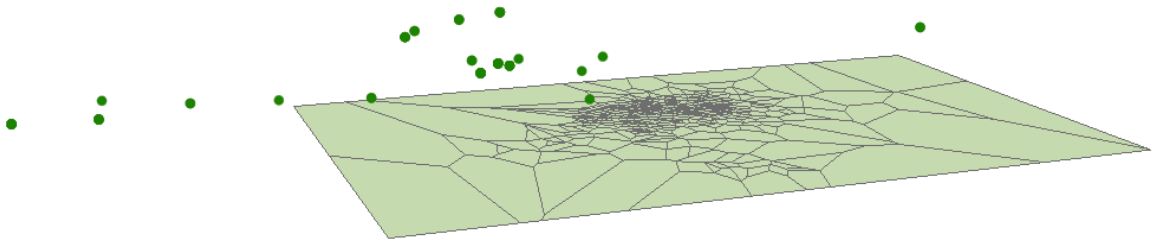


Figure 3. User with the largest average distance.

Another valuable part of the presented method is that it allows for further customizing the cost function based on practical needs. For instance, if researchers are interested in the question "Who's night-hour activity pattern is similar to User A?", they can incorporate the temporal restriction when calculating the cost functions. This type of research question is also one of our future work directions.

4. Conclusion

In this paper we have demonstrated a modified edit distance method to measure the trajectory similarity of mobile phone users. The analysis shows that this method is highly effective in identifying patterns in a cellular environment, as well as providing input for cell phone companies and policy makers. In the future we will further modify the

algorithm to incorporate various types of cost functions. We will also apply this method to other datasets to further validate its robustness.

References

Eagle N, Pentland A and Lazer D, 2009, Inferring friendship network structure by using mobile phone data. *Proceedings of the National Academy of Sciences of the United States of America*, 106: 15274-15278.

Gonzalez MC, Hidalgo CA and Barabasi AL, 2008, Understanding individual human mobility patterns. *Nature*, 453: 779-782.

Kang H-Y, Kim J-S and Li K-J, 2009, Similarity measures for trajectory of moving objects in cellular space. In *SAC*, 1325-1330.

Kolbe TH, Becker T and Nagel C, 2008, 1st Technical Report - Discussion of Euclidean Space and Cellular Space and Proposal of an Integrated Indoor Spatial Data Model. Berlin: Technische Universität Berlin.

Wagner RA and Fischer MJ, 1974, The String-to-String Correction Problem. *Journal of the ACM*, 21: 168-173.

Xia Y, Wang GY, Zhang X, Kim GB and Bae HY, 2011, Spatio-temporal Similarity Measure for Network Constrained Trajectory Data. *International Journal of Computational Intelligence Systems*, 4: 1070-1079.

Yuan Y, Raubal M and Liu Y, 2012, Correlating mobile phone usage and travel behavior - a case study of Harbin, China. *Computers, Environment and Urban Systems*, 36: 118-130.

Zheng Y and Zhou X, 2011, *Computing with spatial trajectories*. New York: Springer.