Similarity Measurement of Moving Object Trajectories

Hechen Liu & Markus Schneider*
Department of Computer & Information Science & Engineering
University of Florida
Gainesville, FL 32611, USA
{heliu, mschneid}@cise.ufl.edu

ABSTRACT

To study the similarity between moving object trajectories is important in many applications, e.g., to find the clusters of moving objects which share the same moving pattern, and infer the future locations of a moving object from its similar trajectories. To define the similarity between moving objects is a challenging task, since not only their locations change but also their speed and semantic features vary. In this paper, we propose a novel approach to measure the similarity between trajectories. The similarity is defined based on both geographic and semantic features of movements. Our approach can be used to detect trajectory clusters and infer future locations of moving objects.

Keywords

Moving objects, trajectory similarity, semantic trajectories

1. INTRODUCTION

The research on moving object databases [2] in recent years has attracted a lot of attention in many applications such as location-based services, traffic management and hurricane research, etc. Some moving objects in the real world share the same moving patterns, which is an important feature and can help people solve real problems. For example, assume that a large number of taxis take the similar routes between two destinations, then we may detect the representative trajectory between these two destinations, and can further infer future locations given the historical movement of a taxi. Intuitively, if two trajectories can be considered as similar to each other, they should satisfy some particular requirements, for example, they should be close enough to each other in the Euclidean space, and further they should have similar directions. Then a challenging problem is: how to measure the closeness?

*This work was partially supported by the National Science Foundation (NSF) under the grant number NSF-IIS-0812194 and by the National Aeronautics and Space Administration (NASA) under the grant number NASA-AIST-08-0081.

IWGS '12 Redondo Beach, California USA

A number of approaches measuring the similarity of the moving object trajectories are cluster-based methods [5] [3]. A partition-and-group algorithm to cluster similar trajectories is discussed in [3]. It defines three measurement, i.e., perpendicular distance, parallel distance and angle distance as the measurements of the similarity. While the above models find the trajectory similarity based on geographic features, some recent approaches introduce the semantic tags to enhance the accuracy of the measurement. The term *semantic trajectories* has been discussed in [1]. The authors consider trajectories as a set of *stops* and *moves*. An approach which mines the similarity of people's trajectories based on location histories is proposed in [4]. It defines a *stay point* as the place where a user stays for a while, and it carries a particular semantic meaning. A recent approach which detects similarity in semantic trajectories is proposed in [7]. The approach uses the longest common subsequence (LCSS) algorithm to find the similarity mainly on the semantics. A continuous work [6] adds geographic feature to measure the similarity and make prediction. It introduces two measurements, i.e. *SemanticScore* and *GeographicScore*. However, this approach first filters the trajectories by semantic similarity and then detect the geographic similarity, therefore two trajectories which are far from each other might have a very high similarity score, if their semantic similarity is high. The difference of our approach is that we compare geographic similarity at the beginning and then measure the semantic similarity.

In this paper, we propose a novel approach to measure the similarity between moving object trajectories. The method is based on both geographic features as well as semantic properties of trajectories. Then we combine both similarity measurements together and define a novel distance function.

The remaining part of the paper is organized as follows. In Section 2 we give the preliminary and an overview of our approach. In Section 3, we give the definition of similarity measurements between trajectories. We draw conclusions and discuss future work in Section 4.

2. PRELIMINARY

In this section, we give preliminary concepts of our research that will be used in further discussion in the rest of the paper. Then we show the framework of our approach.

As a human trajectory can show places that a people visited, we can detect the properties of a place and add semantic tags to it. We call such trajectories semantic trajectories. The semantic trajectory is defined as follows.

Definition 1 (Semantic trajectory) *A semantic trajectory is a list of points* $< (x_1, y_1, t_1, S_1), \ldots, (x_i, y_i, t_i, S_i), \ldots, (x_n, y_n, t_n, S_n) >$*, with the following conditions,*

$i)$ $n \in \mathbb{N}, i \in \{1, 2, ..., n\}$
$ii)$ $\forall 1 < i < j < n : t_i < t_j$
$iii) \forall 1 < i < n, S_i = \{s_{i1}, s_{i2}, \ldots, s_{im}\}, (x_i, y_i) \rightarrow S_i.$

In the above definition, Condition (iii) shows a *semantic mapping* process, i.e., each location that visited by a human can be associated with one or more semantic tags.

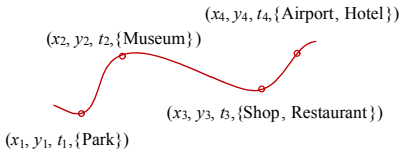

Figure 1: An examples of a semantic trajectory (b)

An example of a semantic trajectory is shown in Figure 1. We notice that the third point contains two semantic tags which are *Shop* and *Restaurant*, while the last point contains semantic tags of *Airport* and *Hotel*.

Given a large number of trajectories with semantic information, we perform the similarity measurement. We define the geographic and semantic similarities respectively, and calculate the similarity scores between all trajectories. In the end we calculate the total distance between two trajectories on top of the similarity scores.

3. DEFINE SIMILAR TRAJECTORIES

In this section, we define the similarity between trajectories. We consider in two aspects, the geographic similarity and semantic similarity.

3.1 Geographic Similarity

First, we define the similarity measurement between trajectories in geometry. A common problem in measuring the similarity of trajectories and trajectory clustering is the handling of sub-trajectories [3]. Given a trajectory which consists of a list of GPS points $tra =< (x_1, y_1, t_1), (x_2, y_2, t_2), \ldots, (x_n, y_n, t_n) >$, a sub-trajectory tra_k is a partial list of tra, and $tra_k =< (x_{k_1}, y_{k_1}, t_{k_1}), \ldots, (x_{k_m}, y_{k_m}, t_{k_m}) >$, where $1 \leq k_1 < k_2 < \ldots < k_m \leq n$. If a sub-trajectory contains only two GPS points, then it is a trajectory *segment*. A problem cost by not identifying sub-trajectories is shown in Figure 2a. Assume that we have a trajectory $tra_1 =< a_1, a_2, a_3, a_4, a_5 >$, where $a_1, a_2, \ldots$ are a list of GPS points, and $tra_2 =< b_1, b_2, b_3 >$, we find that portions of tra_1 and tra_2 are very similar. However, if we compare the entire trajectory of tra_1 with tra_2, it's hard to tell whether tra_1 and tra_2 are similar, because tra_1 has a sudden turn at point a_3. We call such point a *turning point*. But if we partition tra_1 into two sub-trajectories as $tra_{11} =< a_1, a_2, a_3 >$ and $tra_{12} =< a_3, a_4, a_5 >$, we are able to measure the similarity between tra_{11} and tra_2. Therefore, an important task is to detect the turning points of a trajectory and partition it into a list of sub-trajectories.

A turning point is detected by given a specific parameter φ which shows the degree of turns made by a trajectory. We define θ_i as the absolute bearing (direction to north)

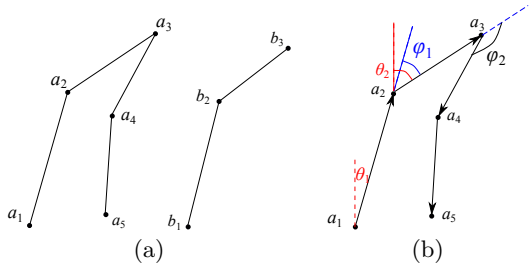

Figure 2: An example that a whole trajectory may not work in identifying clusters (a); using turns to detect the point to partition sub-trajectories (b)

of the i-th segment. Given the latitude and longitude of a spatial point, its bearing can be determined. Let φ_i denote the degree that a trajectory turns at the i-th point, then $\varphi = \theta_i - \theta_{i-1}$, as shown in Figure 2 b. Then we can select a range of φ, so that if the absolute value of the degree of a turn is greater than φ, we split the trajectory at this turning. In one extreme case, if we choose φ to be 0, then each segment will be partitioned. In the other case, if we choose φ to be π, then no trajectories will be split.

The next criteria that we want to consider to partition a trajectory is a temporal constraint. Previous algorithms only consider the similarity between the shapes of trajectories. However, we consider that time is an important feature in the moving pattern of a moving object. For example, a person's trajectory may consist of two parts: in the first part, he walks with the speed of 4 km/h, while in the second part he takes a taxi and travels with the speed of 50 km/h. Since these two parts show different moving patterns, we will split them. Therefore we define a speed ratio of the i-th segment τ_i and a temporal constraint τ as $\tau_i = \frac{v_i}{\bar{v_i}}$, where v_i is the average speed of the moving object at the i-th segment, and $\bar{v_i}$ is the average speed so far. If τ_i exceed the threshold τ we set, then we will partition the trajectory at the i-th point.

After partitioning trajectories into sub-trajectories, we are able to define the similarity between sub-trajectories.

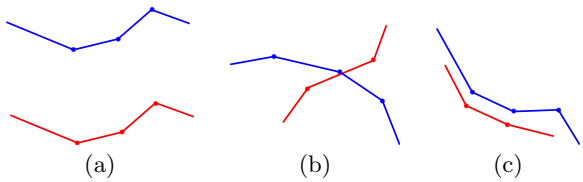

Figure 3: Two trajectories have exactly the same shape and direction while they are far from each other (a); two trajectories have overlap, but their directions are very different (b); two trajectories have the similar direction but different lengths (c).

.

When measuring the similarity between spatial objects, an intuitive way is to measure how close in distance they are to each other. A method is to consider these two spatial objects as two point sets, and find the minimum distance, as shown in Figure 3a. However, this is not a correct assumption. For example, if two trajectories overlap, then the distance will be zero, and these two trajectories will be

considered "similar", shown by Figure 3b, which is not accurate. Therefore, we need to consider their directions in the similarity measurement. Further, we must take into consideration the difference of the lengths between trajectories. A short trajectory will not be similar to a long trajectory (Figure 3c). With the careful consideration of all the above issues, we introduce the following concepts which are necessary for us to define the trajectory similarity.

Center of mass.

The center of mass of an arbitrary 2D shape $(\bar{x}, \bar{y})$ is the geometric center of this shape. We use $ctr(tra)$ to denote the center of mass of the trajectory tra. It is calculated by,

$$ctr(tra) = (\bar{x}, \bar{y}) = (\frac{\int x f(x) dx}{\int f(x) dx}, \frac{\int y f(y) dy}{\int f(y) dy})$$

where $f(x)$ and $f(y)$ denote the density distribution on x-coordinate and y-coordinate respectively.

Assume that we have a sub-trajectory $<(x_1, y_1, t_1), (x_2, y_2, t_2), \ldots (x_n, y_n, t_n) >$, and if we assume a uniform distribution of density, $f(x) = f(y) = 1$, then the center of mass is,

$$(\bar{x}, \bar{y}) = (\frac{\sum_{i=1}^{n-1}(x_{i+1}^2 - x_i^2)}{2\sum_{i=1}^{n}(x_{i+1} - x_i)}, \frac{\sum_{i=1}^{n-1}(y_{i+1}^2 - y_i^2)}{2\sum_{i=1}^{n}(y_{i+1} - y_i)})$$

A special case is to find the center of a segment, i.e., a sub-trajectory just contains two points (x_1, y_1, t_1), (x_2, y_2, t_2). The center is $(\frac{x_1 + x_2}{2}, \frac{y_1 + y_2}{2})$, which is in accordance with the above equation. It is obvious that the center of mass of a trajectory may or may not lie on the trajectory, as shown by Figure 4a and Figure 4b.

Displacement.

The displacement of a trajectory tra is the shortest distance from its origin to its destination (Figure 4c). We introduce this concept because it shows a *short cut* of a trajectory. The short cut always leads to the right destination and is useful in identifying representative trajectories. We use the term s to denote the displacement.

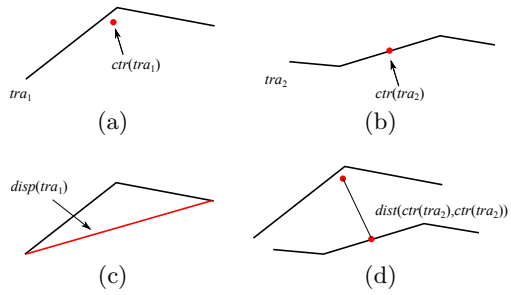

Figure 4: Center of mass lies outside the trajectory (a); center of mass lies on the trajectory (b); the euclidean distance between the center of mass (c); the displacement of a trajectory (d)

Cosine similarity.

The cosine similarity between two vector is a measure of the cosine of the angle between them, and the value is between [-1,1]. Here we use it to measure the similarity between the directions of two sub-trajectories. We calculate the cosine similarity between the displacements of two trajectories. A larger value shows a higher similarity and a smaller distance, therefore we add this metric as an negative term. Let s_1, s_2 denote the displacement of trajectory tra_1 and tra_2 respectively, the cosine similarity is defined as,

$$cos(s_1, s_2) = \frac{s_1 \cdot s_2}{\|s_1\| \|s_2\|}$$

Now we are able give the definition of the geometric distance between two sub-trajectories.

Definition 2 (Geographic distance) *Let tra_1 and tra_2 denote two trajectories. Let s_1 and s_2 denote the displacement of tra_1 and tra_2 respectively, and $\|tra\|$ denote the length of tra. The geographic distance between tra_1 and tra_2 is defined as,*

$$\begin{aligned} geoDist(tra_1, tra_2) &= ctrDist(tra_1, tra_2) \\ &+ ctrDist(tra_1, tra_2) \times \frac{|\ \|tra_1\| - \|tra_2\|\ |}{\max(\|tra_1, \|tra_2\|)} \\ &- \text{avg}(\|s_1\|, \|s_2\|) \times \cos(s_1, s_2) \end{aligned} \quad (1)$$

The first term measures the distance between the centers of mass. The second term measures the difference between the length of the trajectories. As we have shown in Figure 3c, the lengths of the trajectories effect their similarity, so that we consider the ratio of the difference to the maximum length of two trajectories. While the third term is the cosine similarity times the average length of two trajectories, which reduces the distance between trajectories. An advantage of this definition is that the distance function is *symmetric*, i.e., the distance from tra_1 to tra_2 is the same from the distance from tra_2 to tra_1. We show this property in the following lemma.

Lemma 1 *The distance function defined in Definition 2 is symmetric, which satisfies,*

$$geoDist(tra_1, tra_2) = geoDist(tra_2, tra_1)$$

Proof. The first term of Equation (2) is the euclidean distance between two points, which is symmetric. The second term is the first term times a ratio, and the denominate of the ratio is the larger value between the two, which is always the same, while the numerator is the absolute value of the difference of two numbers. The third term is a ratio, where the numerator is the dot product of two vectors, which is commutable, and the denominator will always return the same value no matter which vector comes first. Therefore this distance function is symmetric. □

3.2 Semantic Similarity

Now we discuss the semantic similarity between trajectories. We adopt the longest common subsequence algorithm. Similar method has appeared in [7], where the authors define the semantic similarity between two trajectories as two different ratios. However, we consider that the semantic similarity should be symmetric.

Definition 3 (Semantic similarity) *The semantic ratio between two trajectories measures a degree of semantic similarity between them, and is defined as*

$$semRatio(tra_1, tra_2) = \frac{LCSS(tra_1, tra_2)}{\min(|tra_1|, |tra_2|)} \quad (2)$$

Here we use $|tra|$ to represent the number of GPS points in tra. We use the dynamic programming approach to calculate the LCSS of two trajectories, as shown in Figure 5.

Algorithm Revised longest common subsequence
Input: two semantic trajectories tra_1, tra_2
Output: the semantic similarity ratio between tra_1 and tra_2

```
1   m ← |tra1|, n ← |tra2|,
2   Initialize M[m][n] //matrix to store the LCSS
3   for i ← 1 to m
4     M[i][0] ← 0
5   for j ← 1 to n
6     M[0][j] ← 0
7   for i ← 1 to m
8     for j ← 1 to n
9       max_len ←= M[i − 1][j − 1] + 1
10        if max_len < M[i − 1][j] and
11              M[i − 1][j] > M[i][j − 1]
12          max_len ← M[i − 1][j]
13        else if max_len < M[i][j − 1] and
14              M[i − 1][j] < M[i][j − 1]
15          max_len ← M[i][j − 1]
16        M[i][j] = max_len
17  // the minimum length of two trajectories
18  len = min(m, n)
19  return M[m][n]/len
```

Figure 5: The algorithm of revised longest common subsequence to determine the semantic similarity ratio between two trajectories

We show that this definition on the semantic similarity is also symmetric in the following lemma.

Lemma 2 *The defined semantic ratio function between two trajectories in Equation (3) is symmetric, i.e.,*
$semRatio(tra_1, tra_2) = semRatio(tra_2, tra_1)$

PROOF. As the LCSS in Equation (3) always returns the sub-sequence in common between two trajectories, and the denominator always returns the minimum length, then the ratio is always symmetric. □

Now we need to combine both measurements together. If we take a look at Definition 2, we may find that this is a measurement of the length. However, when we look into Definition 3 we observe that it is a ratio between [0,1]. An idea is to take the second measurement as a multiplier. However, the former is a distance function, but the latter shows a higher score when two trajectories are similar. Thus we should divide the second factor from the first. Therefore we give the following definition.

Definition 4 *The total distance of two trajectories tra_1 and tra_2 measures the similarity between them considering both geographic and sementic features, and is defined as*

$$totalDist(tra_1, tra_2) = geoDist(tra_1, tra_2) \times \frac{1}{1 + \alpha \times semRatio(tra_1, tra_2)} \quad (3)$$

where $0 \leq \alpha \leq 1$

Here we introduce a parameter α to adjust the portion that how much the semantic features can affect the similarity. As α increases, the higher the portion is, as it will shrink the value of the final distance and make two trajectories more similar. If we set α to zero, the distance is measured merely based on geographic distance.

We show that the definition in Equation (5) is symmetric by the following theorem.

Theorem 1 *The total distance function measuring the similarity between two trajectories in Equation (5) is symmetric, i.e.,*
$totalDist(tra_1, tra_2) = totalDist(tra_2, tra_1)$

PROOF. From Lemma 1 and Lemma 2 we have proved that the geographic distance and the semantic ratio are both symmetric, therefore in Equation (5) we will always get the same value no matter which trajectory comes first. □

4. CONCLUSIONS AND FUTURE WORK

In this paper, we draw the problem of finding similar moving object trajectories. We propose a novel approach to measure the similarity between trajectories considering both geometric and semantic properties. We first split a trajectory into sub-trajectories by considering the trajectory turns and temporal constraint and then define novel measurements on geographic similarity between sub-trajectories as well as semantic similarity. After that we define distance functions to measure the geographic similarity as well as semantic similarity. Our distance functions are proved to be symmetric. In the future, we will adopt this approach to cluster similar trajectories and find the representative trajectory to infer future locations of moving objects.

5. REFERENCES

[1] Luis Otavio Alvares, Vania Bogorny, Bart Kuijpers, Bart Moelans, Jose Antonio, Fernandes De Macedo, and Andrey Tietbohl Palma. Towards semantic trajectory knowledge discovery. *Data Mining and Knowledge Discovery*, 2007.

[2] Ralf Hartmut Güting and Markus Schneider. *Moving Objects Databases*. Morgan Kaufmann Publishers, 2005.

[3] Jae-Gil Lee, Jiawei Han, and Kyu-Young Whang. Trajectory clustering: a partition-and-group framework. In *ACM SIGMOD*, pages 593–604, 2007.

[4] Quannan Li, Yu Zheng, Xing Xie, Yukun Chen, Wenyu Liu, and Wei-Ying Ma. Mining user similarity based on location history. In *ACM SIGSPATIAL GIS*, pages 34:1–34:10, 2008.

[5] Jörg Sander, Martin Ester, Hans-Peter Kriegel, and Xiaowei Xu. Density-based clustering in spatial databases: The algorithm gdbscan and its applications. *Data Mining and Knowledge Discovery*, 2:169–194, 1998.

[6] Josh Jia-Ching Ying, Wang-Chien Lee, Tz-Chiao Weng, and Vincent S. Tseng. Semantic trajectory mining for location prediction. In *ACM SIGSPATIAL GIS*, pages 34–43, 2011.

[7] Josh Jia-Ching Ying, Eric Hsueh-Chan Lu, Wang-Chien Lee, Tz-Chiao Weng, and Vincent S. Tseng. Mining user similarity from semantic trajectories. In *LBSN*, pages 19–26, 2010.