IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 6, 2018, accepted December 17, 2018, date of publication December 24, 2018,
date of current version January 29, 2019.

Digital Object Identifier 10.1109/ACCESS.2018.2889475

Approximate Similarity Measurements on Multi-Attributes Trajectories Data

PAN XIAO[1,2], MA ANG[1,2], ZHANG JIAWEI[3], AND WU LEI[1,2]

[1]School of Economics and Management, Shijiazhuang Tiedao University, Shijiazhuang 050043, China
[2]Key Research Base for Humanities and Social Sciences in Hebei Province, Shijiazhuang Tiedao University, Shijiazhuang 050043, China
[3]Department of Computer Science, Florida State University, Tallahassee, FL 32306, USA

Corresponding author: Pan Xiao (smallpx@stdu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61303017, in part by the Natural Science Foundation of Hebei Province of China under Grant F2018210109, in part by the Hebei Provincial Education Department under Grant ZD2018040, in part by the Fourth Outstanding Youth Foundation of Shijiazhuang Tiedao University under Grant Z661250444, and in part by the Introducing Overseas Student under Grant C201822.

ABSTRACT With the development of global positioning technology, sensor networks, and smart mobile terminal, a large number of trajectory data are accumulated. Trajectory data contains a wealth of information, including spatiality, time series, and other external descriptive attributes (i.e., features, travelling mode, and so on). Trajectory analysis and mining show the great value. The research of trajectory similarity measurement is the basis of trajectory data management and mining, which plays an important role in trajectory computing. Most trajectory similarity work only focuses on the spatial-temporal features. The addition of multi-attributes to the trajectories changes the trajectory similarity. However, there are few researches focusing on multi-attributes trajectory similarity. In this paper, we propose two novel trajectory similarity measurements, i.e. maximum-minimum trajectory distance and sum of minimum trajectory distance and analyze the correlation among the spatial-temporal similarity and textual similarity. Finally, the measurement validity is verified and visualized through clustering, by both a simulation dataset and a real dataset.

INDEX TERMS Trajectory similarity measurement, clustering, big trajectory data, trajectory computing.

I. INTRODUCTION

With the development of wireless communication technology, global positioning system, and smart mobile terminal, trillion byte or even peta byte trajectory data accumulates rapidly. DiDi[1] announced that more than 70TB spatial-temporal data is generated per day and the processing data size is up to 4500TB daily [1]. In the meantime, with the popularity of social media, such as Twitter, Facebook, and Foursquare, a wealth of external information are embedded into the trajectories. As a result, the raw trajectories are enriched by a wide variety of semantics. For example, the geo-textual objects in trajectories are associated with the location name, the point category (i.e., restaurants, museums) and etc. We unify the semantics[2] as the attributes of the trajectories into the form of a collection of textual keywords. The trajectories including the locations, timestamps, and descriptive attributers are called multi-attributes trajectories.

There have been some literatures [2]–[5] on multi-attribute trajectories. The similarity study is basic and important in trajectory management system. For instance, a similar trajectory can be recommended to a user as a tour reference in trajectory recommendation system. For another example, similarity trajectories can be gathered together to find out the trajectory pattern. The existing trajectory similarity measurements can be divided into two categories w.r.t. the trajectory representation. If a trajectory is represented as a segments sequence, the well-known metrics include Line Segments Distance [6], One Way Distance [7], Edit Distance on Segment [8], etc. If a trajectory is a points sequence, the trajectory similarity is defined as Lp-norm Distance [9], LCSS [10], and ERP [10], Hausdorff Distance [11], Fréchet Distance [12], DTW [13], etc. However, the above measurements take only the spatial information and the temporal information into account. The textual similarity is not included. The trajectory similarity changes when the textual similarity is taken into account.

Let's consider the three trajectories in Figure 1 as an example. Figure 2 shows the details of the three trajectories. The

[1]https://www.didiglobal.com/

[2]In general, the concept of semantics is hierarchical. However, our paper only considers the fine-grained information at the lowest level.

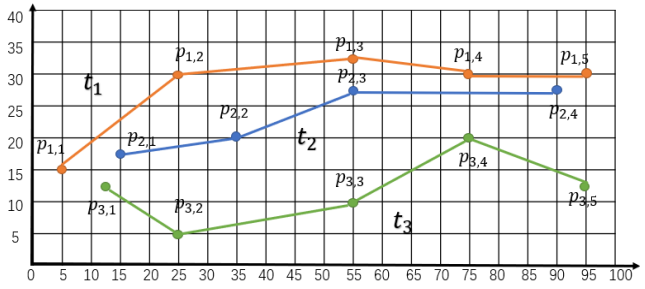

FIGURE 1. Trajectory example.

	Point id	Location	Timestamp	Multi-attribute		
				Text 1	Text 2	Text 3
t_1	$p_{1,1}$	(5,15)	0	Breakfast	Coffee	Spot
	$p_{1,2}$	(25,30)	3	Gay	Bar	Restaurant
	$p_{1,3}$	(55,33)	6	Dessert	Shop	Bar
	$p_{1,4}$	(75,30)	9	Science Museum	Bar	
	$p_{1,5}$	(95,30)	10	Asian Restaurant		
t_2	$p_{2,1}$	(15,18)	0	**Bookstore**	**American Restaurant**	
	$p_{2,2}$	(35,20)	3	**Department Store**		
	$p_{2,3}$	(55,23)	6	Dessert	Shop	**Restaurant**
	$p_{2,4}$	(90,28)	10	**Doctor's Office**	**Bookstore**	**Pub**
t_3	$p_{3,1}$	(13,13)	1	Breakfast	Coffee	
	$p_{3,2}$	(25,5)	2	Bar	Restaurant	
	$p_{3,3}$	(55,10)	6	Dessert	Shop	Bar
	$p_{3,4}$	(75,20)	9	Science Museum	Bar	
	$p_{3,5}$	(95,13)	10	**Restaurant**		

FIGURE 2. Trajectory information.

bold texts in t_2 and t_3 are different from the keywords of t_1. In previous work, t_1 and t_2 are the most similar trajectories considering the spatial-temporal similarity. However, we can observe that the keywords on t_1 are almost consistent with the ones on t_3. On the other hand, the keywords of t_2 are significantly different from the ones of t_1. Thus, t_1 and t_3 are most similar when the textual similarity is blended into the similarity measurement.

Reference [5] is a representative work measuring the similarity on the spatial dimension, the temporal dimension, and the textual dimension all at once. For the spatial fold, the spatial distance is the average number of matching pairs of two trajectories whose spatial distances are less than a threshold. For the temporal fold, the ratio of the common time period to the total time length of the two trajectories is used as the temporal distance. For the textual fold, the textual distance is the average number of exact textual matching pairs of two trajectories. However, reference [5] suffers from two main drawbacks. First, reference [5] only regards two keywords are similar when the two words are exactly same. This is not practical. Uncertain data or misspell data indeed exist in the real world (i.e., theatre and theater). Second, reference [5] does not consider trajectories time alignment, where the spatial dimension and the temporal dimension are separated. As we know, the two dimensions depend on each other.

To overcome the limitations, we propose two approximate similarity measures on trajectories (that is, MMTD and SUMTD). Both measurements resolve the problem of trajectory time alignment, and support approximate similarity using the edit distance. Besides, we analyze the correlations among the spatial-temporal similarity and the textual similarity using a real dataset. We find that the spatial-temporal similarity and the textual similarity are weak correlation. In order to verify the effectiveness of MMTD and SUMTD, we apply the two similarity measurements in a classical clustering algorithm (i.e. k-medoids) and visualize the clustering results.

To sum up, the main contributions of this paper are as follows:

- We propose two trajectory similarity measures, MMTD and SUMTD, for multi-attributes trajectories which support approximate similarity.
- To the best of our knowledge, we are the first to prove that the spatial-temporal similarity and the textual similarity are weak correlated with each other using the correlation analysis.
- We demonstrate the effectiveness of our proposed similarity metrics in the k-medoids clustering using a simulated dataset and a real dataset.

The rest of the paper is organized as follows. Related work is reviewed in Section II. Section III introduces the preliminary. Section IV and Section V propose the trajectory similarity measurements. The correlations of the spatial-temporal similarity and the textual similarity are analyzed in Section VI. Section VII reports the experiment evaluation. Finally, Section VIII concludes this paper and discusses the future work.

II. RELATED WORK

The similarity measurements on multi-attributes trajectories can be divided into three types: spatial-temporal similarity, spatial-textual similarity, and spatial-temporal-textual similarity.

A. SPATIAL-TEMPORAL SIMILARITY MEASUREMENTS

Reference [12]–[15] consider the spatial-temporal similarity of trajectories. Longest Common Subsequence (LCSS) [14] takes the number of matching points pairs as the trajectory distance ignoring the far-away points. LCSS needs a manual matching threshold to determine the distance. Discrete Fréchet Distance (DFD) [12] considers the locations and ordering of the points along the curve using the "shortest dog leash distance". Dynamic Time Warping (DTW) [13] allows repeating some points to achieve the best alignment. Different from the above work, [15] computes the spatial distance by Euclidean distance and temporal distance by timestamp difference using an exponential function. The linear combination of the spatial and temporal distances is the final similarity.

B. SPATIAL-TEXTUAL SIMILARITY MEASUREMENTS

Reference [16]–[18] concerned both the spatial and textual similarity on two trajectories. Reference [16] computes the spatial distance using Euclidean distance and the textual distance by Edit distance. In [17], the spatial distances consist of trajectory geometric center distance, trajectory length difference, and direction. The textual distances are based on the longest common subsequence of visited points, which only consider the full match. Reference [18] considers the two trajectories are similar if they share a common points sequence with the similar travel time. Their approach is different from the existing similarity measures due to considering the visit frequency.

C. SPATIAL-TEMPORAL-TEXTUAL SIMILARITY MEASUREMENTS

There are few works [4], [5] consider the trajectory distance from the spatial aspect, the temporal aspect and the textual aspect all at once. The similarity measurement proposed by [5] is a linear combination of the Euclidean distance, the time interval intersection, and the number of full matching pair. Reference [4] improves [5] by defining a new textual similarity by considering the hierarchy of the label semantics. A category tree is defined for the text classification, and different weights are assigned to the nodes for establishing the importance. In this paper, we propose two new similarity measurements for multi-attributes trajectories. The two new similarities support trajectory time alignment and the approximate textual similarity.

III. PRELIMINARY

Definition 1 (Multi-Attributes Trajectory): A multi-attributes trajectory t is a points sequence $\langle p_{1,1}, p_{1,2}, \ldots, p_{1,n}\rangle$. Each point is in the form of a tuple $\langle l,\ i,\ ky\rangle$, representing that the user locates at $l = (x, y)$ at timestamp i with a textual collection ky such as travel modes, weather conditions or road conditions, etc.

Definition 2 (Trajectory Similarity): Given two trajectories t_1, t_2, the similarity of the two trajectories is

$$S(t_1, t_2) = 1 - (\omega_1, \omega_2, \omega_3)\begin{pmatrix} dist_1(t_1, t_2) \\ dist_2(t_1, t_2) \\ dist_3(t_1, t_2) \end{pmatrix} \quad (1)$$

where $\omega_i (i = 1, 2, 3)$ and $0 \le \omega_i \le 1$ representing the spatial, temporal, and textual similarity weights respectively. $\mathrm{dist_i}(t_1, t_2)$ $(i = 1, 2, 3)$ is the spatial (temporal and textual) distance between t_1 and t_2, and $\mathrm{dist_i}(t_1, t_2)$ $(i = 1, 2, 3)$ is in [0, 1]. The larger S is, the more similar the two trajectories are.

IV. MINTD: TRAJECTORY SIMILARITY MEASUREMENT BASED ON DISCRETE POINTS

A. ANALYSIS

Intuitively, the trajectory similarity can be defined as an aggregate distance of any two points on the trajectory. In particular, there are four ways.

1) **The minimum point-to-point distance**: The simplest method is to use the minimum distance of any two points on the trajectories. That considers the best case. For each point $p_{1,\mathrm{i}}$ on t_1, we can find the nearest point $p_{2,\mathrm{j}}$ on t_2. Then, the minimum distance between all points pairs on the two trajectories is the distance between t_1 and t_2. Formally,

$$dist_i(t_1, t_2) = \min_{p_{1,i}\in t_1, p_{2,j}\in t_2}\left\{d_i(p_{1,i}, p_{2,j})\right\} \quad (2)$$

where $\mathrm{d_i}(p_{1,\mathrm{i}}, p_{2,\mathrm{j}})$ is the distance function. For instance, we can use the Euclidean distance to compute the spatial distance, the timestamp difference to represent temporal distance, and the edit distance for textual distance. Then, in Figure 1, the spatial distance between t_1 and t_2 is $\mathrm{dist_1}(t_1, t_2)=\mathrm{d_1}(p_{1,5}, p_{2,4})=((95-90)^2+(30-28)^2)^{0.5}=5.4$, the temporal distance between t_1 and t_2 is $\mathrm{dist_2}(t_1, t_2)=\mathrm{d_2}(p_{1,1}, p_{2,1})=0$, and the textual distance between t_1 and t_2 is $\mathrm{dist_3}(t_1, t_2)=\mathrm{d_3}(p_{1,3}, p_{2,3})=4$. If the weight vector is (0.3, 0.4, 0.3), the similarity after normalization between t_1 and t_2 is 0.95.

2) **The maximum point-to-point distance**: We can also use the maximum distance of any two points on the two trajectories as the trajectory similarity, which pays attention to the worst case. That is, for each point $p_{1,\mathrm{i}}$ on t_1, the furthest point $p_{2,\mathrm{j}}$ on t_2 is found. Then, the maximum distance of all points pair is used as the distance between t_1 and t_2. That is,

$$dist_i(t_1, t_2) = \max_{p_{1,i}\in t_1, p_{2,j}\in t_2}\left\{d_i(p_{1,i}, p_{2,j})\right\} \quad (3)$$

In Figure 1, the spatial distance between t_1 and t_2 is $\mathrm{dist_1}(t_1, t_2)=\mathrm{d_1}(p_{1,1}, p_{2,4})=((90\text{-}5)^2+(28\text{-}15)^2)^{0.5}=86$, the temporal distance between t_1 and t_2 is $\mathrm{dist_2}(t_1, t_2)=\mathrm{d_2}(p_{1,1}, p_{2,1})=10$, and the textual distance between t_1 and t_2 is $\mathrm{dist_3}(t_1, t_2)=\mathrm{d_3}(p_{1,1}, p_{2,1})=24$. After the normalization, the similarity between t_1 and t_2 is 0.01.

3) **The sum-min distance** [15], [19], [20]: The sum of the minimum point-point distances focuses on the average-best case. That is, for each point $p_{1,\mathrm{i}}$ on t_1, we find the nearest point $p_{2,\mathrm{j}}$ on t_2. Then, the distance from t_1 to t_2 is formally represented as,

$$dis_i(t_1, t_2) = \sum_{p_{1,i}\in t_1} \min_{p_{2,j}\in t_2}\left\{d_i(p_{1,i}, p_{2,j})\right\} \quad (4)$$

Note that $\mathrm{dis_i}(t_1, t_2) \neq \mathrm{dis_i}(t_2, t_1)$. Then,

$$dist_i(t_1, t_2) = \frac{1}{2}\left(\frac{dis_i(t_1, t_2)}{|t_1|} + \frac{dis_i(t_2, t_1)}{|t_2|}\right) \quad (5)$$

The spatial distance between t_1 and t_2 is $\mathrm{dist_1}(t_1, t_2)$ $=10.51$, and the temporal distance between t_1 and t_2 is $\mathrm{dist_2}(t_1, t_2)=0.1$, and the textual distance between t_1 and t_2 is $\mathrm{dist_3}(t_1, t_2)=7.65$. Then, the final normalized similarity is 0.90.

4) **The sum-max distance:**Contrary to the sum-min distance, the average-worst case is evaluated in the fourth

TABLE 1. The spatial-temporal-textual similarity results by different cases.

	Case (1)	Case (2)	Case (3)	Case (4)
t_1, t_2	0.95 [2]	0.01 [3]	0.90 [2]	0.24 [3]
t_1, t_3	0.97 [1]	0.70 [1]	0.91 [1]	0.30 [1]
t_2, t_3	0.93 [3]	0.10 [2]	0.84 [3]	0.29 [2]

case. For each point $p_{1,\mathrm{i}}$ on t_1, we can find the furthest point $p_{2,\mathrm{j}}$ on t_2. Then, the sum of the maximum distance of all points on t_1 to the furthest points on t_2 is used as the distance from t_1 to t_2. That is,

$$dis_i(t_1, t_2) = \sum_{p_{1,i} \in t_1} \max_{p_{2,j} \in t_2} \left\{ d_i(p_{1,i}, p_{2,j}) \right\} \quad (6)$$

Then,

$$dist_i(t_1, t_2) = \frac{1}{2} \left(\frac{dis_i(t_1, t_2)}{|t_1|} + \frac{dis_i(t_2, t_1)}{|t_2|} \right) \quad (7)$$

Continuing to take Figure 1 as an example, $\mathrm{dist}_1(t_1, t_2)$= 68.38, $\mathrm{dist}_2(t_1, t_2)$= 8.32, and $\mathrm{dist}_3(t_1, t_2)$=18.15. The normalized similarity between t_1 and t_2 is 0.24.

The minimum point-to-point distance (i.e., Case (1)) and the maximum point-to-point distance (i.e., Case (2)) represent the best and the worst cases of two trajectories, which are sensitive to anomalies. Both the sum-min distance (i.e., Case (3)) and the sum-max distance (i.e., Case (4)) consider the average situation. The sum-min distance is more representative than the sum-max distance. For instance, Table 1 shows the similarities between any two trajectories in Figure 1 with the different similarity metrics. Both Case (3) and Case (4) consider t_1, t_3 are most similar. The difference between the two cases is the similarity ranks of $\mathrm{S}(t_1, t_2)$ and $\mathrm{S}(t_3, t_2)$. In Figure 2, we observe that the texts of t_1 and t_3 are almost same, so the similarity difference between t_1, t_2 and t_3, t_2 depends mainly on the spatial-temporal similarities. From Figure 2, t_1, t_2 are more similar from the spatial-temporal aspect. In Case (3), $\mathrm{S}(t_1, t_2) > \mathrm{S}(t_3, t_2)$. However, Case (4) is the opposite. The drawback of Case (4) is that the spatial distance between t_1 and t_2 is dominated by the further points pair (i.e., $p_{1,1}$ and $p_{2,4}$). In an extreme case, two same trajectories overlap with each other. The spatial distance is the distance between the first point to the last point, which is unreasonable.

B. BASELINE: MINTD

According to the above analysis, we propose MINTD under the average-best case (i.e., Case (3)).

(Definition 3 (MINTD): Given two trajectories $t_1 = \{p_{1,1}, p_{1,2}, \ldots, p_{1,\mathrm{n}}\}$, $t_2 = \{p_{2,1}, p_{2,2}, \ldots, p_{2,\mathrm{n}}\}$, MINTD between t_1 and t_2 is (8), as shown at the bottom of the next page.

The spatial distance (denoted as d_1) is the Euclidean distance. The temporal distance (denoted as d_2) is determined by the timestamp difference of two points. The textual distance (denoted as d_3) is the edit distance of keywords associated with the two points.

We call MINTD as the base line since MINTD doesn't resolve the time alignment problem. As a result, the points pair for computing the spatial similarity, the temporal similarity and the textual similarity are not the same pair. The practicability of the method is limited. In Section V, we will introduce two time alignment aware trajectory similarity measurements.

V. MMTD AND SUMTD: TIME ALIGNMENT AWARE TRAJECTORY SIMILARITY MEASUREMENT

Time alignment is important for measuring the trajectory similarity. Before presenting the two new time alignment aware similarity measurements, we define anchor points and aligned points first.

Definition 4 (Anchor Points and Aligned Points): Given two trajectories $t_1 = \{p_{1,1}, p_{1,2}, \ldots, p_{1,\mathrm{n}}\}$, $t_2 = \{p_{2,1}, p_{2,2}, \ldots, p_{2,\mathrm{n}}\}$, $p_{1,\mathrm{i}} \in t_1$ and $p_{2,\mathrm{j}} \in t_2$ are defined as the **anchor points** at the timestamp k. Generally, an aligned point on one trajectory is defined as the previous point of anchor points on the other trajectory. That is, the **aligned point**on t_2 w.r.t. $\mathrm{p}_{1,\mathrm{i}}(\in t_1)$ is $p_{2,\mathrm{j}-1}$. Symmetrically, the aligned point on t_1 w.r.t. $\mathrm{p}_{2,\mathrm{j}}(\in t_2)$ is $p_{1,\mathrm{i}-1}$. Then, the aligned points pairs at timestamp k are $\{(p_{1,\mathrm{i}-1}, p_{2,\mathrm{j}-1}), (p_{1,\mathrm{i}-1}, p_{2,\mathrm{j}}), (p_{1,\mathrm{i}}, p_{2,\mathrm{j}-1})\}$.

For example, in Figure 1, assume $p_{1,2}$ and $p_{2,2}$ are the anchor points. Then, the aligned point on t_2 w.r.t $p_{1,2}$ on t_1 is $p_{2,1}$. The aligned point on t_1 w.r.t $p_{2,2}$ on t_2 is $p_{1,1}$. Then, the aligned points pairs are $\{(p_{1,1}, p_{2,1}), (p_{1,1}, p_{2,2}), (p_{1,2}, p_{2,1})\}$.

We use the aggregate distance between the aligned points and the anchor points as the trajectory similarity. MMTD uses the maximum of minimum distances between the aligned points and the anchor points (in Section V.A), which focuses on the worst-best case. SUMTD is the sum of the distances between the anchor points and the minimum distance between the aligned points (in Section V.B), which concerns the average case.

A. MMTD: MAXIMUM-MINIMUM DISTANCE

The trajectory is regarded as a whole. The keywords associated with each point constitute a word bag for the trajectory. The similarity of the two word-bags is the textual similarity of the two trajectories. The spatial-temporal similarities are considered together. Generally, the minimum distance of any two aligned points (i.e., $\min\{\mathrm{d}_1(p_{1,\mathrm{i}-1}, p_{2,\mathrm{j}-1}), \mathrm{d}_1(p_{1,\mathrm{i}-1}, p_{2,\mathrm{j}}), \mathrm{d}_1(p_{1,\mathrm{i}}, p_{2,\mathrm{j}-1})\}$) and the distance of anchor points $\mathrm{d}_1(p_{2,\mathrm{i}}, p_{2,\mathrm{j}})$ are computed respectively. Then, the distance of the two trajectories is the maximum one. Specially, for the starting point (e.g. $p_{1,\mathrm{i}}(i{=}1)$ of t_1), we compute the maximum distance between the points on t_2 with the starting point $(p_{1,1})$ on t_1, denoted $\max\{\mathrm{d}_1(p_{1,1}, p_{2,\mathrm{j}}), \mathrm{d}_1(p_{1,1}, p_{2,\mathrm{j}-1})\}$ $(j \in [2, m])$. We do the same calculation for the starting point $p_{2,1}$ on t_2. The temporal-spatial distance $\mathrm{dist}_{1,2}(t_1, t_2){=}\mathrm{d}_1(p_{1,\mathrm{i}}, p_{2,\mathrm{j}})$ $(i \in [2, n], j \in [2, m])$ is defined as following,

$$d_1(p_{1,i}, p_{2,j}) = \max \begin{cases} d_1(p_{1,i}, p_{2,j}) \\ \min \begin{cases} d_1(p_{1,i-1}, p_{2,j}) \\ d_1(p_{1,i}, p_{2,j-1}) \\ d_1(p_{1,i-1}, p_{2,j-1}) \end{cases} \end{cases} \quad (9)$$

If $i = 1, j = 1$, the distance is $d_1(p_{1,1}, p_{2,1})$. If $i = 1, j \in [2, m]$, the distance equals $\max\{d_1(p_{1,i}, p_{2,j}), d_1(p_{1,i}, p_{2,j-1})\}$. If $i \in [2, n]$, $j = 1$, the distance is $\max\{d_1(p_{1,i}, p_{2,j}), d_1(p_{1,i-1}, p_{2,j})\}$. Equation (9) is the famous discrete Fréchet Distance [12], where the temporal similarity is implied in the equation. In order to support approximate textual similarity, edit distance is employed in MMTD.

Definition 5 (MMTD): Given two trajectories $t_1 = \{p_{1,1}, p_{1,2}, \ldots, p_{1,n}\}$, $t_2 = \{p_{2,1}, p_{2,2}, \ldots, p_{2,n}\}$,

$$MMTD(t_1, t_2) = 1 - (\omega_{1,2}, \omega_3) \begin{pmatrix} dist_{1,2}(t_1, t_2) \\ dist_3(t_1, t_2) \end{pmatrix} \quad (10)$$

where $dist_{1,2}$ is the discrete Fréchet Distance, and the textual distance $dist_3$ is the Edit distance of the two words bags of the two trajectories.

B. SUMTD: TRAJECTORY SUM OF MINIMUM DISTANCE

MMTD is sensitive to outliers since MMTD is the distance between the two points. The average distance will help reducing the influence of outliers. Therefore, we propose SUMTD.

Definition 6 (SUMTD): Given two trajectories $t_1 = \{p_{1,1}, p_{1,2}, \ldots, p_{1,n}\}$, $t_2 = \{p_{2,1}, p_{2,2}, \ldots, p_{2,n}\}$, SUMTD between t_1, t_2 is

$$SUMTD(t_1, t_2) = 1 - (\omega_{1,2}, \omega_3) \begin{pmatrix} dist_{1,2}(t_1, t_2) \\ dist_3(t_1, t_2) \end{pmatrix} \quad (11)$$

$dist_3$ is the Edit Distance of the two word-bags for the two trajectories. The spatial-temporal distance $dist_{1,2}(t_1, t_2) = d'_1(p_{1,i}, p_{2,j})$ $(i \in [2, n], j \in [2, m])$ is computed by Equation (12),

$$d'_1(p_{1,i}, p_{2,j}) = d_1(p_{1,i}, p_{2,j}) + \min \begin{cases} d_1(p_{1,i-1}, p_{2,j}) \\ d_1(p_{1,i}, p_{2,j-1}) \\ d_1(p_{1,i-1}, p_{2,j-1}) \end{cases} \quad (12)$$

If $i = 1, j = 1$, the distance is $d_1(p_{1,1}, p_{2,1})$. If $i = 1$, $j \in [2, m]$, the distance equals to $d_1(p_{1,i}, p_{2,j}) + d_1(p_{1,i}, p_{2,j-1})$. If $i \in [2, n]$, $j = 1$, the distance is $d_1(p_{1,i}, p_{2,j}) + d_1(p_{1,i-1}, p_{2,j})$. Generally, the minimum distance of any two aligned points $\min\{d_1(p_{1,i-1}, p_{2,j-1}), d_1(p_{1,i-1}, p_{2,j}), d_1(p_{1,i}, p_{2,j-1})\}$ and the distance of anchor points $d_1(p_{2,i}, p_{2,j})$ are computed respectively. Then, the distance of two trajectories is the sum of the two distances obtained earlier. Similar to MMTD, the spatial-temporal similarities are considered together.

TABLE 2. The spatial-temporal-textual similarity rank by different measurement.

Rank	MINTD	MMTD	SUMTD
1	t_1, t_3	t_1, t_3	t_1, t_3
2	t_1, t_2	t_1, t_2	t_1, t_2
3	t_2, t_3	t_2, t_3	t_2, t_3

C. A TOY EXAMPLE

We continue to use the example in Figure 1 and Figure 2 as the toy example to elaborate MINTD, MMTD, SUMTD. The weight vector is (0.4,0.3,0.3) in Equation (1) and (0.5,0.5) in Equations (10,11). Table 2 shows that the spatial-temporal-textual similarity rank of any two trajectories among $\{t_1, t_2, t_3\}$. From Table 2, we observe that the lists of similarity rank are consistent though the similarity measurements are different. Specifically, t_1 and t_3 are the most similar. t_2 and t_3 are the least similar.

VI. CORRELATION ANALYSIS

Most of the existing work assume that the spatial similarity, the temporal similarity and the textual similarity are independent. However, to the best of our knowledge, none of the work gives a strict proof. In this section, we evaluate the relevance of the three similarities using a real dataset by the correlation analysis in statistics.

A. DATASETS

We use a real dataset scrawled from Foursquare [21], [22]. The real dataset obtains 49,062 users in New York City (NYC) and 31,544 users in Los Angeles (LA). Each check-in record contains users' ID, venues with the geo-locations, check-in timestamp, and tips. We align each user's check-in points by the timestamps, such that a user check-in trajectory is formed. The textual set for each venue is generated by the

$$MINTD(t_1, t_2) = (\omega_1, \omega_2, \omega_3) \cdot \begin{pmatrix} \frac{1}{2}\left(\frac{\sum_{p_{1,i} \in t_1} \min_{p_{2,j} \in t_2}\{d_1(p_{1,i}, p_{2,j})\}}{|t_1|} + \frac{\sum_{p_{2,j} \in t_2} \min_{p_{1,i} \in t_1}\{d_1(p_{1,i}, p_{2,j})\}}{|t_2|}\right) \\ \frac{1}{2}\left(\frac{\sum_{p_{1,i} \in t_1} \min_{p_{2,j} \in t_2}\{d_2(p_{1,i}, p_{2,j})\}}{|t_1|} + \frac{\sum_{p_{2,j} \in t_2} \min_{p_{1,i} \in t_1}\{d_2(p_{1,i}, p_{2,j})\}}{|t_2|}\right) \\ \frac{1}{2}\left(\frac{\sum_{p_{1,i} \in t_1} \min_{p_{2,j} \in t_2}\{d_3(p_{1,i}, p_{2,j})\}}{|t_1|} + \frac{\sum_{p_{2,j} \in t_2} \min_{p_{1,i} \in t_1}\{d_3(p_{1,i}, p_{2,j})\}}{|t_2|}\right) \end{pmatrix} \quad (8)$$

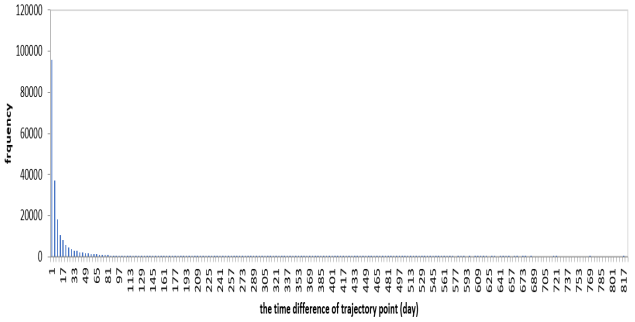

FIGURE 3. Trajectory time distribution (LA).

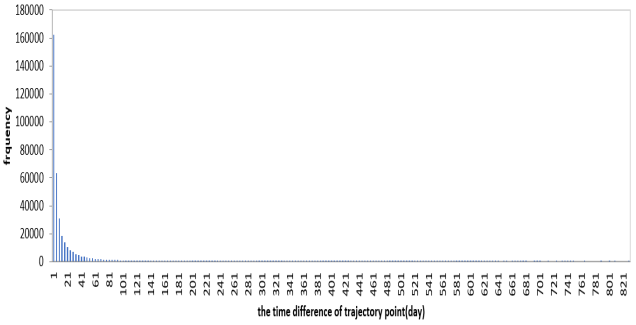

FIGURE 4. Trajectory time distribution (NYC).

TABLE 3. Statistics of real data sets.

	Before time division		After time division	
	LA	NYC	LA	NYC
Number of Trajectories	31,544	49,062	18,024	29,784
Number of Check-in points	267,579	424,649	157,931	247,660
Number of keywords	395,738	630,691	372,422	596,312
Max number of points/ trajectory	200	200	199	199
Min number of points/ trajectory	1	1	3	3
Average number of points/ trajectory	9	9	8.7	8.3

keywords in the tips associated with the venue. The number of textual keywords associated with a venue is between [0,18]. We select 3 texts randomly for each check-in venue as the textual attributes.

The variation of the timestamps difference of any two consecutive checked-in venues is large. Figure 3 and Figure 4 show the timestamps difference distribution. From the two figures, we can see almost 50% of the check-in timestamps difference are less than 14 or 15 days. Thus, we divide the trajectories by two weeks (i.e. 14 days). Then, the trajectories, whose length is greater than 2, are left. The statistical information of the trajectories before the time division and after the time division is shown in Table 3.

B. CORRELATION

1,200 trajectories are selected by the simple random sample. The spatial-temporal similarity and the textual similarity between any two trajectories are computed as MMTD and SUMTD (i.e., Definition 5 and Definition 6) respectively. The correlation matrix is generated from SPSS through Kendall test in correlation analysis. The Kendall test [23] is a non-parameter hypothesis test that uses the calculated correlation coefficients to test the statistical dependence of two variables. The Kendall correlation coefficient ranges from -1 to 1. When the absolute value of τ is closer to 1, the correlation of the two variables is strong. When the absolute value of τ is closer to 0, the two random variables are independent. We take the MMTD as an example. The result of spatial-temporal and textual correlation of MMTD is shown in Table 4. From Table 4, we observe that the correlation coefficient is 0.30, so the spatial-temporal similarity and the textual similarity are weak correlation.

TABLE 4. Spatial-temporal and textual correlation analysis(MMTD).

		Spatial-temporal similarity	Textual similarity
Spatial-temporal similarity	Correlation Coefficient	1.00	0.30**
	Sig.(2-tailed)		0.00
Textual similarity	Correlation Coefficient	0.30**	1.00
	Sig.(2-tailed)	0.00	

**.correlation is significant at the 0.01 level (2-tailed).

In order to determine whether the relevance of the samples can represent the relevance of the population, we performed the hypothesis test with the confidence level a=0.01.

Proof:

H_0: There is no significant effect between the spatial-temporal similarity and the text similarity.

H_1: There is a significant effect between the spatial-temporal similarity and the text similarity.

In Table 4, the two-sided test Sig = 0.00 <0.01. Therefore, we reject the hypothesis H_0. In other words, the spatial-temporal similarity and the text similarity have a significant effect.

Proof done.

VII. EXPERIMENTS

A. SETTINGS

In the experimental evaluation, we use a simulation dataset and the real dataset in Section VI. The simulation dataset is generated by the well-known Thomas Brinkhoff Generator. Each data includes an object status, an object id, a timestamp, and a location. A set of keywords is assigned to a simulated object based on the text distribution of the real data set. We generated 1,290 trajectories including 15,181 points. The weights of spatial-temporal similarity and textual similarity are (0.5, 0.5) in default.

We implement all the experiments using Java on a PC with an Intel Core i7-7500U 2.90GHz processor. MINTD, MMTD, SUMTD and MSM are compared. MINTD, MMTD and SUMTD are our proposed methods. MSM is revised from [5] with a specific distance instead of the number of

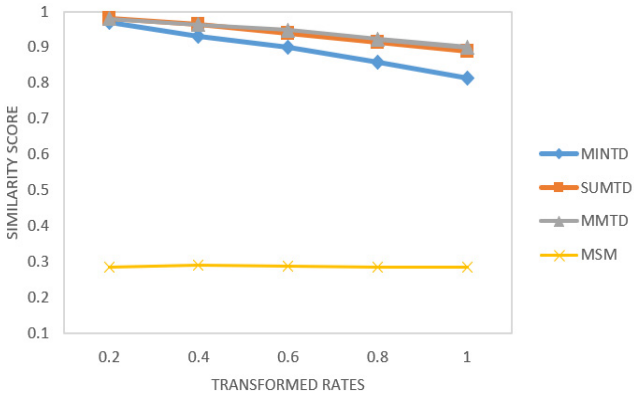

FIGURE 5. Addition of interpolated points.

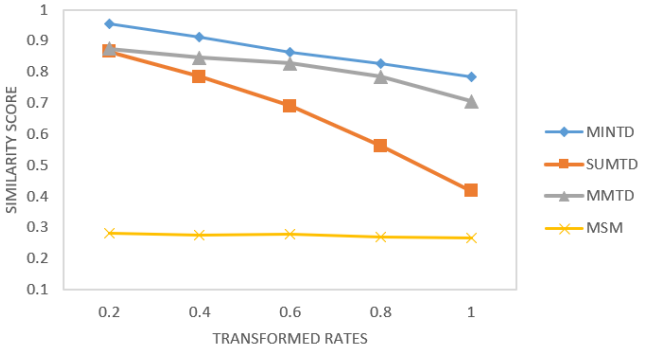

FIGURE 6. Addition of random points.

matching pairs. Finally, we use the classical clustering techniques to visualize the similarity of trajectories.

B. SIMULATED EXPERIMENTS

We choose 10 trajectories randomly from 1,290 simulated trajectories as the seed trajectories. 50 trajectories are generated by the trajectory transformation approach in [5] with the different transformed rates and the transformed type (e.g. addition of interpolated points, addition of random points, points removing, points order change, and points replacement). For example, if we remove points from the seed trajectory with 10 points and $r = 0.2$, 2 points will be removed from the seed trajectory. Comparing the similarity between the seed trajectory and the transformed trajectory is for evaluating the impact of every change on the different similarity measurement.

Figure 5 shows the average similarity change trend when r interpolated points are added. We randomly select two consecutive points from the seed trajectory and use the linear interpolation method to add a new point. The texts of the new point take the intersection of the keywords from the two consecutive points. After inserting interpolated points, we expect that the similarity of two trajectories are high and the similarity changes little with r increasing. From Figure 5, we observe that the similarity scores of MINTD, MMTD, SUMTD decrease, while the similarity score of MSM remains stable as the transformed rate increases. Since MSM uses the most matched points to compute the distances without time alignment, inserting interpolated points to the trajectories has no effect on MSM. However, since MSM requires the exact match for calculating text similarity, the similarity scores are lowest among the four methods. The similarity scores of MMTD, SUMTD and MINTD decrease resulting from the increase of the spatial distances. For the three measurements, the textual similarities between the seed trajectory and the transformed trajectory doesn't change since the whole word bags for two trajectories are same. The spatial distances increase as the result of the increased distance between the interpolated points and the points on the seed trajectories. Comparing with MMTD and SUMTD, the change of similarity scores of MINTD is most obvious. That is because MMTD and SUMTD consider the time alignment, only the distances between aligned points and the anchor points contribute to the final similarity. However, MINTD is determined by the distances between any two points on the trajectories. The distance from the transformed trajectory to the seed trajectory increase obviously.

Figure 6 shows the changes of the average trajectory similarity with random points being added into the seed trajectory. The locations of new points are generated from the spatial range and the temporal range randomly. Three texts of the new points are randomly selected from the keyword's vocabulary expect the keywords set on the trajectory. From Figure 6, the trajectory similarities of MMTD, SUMTD and MINTD decrease obviously with r increasing. SUMTD is most sensitive to the random points insertion. The change trend of MMTD is more sensitive than the one of MINTD since MMTD is determined by the maximum one between anchors distance and aligned points distances. However, MINTD uses the best cases among any points pair. The large distances generated by the random points to the seed trajectories are ignored. The similarity scores of MSM are smallest among all the measurements. Meanwhile, MSM fails to respond to similarity change with the random points being inserted.

Figure 7 shows the changes of the average trajectory similarity with removing points randomly from the seed trajectories. Removing points indicate that both the trajectory length and the trajectory shape change. Thus, the changes on the similarity should be distinct. Figure 7 shows the trajectory similarity decreases as r increasing. When r reduces to 1, the similarity decreases to 0, since there is no points on the transformed trajectory. When $r \leq 0.8$, the curve of MINTD and MSM decrease smoothly. In contrast, SUMTD and MMTD change sharply. The reason is the same as the previous analysis. Both MSM and MINTD use the distances between any two points, which slow the reducing trend. With the points removing, the aligned points pairs and the anchors pairs change. Thus, SUMTD and MMTD grasp the change. However, the changes of SUMTD is slightly obvious than the one of MMTD.

In Figure 8, we verify the four measurements under the points order change. The timestamps of the new trajectories

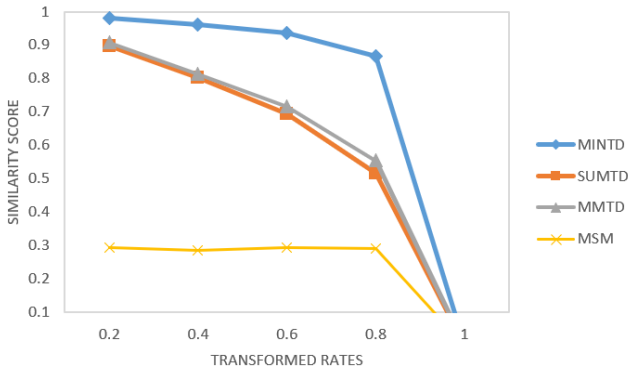

FIGURE 7. Removing points.

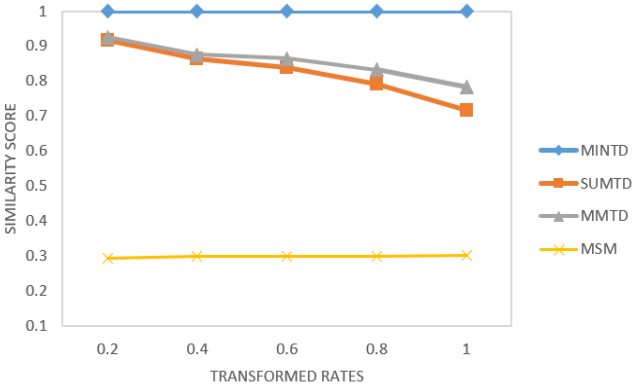

FIGURE 8. Points order change.

are generated through switching the timestamps of several points from the seed trajectories randomly, while the geographic locations and the keywords remain same. With points order change, the trajectory shapes change as well. From Figure 8, we observe that MINTD and MSM don't change with different transformed rates. That is because both of them ignore the time alignment. Furthermore, MINTD regards two trajectories are same (i.e., the similarity scores are 1). Since MINTD separates the relation between the location, the keyword and the timestamp on a point, the best similarity is counted. Thus, points order changes have no effect on MINTD, which is not practical. Though MSM also concerns the best case, the requirement for keywords exact match lowers down the similarity score. The curve of MSM is below the other three methods. The similarity scores of SUMTD and MMTD are similar when r is less than 0.6. SUMTD shows better performance than MMTD when r becomes large.

In Figure 9, we replace the points in the seed trajectory with random points into a new trajectory. In this case, the trajectory length remains the same, while both the trajectory shape and the location changes. Comparing the curves of MSM in Figure 5 to Figure 9, the similarities of MSM almost are same. That indicates that MSM hardly perceives any changes on the seed trajectory. The similarities of MINTD, SUMTD and MMTD decrease linearly with r increasing. SUMTD outperforms the other two measurements. In Figure 6 and

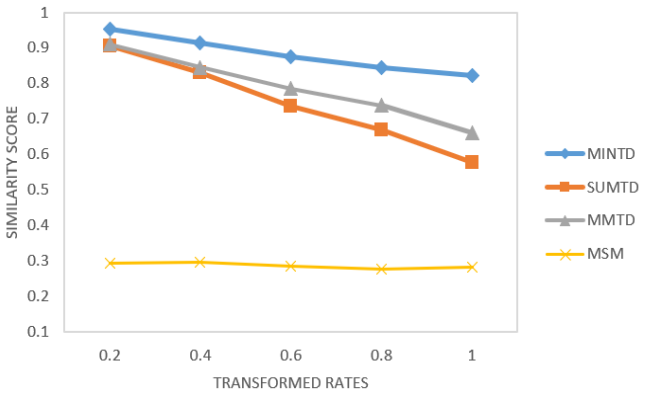

FIGURE 9. Points replacement.

Figure 9, the new trajectories both contain random points. The difference between them is the trajectory length. We can see that SUMTD changes more obviously in Figure 6. SUMTD is the distance sum between all points pairs after repeating some points. Hence, SUMTD is more sensitive on trajectory length. However, the similarity scores of MMTD in both figures are similar. We can see that MMTD changes a lot in Figure 9. That is because MMTD is determined by the distance between two points. The new random points deteriorate the similarity score of MMTD.

C. VISUALIZATION WITH THE SIMULATED DATA

As we know, the k-medoids [24] algorithm is immune to outliers since it improved from k-means and chooses the objects in clusters as the center. Hence, we choose the k-medoids to visualize the effectiveness of the similarity measurements with the 1,290 simulated trajectories. We use the silhouette coefficient [24] to specify the parameter k, which is widely used especially when the ground truth of a dataset is not available. Figure 10 and Figure 11 show the cluster result of the simulated data. One color represents one cluster. By default, $k = 3$.

In Figure 10, we only consider the spatial-temporal similarity. The weights are set as $w_{1,2} = 1$ and $w_3 = 0$ in SUMTD and MMTD, and the weights are set as $w_1 = w_2 = 0.5$ and $w_3 = 0$ in MINTD and MSM. Since the spatial-temporal locality, the objects with large spatial-temporal similarities tend to be clustered together. We observe that the cluster are similar in SUMTD and MMTD, which shows the explicit spatial distribution. However, the clusters in MINTD and MSM are not distinguishable.

Figure 11 shows the clusters results after the addition of the text similarity. We observe that the clusters generated by MINTD and MSM are similar in Figure 10 (a) (d) and Figure 11 (a) (d). Since MINTD and MSM separate the spatial similarity computing, the temporal similarity computing and the textual similarity computing, the points which contributes to the final similarity are not the same pairs. As a result, the clusters in MINTD and the ones in MSM distribute over the whole space.

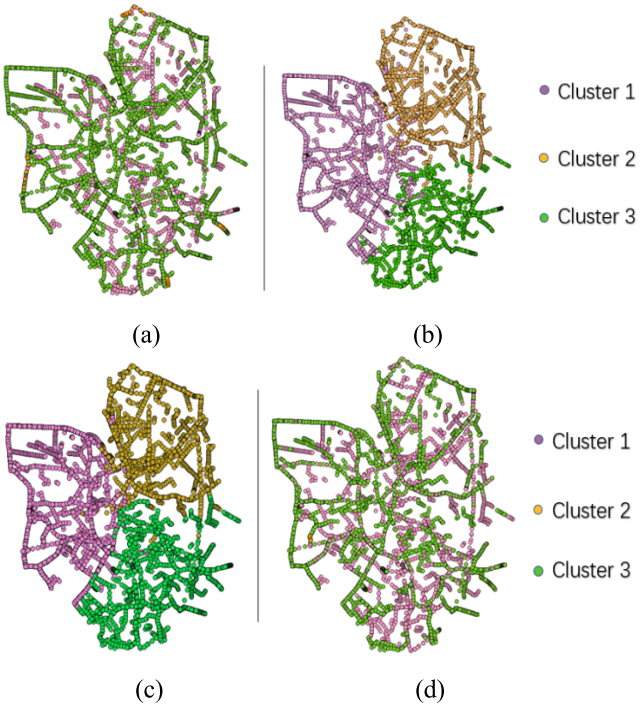

FIGURE 10. Trajectory spatial-temporal clustering (k=3). (a) MINTD. (b) SUMTD. (c) MMTD. (d) MSM.

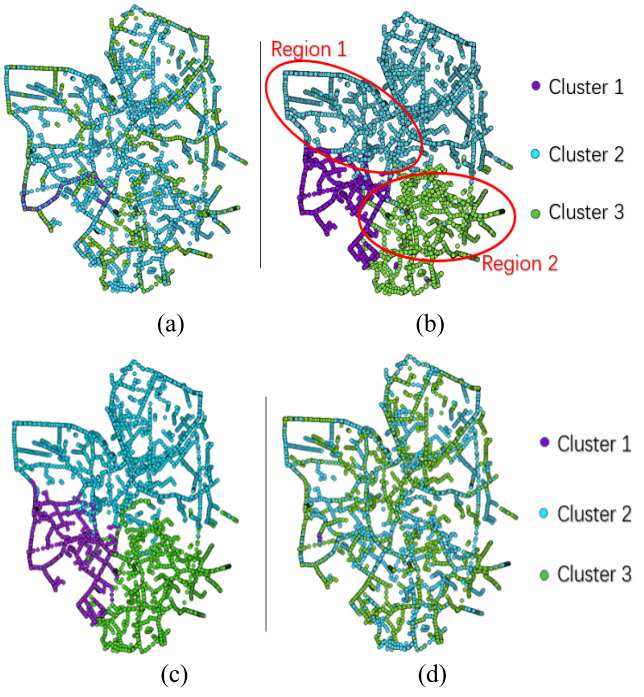

FIGURE 11. Trajectory spatial-temporal-textual clustering (k=3). (a) MINTD. (b) SUMTD. (c) MMTD. (d) MSM.

D. VISUALIZATION WITH THE REAL DATA

In this section, we apply k-medoids with the real dataset with $k = 3$. We select two trajectory sets from LA. The trajectory length in one set is 3, and the one in the other set is between 3 to 15. Similar to the results in Figure 10(a)(d) and Figure 11(a)(d), MINTD and MSM is weak distinguish. Hence, in this section, we only visualize the results of MMTD and SUMTD. Figure 12 and Figure 13 show the clustering results. Each color (i.e., green, red, yellow) represents a cluster. In order to evaluate similarity distribution of the clusters, we also tested the discrete coefficient.

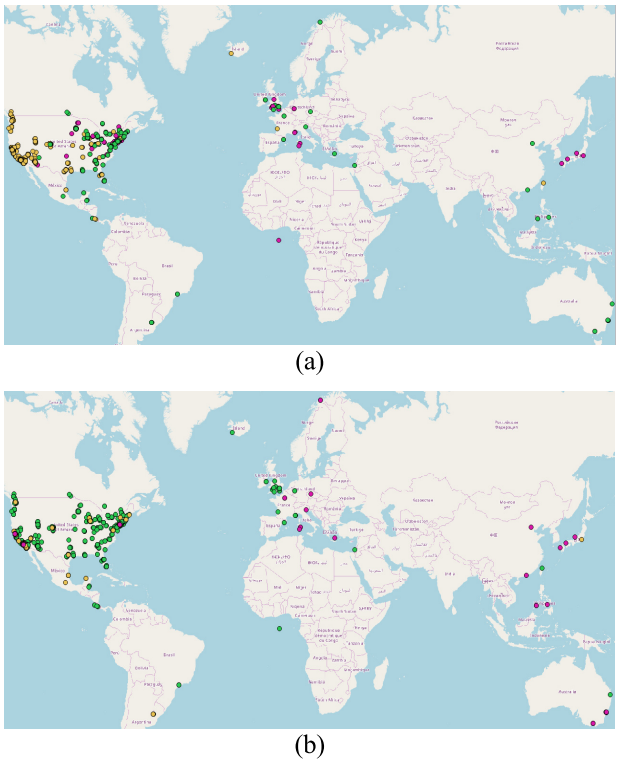

FIGURE 12. Clusters of trajectories with length equal to 3. (a) SUMTD. (b) MMTD.

Figure 12(a) and Figure 12 (b) show the clustering results of trajectories with equal length (i.e. 3) under SUMTD and MMTD, respectively. We observe that the distribution of clusters is a little different. The right sides of Figure 12(a) and Figure 12(b) are very sparse and almost in green and red. In the left side, the distribution of the green, red and yellow clusters of Figure 12(a) are more uniform, while in Figure 12 (b), the distribution of green clusters accounts for a large proportion. The average discrete coefficients of SUMTD and MMTD are the same (i.e., 0.72), though it seems that the clusters in Figure 12(a) is more compact than the ones in Figure 12 (b). The same discrete coefficients is because the trajectories in the green clusters in Figure 12(a) are more dispersed.

Figure 13 (a) and Figure 13(b) show the clustering results of SUMTD and MMTD with trajectory length from 3 to 15. We observe that the distribution of clusters is similar, especially for the green and yellow clusters. The red cluster in Figure 13(a) is sparser than the red cluster in Figure 13(b). The average discrete coefficients of SUMTD and MMTD are 0.76 and 0.7 respectively. MMTD shows better stability than SUMTD when the length of the trajectories varies.

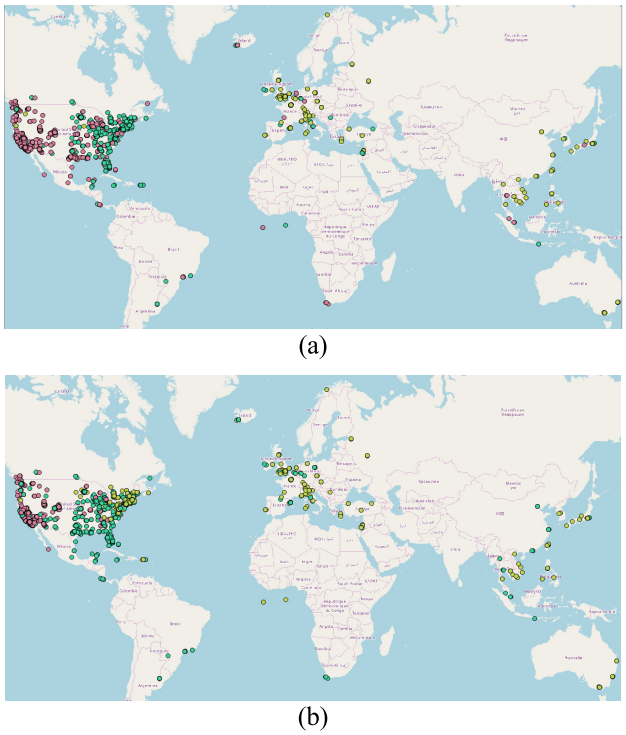

FIGURE 13. Clusters of trajectories with length between 3 to 15. (a) SUMTD. (b) MMTD.

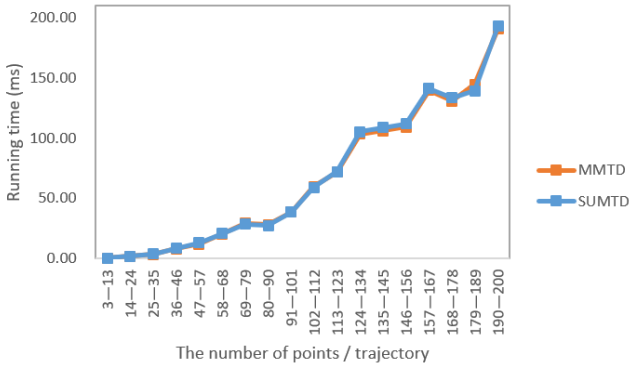

FIGURE 14. The performance test of trajectory similarity methods.

E. EFFICIENCY

In this section, we tested the performance with varying trajectory size. The trajectory size is the points number in a trajectory. We increase the trajectory size from [3, 13] to [190, 200]. Taking [3, 13] as example, it means the minimum number of points in a trajectory is 3 and the maximum one is 13. Figure 14 shows the average running time of MMTD and SUMTD. We observe that the running time increases when the average trajectory size (points number) increases, which mainly results from the spatial-temporal distance computation. The running time of the two methods are almost same, since both of them require traversing the whole trajectory. However, even if the trajectory size becomes 190-200, the average running time is only 200ms.

VIII. CONCLUSION

The booming social media enrich the trajectories with multi-attributions including spatial information, temporal information and other external information. However, most of existing works only focus on the similarity measures on the spatial-temporal feature. In this paper, we proposed two trajectory similarity measurements that measure the spatial-temporal-textual trajectory similarity at the same time. MMTD evaluates the worst of the best cases of trajectory, while SUMTD is the average similarity of trajectories. Both of MMTD and SUMTD resolve the problem of trajectory time alignment and support approximate similarity for multi-attributes trajectories. After the series of the experiments, SUMTD is most effective to various trajectory changes. We also prove that the temporal-spatial similarity and the semantic textual similarity are weak correlation. In our future work, we will work on the new similarity measurement on the road network employing multi-source data fusion and supporting a hierarchal semantics among trajectories into a city. We also will apply our proposed measurements in trajectory recommendation and trajectory mining.

REFERENCES

[1] D. Chuxing, "DiDi aids urban regulators with big data," DiDi, Beijing, China, May 2017. [Online]. Available: http://www.didichuxing.com/en/press-news/wu6uiayf.html

[2] J. Xu, H. Lu, and R. H. Güting, "Range queries on multi-attribute trajectories," *IEEE Trans. Knowl. Data Eng.*, vol. 30, no. 6, pp. 1206–1211, Jun. 2018.

[3] C. A. Ferrero, L. O. Alvares, and V. Bogorny, "Multiple aspect trajectory data analysis: Research challenges and opportunities," in *Proc. 17th Brazilian Symp. Geoinformatics*, Campos do Jordão, Brazil, 2016, pp. 56–67.

[4] F. J. M. Arboleda, S. R. Fernández, and V. Bogorny, "Towards a semantic trajectory similarity measuring," *Indian J. Sci. Technol.*, vol. 10, no. 18, pp. 1–14, 2017.

[5] A. S. Furtado, D. Kopanaki, L. O. Alvares, and V. Bogorny, "Multidimensional similarity measuring for semantic trajectories," *Trans. GIS*, vol. 20, no. 2, pp. 280–298, Apr. 2016.

[6] Y. Chen, K. Jiang, Y. Zheng, and C. Li, "Trajectory simplification method for location-based social networking services," in *Proc. Int. Workshop Location Based Social Netw.*, Seattle, WA, USA, 2009, pp. 33–40.

[7] B. Lin and J. W. Su, "One way distance: For shape based similarity search of moving object trajectories," *Geoinformatics*, vol. 12, no. 2, pp. 117–142, Jun. 2008.

[8] M. Xie, "EDS: A segment-based distance measure for sub-trajectory similarity search," in *Proc. Int. Conf. Manage. Data*, Snowbird, UT, USA, 2014, pp. 1609–1610.

[9] L. Chen and R. Ng, "On the marriage of lp-norms and edit distance," in *Proc. 13th Int. Conf. Very Large Data Bases*, Toronto, ON, Canada, 2004, pp. 792–803.

[10] H. Wang, H. Su, K. Zheng, S. Sadiq, and X. Zhou, "An effectiveness study on trajectory similarity measures," in *Proc. 24th Australas. Database Conf.*, Adelaide, SA, Australia, 2013, pp. 13–22.

[11] M. R. Evans, D. Oliver, and S. Shekhar, "Fast and exact network trajectory similarity computation: A case-study on bicycle corridor planning," in *Proc. 2nd ACM SIGKDD Int. Workshop Urban Comput.*, Chicago, IL, USA, 2013, pp. 1–8.

[12] B. Tang, M. L. Yiu, K. Mouratidis, and K. Wang, "Efficient motif discovery in spatial trajectories using discrete Fréchet distance," in *Proc. 20th Int. Conf. Extending Database Technol.*, Venice, Italy, 2017, pp. 378–389.

[13] E. J. Keogh and M. J. Pazzani, "Scaling up dynamic time warping for datamining applications," in *Proc. 6th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Boston, MA, USA, 2000, pp. 285–289.

[14] M. Vlachos, G. Kollios, and D. Gunopulos, "Discovering similar multidimensional trajectories," in *Proc. 18th Int. Conf. Data Eng.*, San Jose, CA, USA, Feb./Mar. 2002, pp. 673–684.
[15] S. Shang, L. Chen, and Z. Wei, "Trajectory similarity join in spatial networks," *Proc. VLDB Endowment*, vol. 10, no. 11, pp. 1178–1189, Aug. 2017.
[16] B. Zheng, N. J. Yuan, K. Zheng, X. Xie, S. Sadiq, and X. Zhou, "Approximate keyword search in semantic trajectory database," in *Proc. 31th IEEE Int. Conf. Data Eng.*, Seoul, South Korea, Apr. 2015, pp. 975–986.
[17] H. Liu and M. Schneider, "Similarity measurement of moving object trajectories," in *Proc. 3rd ACM SIGSPATIAL Int. Workshop GeoStreaming*, Redondo Beach, CA, USA, 2012, pp. 19–22.
[18] X. Xiao, Y. Zheng, Q. Luo, and X. Xie, "Inferring social ties between users with human location history," *J. Ambient Intell. Humanized Comput.*, vol. 5, no. 1, pp. 3–19, Dec. 2012.
[19] A. Ismail and A. Vigneron, "A new trajectory similarity measure for GPS data," in *Proc. ACM Sigspatial Int. Workshop Geostreaming*, Bellevue, WA, USA, 2015, pp. 19–22.
[20] Z. Chen, H. T. Shen, X. Zhou, Y. Zheng, and X. Xie, "Searching trajectories by locations: An efficiency study," in *Proc. Int. Conf. Assoc. Comput. Mach. Special Interest Group Manage. Data*, Indianapolis, IN, USA, 2010, pp. 255–266.
[21] J. Bao, Y. Zheng, and M. F. Mokbel, "Location-based and preference-aware recommendation using sparse geo-social networking data," in *Proc. SIGSPATIAL Int. Conf. Adv. Geographic Inf. Syst.*, Redondo Beach, CA, USA, 2012, pp. 199–208.
[22] L. Y. Wei, Y. Zheng, and W. C. Peng, "Constructing popular routes from uncertain trajectories," in *Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Beijing, China, 2012, pp. 195–203.
[23] M. Fisz, *Probability Theory and Mathematical Statistics*. Moscow, Russia: Mir Publishers, 1986.
[24] D.-W. Choi and C.-W. Chung, "A K-partitioning algorithm for clustering large-scale spatio-textual data," *Inf. Syst.*, vol. 64, pp. 1–11, Mar. 2017.

PAN XIAO was born in Xingtai, Hebei, China, in 1981. She received the B.S. and M.S. degrees in computer science and technology from the University of Yanshan, in 2003 and 2006, respectively, and the PhD. degree in computer application from the Renmin University of China, Beijing, China, in 2010. In 2008, she was a Research Assistant with the Department of Computer Science, Hong Kong Baptist University. From 2015 to 2016, she visited the lab supervised by Prof. Philips, who is with the University of Illinois at Chicago, USA. She is currently an Associate Professor with Shijiazhuang Tiedao University. She has published more than 20 papers in journals and academic conferences, which are indexed by SCI and EI. Her research interests include data management, mobile computing, and privacy protection. She is a CCF Member. She was a recipient of the Talent Project of Hebei Province (third level) and Beijing Science and Technology Award.

MA ANG was born in Shijiazhuang, Hebei, China, in 1992. She received the B.S. degree in e-commerce from Shijiazhuang Tiedao University, Hebei, China, in 2016, where she is currently pursuing the master's degree. Her research interests include data mining and mobile computing.

ZHANG JIAWEI received the bachelor's degree in computer science from Nanjing University, China, in 2012, and the Ph.D. degree from the Department of Computer Science, University of Illinois at Chicago. He founded the IFM Lab, in 2017, which is a research oriented academic laboratory, providing the latest information on fusion learning and data mining research works and application tools to both academia and industry, where he has been the Director, since 2018. He is currently an Assistant Professor with the Department of Computer Science, Florida State University. His current research interests include data mining and machine learning, especially multiple aligned social networks studies.

WU LEI was born in Xingtai, Hebei, China, in 1980. He is currently the Master Lecturer with Shijiazhuang Tiedao University and also a member of the Soft Science Research Institute on Engineering and Construction Management, Hebei. His research interests include data management on moving objects and location based social networks.