Fuzzy Trajectory Clustering Technique Based on Fast Reduced Measure of Semantic Similarity

Chunchun Hu Nianxue Luo Qiansheng Zhao

School of Geodesy and Geomatics, Wuhan University, 129 Luoyu Road,Wuhan 430079, China
{chchhu, xnluo, qshzhao}@sgg.whu.edu.cn

Received 8 December 2015; Revised 24 December 2015 ; Accepted 13 January 2016

Abstract. The data form of trajectory by introducing time and spatial dimension totally differs from traditional static data and has greatly enriched the content of data itself. Clustering spatio-temporal trajectory can find the motion and behavior patterns of moving object as it evolves over time. Generally, an object moves along a straight line at certain speed till it changes its direction and/or speed. However, the motion way, special semantics and fuzziness in trajectory data has not completely been taken yet into account during the clustering process. In this paper, we study the semantic extension of trajectory model, similarity measure and fuzzy trajectory clustering algorithm based on data reduction. First, a coarse-grained measure function of similarity based on data reduction, which includes location similarity, motion direction and speed similarity, is defined. Second, we propose a new fuzzy clustering algorithm named TraFCM for spatio-temporal trajectory data based on the defined similarity measure which can enhance the computing efficiency observably. The evaluation of experiment performed on synthetic and real trajectory indicate the effectiveness and efficiency of our approach. And the new clustering method shows excellent performance even if trajectories data are reduced to a half of raw data.

Keywords: Spatio-temporal trajectory, Semantic similarity measure, Fuzzy clustering, Data reduction

1 Introduction

Trajectory clustering is one of temporal data mining methods based on the cluster concept, and its main research objects are the trajectories produced by the moving objects. By exploring similar trajectories and extracting feature trajectories, we can find the motion and behavior patterns of moving objects which produce trajectory data by carrying location-aware devices.

However, trajectory clustering can partition spatio-temporal objects with similar behavior into together, and separate the spatio-temporal objects with different behavior as far as possible. A trajectory keep a record of position information of a moving object as it evolves over time, the concept of uncertainty appears in various ways [1]. Although some clustering techniques can deal with noise in clustering spatio-temporal data [2], but they usually ignore the fuzziness of data itself. Clustering methods introduce fuzzy logic, such as FCM [3], which allow each data element to belong to different clusters by a certain degree of membership. In addition, design and define the similarity measure between different trajectories is crucial issue in the trajectory clustering [4]. Many approaches have been introduced in the literatures that try to quantify the dissimilarity between trajectories.

Specially, with the technology development of wireless communication, satellite positioning and earth observation, the capability of acquiring data is greatly enhanced. Huge amount of data has been an important problem in the field of GIS over the years, and there are also speed, heterogeneous, analysis and mining issues except the huge amount for achieving across from the huge amount data to big data [5]. How to improve the computational efficiency and reduce the cost of time is also a key issue when clustering massive data. In this paper, our work mainly focus on the following: 1) measure similarity between spatio-temporal trajectories. Hausdorff distance is employed to measure trajectory similarity by semantic extension especially for the trajectories with the different time ranges; 2) introducing the data reduction method into measure similarity; 3) building the more adaptive trajectory fuzzy clustering. We propose a novel modification of the FCM algorithm for clustering spatio-temporal trajectory based on the new similarity measure.

2 Related Work

Spatio-temporal trajectory clustering technique is still a hot issue. Previous research mostly focused on clustering point data and static data, but could not be finer for the spatio-temporal data. Some new solution can handle the

clustering problems. Representative algorithms include TRACLUS [6], MMC [7] and so on. Another clustering techniques mainly bases on distance similarity. The well-known clustering algorithms can be expanded to satisfy trajectory data, such as k-means [8], BIRCH [9] and so on. Specially, the DBSCAN [10] is expanded to cluster the ship trajectory [11] and GPS trajectory segment [12]. However, the proximate trajectories are grouped together according to similar measure. And these measure techniques mainly focus on the following categories.

Firstly, the Euclidean distance is one of the most original trajectory similarity measure methods. However, Euclidean distance only is used to compute pairwise points between two trajectories. When the two trajectories scales are inconsistent, calculation Euclidean distance will fail if we lack data points. Secondly, DTW [13] is used to calculate the similarity of two sequences. Compared with other approaches, its advantage is not limited by the length of time sequence. Given two time sequence Q and C, which their data lengths are *n* and *m* respectively, we can compute the similarity distance between them [14]. However, DTW is very sensitive for isolated points and noise. Thirdly, LCSS distance can measure the similarity of two trajectories by obtaining the longest common subsequence between them [15]. LCSS distance only calculates similar parts between trajectories. However, LCSS only focuses on the similar parts between trajectories, but not to compute the dissimilar parts. So the similarity measure of the trajectories by computing LCSS distance is too "rough ". While other similarities measure between the trajectories present good performance like MBB[16]. A minimal bounding box (MBB) represents an interval bounded by limits of time and location [16].

3 Coarse- Grained Trajectory Similarity Measure Based on Semantic Expansion

Spatio-temporal trajectory is a sequence which keep a record of locations and time of a moving object [4], thus it has the spatial and temporal features. And it is usually expressed using a group of spatio-temporal sampled point sequences. Generally, it is key to measure the similarity between the trajectories during the clustering process. In this paper, Hausdorff distance is employed to measure the similarity. Hausdorff distance, which is improved based on the Euclidean distance, can guarantee the accuracy of trajectory distance measure. Meanwhile, it is robustness and also can be used to calculate trajectory distance under different time scale.

3.1 Similarity Measure Based on Semantic Expansion

Through the Hausdorff distance can be a very good solution to compute spatial distance between trajectories. But it can only solve the mismatch degree of scalar data. While moving objects are assumed to move straight between the observed points at a constant speed [16]. Usually, we group objects which perform similar movements like going in the same direction or perform the same turns. Therefore, the trajectories have directional properties. Adding orientation variable can achieve spatial semantic expansion. However, most simplified trajectory data models only include spatial and temporal information. And it can be restored by using the linear interpolation. So the direction of each location can be fixed according to time stamp on the trajectory. In addition, an object moves along a straight line at some constant speed till it changes its direction and/or speed [16]. Each moving object *o* was presented by a 5-tuple (x_o,y_o,v_{xo},v_{yo},t_o) [17] which included location and speed. In the paper [18], a set of triples (*id*; *loc*; *t*) are used to represent the trajectory. In this paper, the trajectory structure is expanded and defined as a set of triples (*Location*, *t*, *Direction*, *Speed*). They directly impact on the calculation of the trajectory similarity.

3.2 Coarse-Grained Trajectory Similarity Measure

However, the time complexity of the Hausdorff distance is *o*(*nm*), where *n* and *m* is the point numbers of two trajectories respectively. It will increase the time cost and reduce the calculation efficiency heavily. The length of original trajectory sequence can be cut down from *n* to *l* (*l*<*n*) if we explore data reduction method. The vertical distance method will be introduced to achieve data reduction in this paper. The reduced process of vertical distance method is as follows.

Take three points for each trajectory curve successively, and calculate the vertical distance *d* between intermediate point and the link line of other two points, and compare *d* with the given tolerance D. If *d*<D, then delete the intermediate point; If *d*>D, then keep the intermediate point; Then take the next three points successively until finishing the curve scan.

Obviously, the granularity size of data reduction depends on the tolerance D. When compute the distance between the point *i* on the trajectory Q and the point *j* on the trajectory C, we need to determine whether or not keep them by employing the vertical distance method. As shown in the Fig.1, the points q_4 and q_6 with red colored label on the sequence Q will be deleted because the vertical distance *d* between the points and the link line of its

two adjacent point is less than the tolerance D. And point c_5 will be deleted on the sequence C similarly. The time complexity of Hausdorff distance will become $o(ls)$ ($l<n$, $s<m$) after data reduction. When the length of trajectory sequences are very long and $l<<n$ and $s<<m$, the computing efficiency will be enhanced significantly.

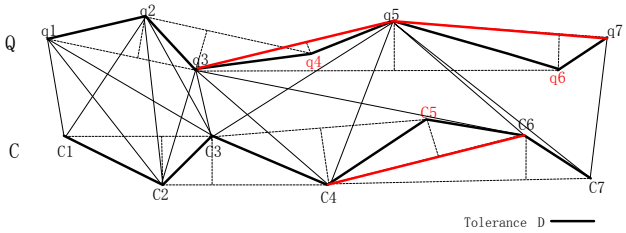

Fig. 1. Data reduction based on vertical distance method

To sum up, the similarity measure which put together the differences of location, direction and speed will be refined based on reduced Hausdorff distance as follows.

The Minimum Spatial Distance. The minimum spatial distance of any point r_i kept by data reduction on trajectory R to each point kept by data reduction on S is defined as follow:

$$dist_{\min}(r_i, S) = \min_{s_j \in S}\{\sqrt{(x_{r_i} - x_{s_j})^2 + (y_{r_i} - y_{s_j})^2}\} \quad (1)$$

$$H_{dist}(R,S) = \max\{h_{dist}(R,S), h_{dist}(S,R)\}, h_{dist}(R,S) = \max_{r \in R}(dist_{\min}) \quad (2)$$

The Minimum Spatial Difference of Direction. The minimum spatial difference of direction of any point r_i kept by data reduction on trajectory R to each point kept by data reduction on S is defined as follow:

$$dist\theta_{\min}(r_i, S) = \min_{s \in S}\{\arccos(\frac{direction_{r_i}^t . direction_{s_j}^n}{\left|direction_{r_i}^t\right| . \left|direction_{s_j}^n\right|})\} \quad (3)$$

$$direction_i^r = (x_{r_{i+1}}^t - x_{r_i}^t, y_{r_{i+1}}^t - y_{r_i}^t) \quad (4)$$

where the above equation represents the vector of point r_i to r_{i+1} at time t.

$$H_{direc}(R,S) = \max\{h_{direc}(R,S), h_{direc}(S,R)\}, h_{direc}(R,S) = \max_{r \in R}(dist\theta_{\min}) \quad (5)$$

The Minimum Speed Difference. The minimum speed difference of any point r_i kept by data reduction on trajectory R to each point kept by data reduction on S is defined as follow:

$$distSpeed_{\min} = \min_{s_j \in S}\{\left|speed_{r_i} - speed_{s_j}\right|\} \quad (6)$$

$$H_{speed}(R,S) = \max\{h_{speed}(R,S), h_{speed}(S,R)\}, h_{speed}(R,S) = \max_{r \in R}(distSpeed_{\min}) \quad (7)$$

4 New Fuzzy Trajectory Clustering Algorithm

4.1 The New Objective Function of New Clustering Algorithm

Basically, cluster analysis techniques could be divided into two categories, namely, crisp and fuzzy clustering. The crisp clustering could cut off the link between objects and cause more deviation of clustering results. A common fuzzy clustering algorithm is the Fuzzy C-Means (FCM). In order to cluster trajectories and extract

meaningful trajectory distribution pattern, the new objective function of new clustering algorithm named TraFCM was defined as follows:

$$J_m = \sum_{i=1}^{c}\sum_{j=1}^{n}(u_{ij})^m\{H^2{}_{dist}(R,S)\,|H^2{}_{direc}(R,S)\,|\,H^2{}_{speed}(R,S)\} \quad (8)$$

where $H_{dist}(R,S)$, $H_{direc}(R,S)$ and $H_{speed}(R,S)$ are respectively the similarity measure function of the location, direction and speed. And n is the number of trajectories.

4.2 New Clustering Algorithm

The new fuzzy trajectory algorithm is executed in the following steps.

Set Tolerance D, Clustering Parameters and Initialize Cluster Centroid. We set different tolerance D in order to obtain the reduced data with different compression ratio. The settings of clustering parameters include the threshold ε ,the cluster number C, weight coefficient k_1, k_2 and k_3, and record number n_c of cluster centroid. Each cluster centroid represents a trajectory grouped by multiple points.

Calculate Membership Degree. Compute matrix of the U(k) =[u_{ij}] for i=1,2,…,c using the following formula.

$$u_{ij} = k1\cdot\frac{(H_{dist}(R_j,S_i))^{(-\frac{1}{m-1})}}{\sum_{i=1}^{c}(H_{dist}(R_j,S_i))^{(-\frac{1}{m-1})}} + k2\cdot\frac{(H_{direc}(R_j,S_i))^{(-\frac{1}{m-1})}}{\sum_{i=1}^{c}(H_{direc}(R_j,S_i))^{(-\frac{1}{m-1})}} + k3\cdot\frac{(H_{speed}(R_j,S_i))^{(-\frac{1}{m-1})}}{\sum_{i=1}^{c}(H_{speed}(R_j,S_i))^{(-\frac{1}{m-1})}} \quad (9)$$

where the similarity distance between the jth trajectory and ith trajectory centroid can be computed by using the formula (1) to (7). And k_1, k_2 and k_3 respectively are the weight of the location, direction and speed. To change these weights can adjust the contribution of each factor for trajectory similarity under the condition $k_1+k_2+k_3=1$.

Update Fuzzy Cluster Centroid. Update the fuzzy cluster centroid $S_i(k+1)$. We can compute new cluster centroid using the following formula.

$$S_i\ (k+1)(r_j(Variable)) = \frac{\sum_{s=1}^{n}u_{ij}^m(k).R_j^s(Variable))}{\sum_{s=1}^{n}u_{ij}^m(k)} \quad (10)$$

where *Variable=Location/direction/speed*, and the expression $R_j^s(Variable)$ represents the location, direction and speed of the sth point on the jth trajectory.

Repeat and Halt Iteration. If meet the following formula, then iteration halts; Otherwise return the second step.

$$\|S(k)-S(k+1)\| < \varepsilon \quad (11)$$

The new FCM algorithm always converge a local minimum value through above iteration calculation. And the time complexity of new FCM is $o(nl)$, where n is the number of trajectories and l is the max number of points on each reduced trajectory($l<m$, m is the max numbers of points on each trajectory and $m>>n$). If we chose the high value of tolerance D, the time complexity of new FCM will decrease greatly.

5 Experimental Results

The experimental data consists of two trajectory sets. One group includes 20 synthetic trajectory, the other group include real trajectory sets of 50 trucks within 37 days (http://www.chorochronos.org/). In order to obtain credible experiment results, the new algorithm parameter is set to m=2, the iterative conditions for ε =0.00001, k_1, k_2 and $k3$ are set to 0.4, 0.3 and 0.3 respectively.

5.1 Experimental Results of New Trajectory Clustering

The experimental result of synthetic trajectory data is shown in Fig.2. According to the semantic features like position and direction, the 20 trajectories are mainly divided into two clusters, which the red line is not only characterized trajectory but also clustering center of each cluster. The clustering result is very similar to the position and direction of each trajectory as shown in Fig.2.

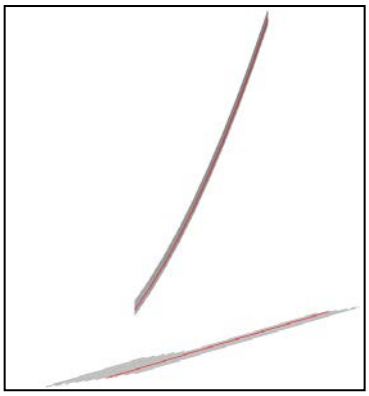

Fig. 2. The clustering result of synthetic trajectory at C=2

However, fuzzy clustering algorithm is an unsupervised classification method, the number of cluster C set respectively to 2, 3, 4, 5, 6 to observe the experimental results which implement on the real trajectory. The clustering results of C=2, which trajectory data did not be reduced during measure the similarity between trajectories, are shown in the left of Fig.3. While the right part in Fig.3 show the result produced by new reduced trajectory clustering when C=2. Different colors represent different trajectory clusters in the figure. As shown in the right result of Fig.3, there are two trajectories grouped into the second cluster after data reduction compared with the left result. And there is the relation between the clustering results and spatial position and direction of the trajectory when C=2. And the results are consistent with the coefficient settings of k_1, k_2 and k_3 in experiments, which the k_1=0.4 shows that the spatial location has a larger weight than other coefficient.

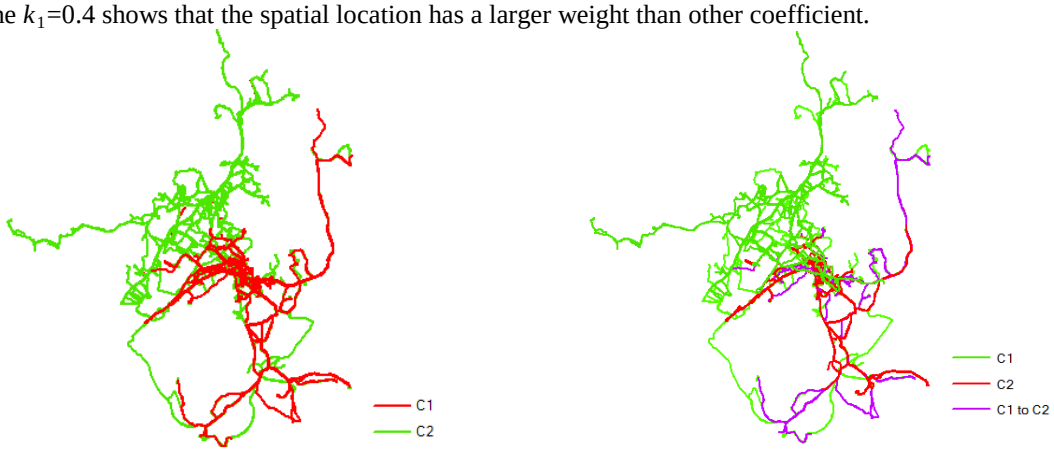

Fig. 2. The clustering results at C=2 (the left result produced by new clustering without data reduction and the right result produced by new reduced clustering algorithm)

5.2 Clustering Validity and Performance Evaluation under Different Data Reduced Ratio

In order to obtain the optimal clustering result, the validity of clustering results need be evaluated. The evaluation results by exploring the cluster validity index XB [19] are shown in Table 1. The smaller the value of XB is, the more optimal the clustering result is. From the Table 1, the XB index achieves the minimum value when data reduced ratio is less than 0.6 at C=2. As shown in the left result of Fig.4, the points on all polylines locate the lowest position at C=2 except the polyline which reduced ratio is equal to 0.6. From the left result of Fig.4, we can see that the XB index obtain minimum value at C=5, to where the red arrow point, when reduced ratio is equal to 0.6.

Meanwhile, in order to validate the performance of new trajectory clustering, we compared the runtime of the new trajectory clustering between different data size. The runtime under the different data reduction ratio are shown in the right figure of Fig.4. When the data size is cut down to a half of original data size, the runtime become very short and is less than a half of the original runtime. And the optimal clustering partition also can be obtained at C=2. When data size is cut down to 40 percent of original data size, we can't get the optimal clustering result at C=2 instead of C=5 as shown in the right figure of Fig.4. However, the new method shows excellent performance within the reduced range of 50 percent.

Table 1. The value of XB index under different reduced ratio

C	Value of XB different data reduced ratio				
	0	0.02	0.2	0.5	0.6
2	***0.91***	***0.91***	***0.54***	***0.75***	1.39
3	1.57	1.18	1.38	0.94	1.46
4	2.39	1.55	2.10	0.87	1.75
5	0.99	1.78	2.83	1.31	***1.31***
6	1.07	1.36	1.79	1.48	1.62

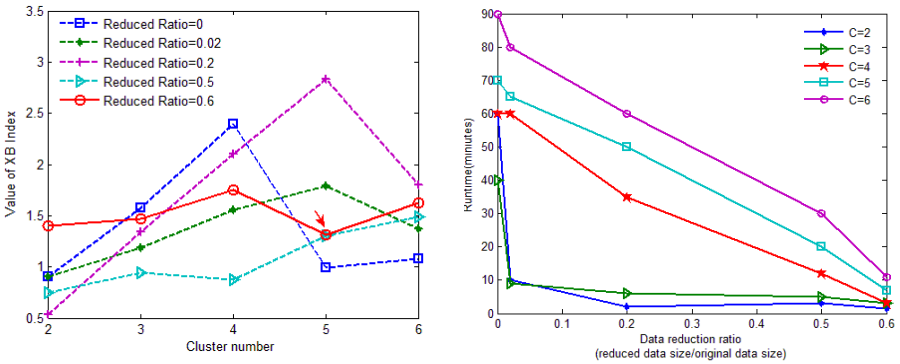

Fig. 3. Clustering validity evaluation and performance evaluation

6 Conclusion

In this paper, we present a new fuzzy clustering algorithm for grouping trajectories of moving objects, which cluster trajectories with the similar semantic feature. Based on our semantic extension of trajectory model, Hausdorff distance and data reduction method, we defined a coarse-grained measure function of similarity which includes the location similarity, the direction similarity and the speed similarity. And the similarity measure is used to build the new fuzzy clustering algorithm for trajectory data set. Moreover, the proposed algorithm can discover the clusters of a bundle of trajectories. The experimental results implement on synthetic and real trajectory data show the effectiveness and efficiency of our approach. Future work will include performing extensive experiment for efficiency purposes and the performance using large trajectory datasets, while the second work includes the improvement of the proposed algorithm and evaluation methods.

Acknowledgement

This research was supported by research project of National Natural Science Foundation under Grant 41401452 and 91224008.

References

[1] N. Pelekis, , I. Kopanakis et al., "Clustering Uncertain Trajectories," *Knowl Inf Syst.*, Vol. 28, No. 1, pp. 117-147, 2011.

[2] I. Assent, R. Krieger, B. Glavic, T. Seidl, "Clustering Multidimensional Sequences in Spatial and Temporal Databases, " *Knowl Inf Sys.*, Vol. 16, No. 1, pp. 29–51, 2008.

[3] J. C. Bezdek, R. Ehrlich, W. Full, "FCM: the Fuzzy C-Means Clustering Algorithm," *Comput Geosci.*, Vol. 10, No. 2-3, pp. 191–203, 1984.

[4] Gong Xi, Pei Tao, Sun Jia, Luo Ming, "Review of the Research Progresses in Trajectory Clustering Methods," *Process in Geography*, Vol. 30, No. 5, pp. 522-534, 2011.

[5] Qingquan Li, Deren Li, "Big Data GIS," *Geomatics and Information Science of Wuhan University*, Vol. 39, No. 6, pp. 641-644, 2014.

[6] J. G. Lee, J. Han, K.Y. Whang, "Trajectory Clustering: A partition-and-group Framework," in *Proceedings of the 2007 ACM SIGMOD*, pp.593-604, 2007.

[7] Y. F. Li, ,J. W. Han, J. Yang, "Clustering Moving Objects," in *Proceedings of the 10th ACM SIGKDD*, pp.617-622, 2004.

[8] P. Bradley, U. Fayyad, C. Reina, "Scaling Clustering Algorithms to Large Databases," in *Proceedings of the 4th Int. Conf. On KDD*, pp.9-15, 1998.

[9] T. Zhang, R. Ramakrishnan, M. Livny, "BIRCH: An Efficient Data Clustering Method for Very Large Databases," in *Proceedings of 1996 ACM-SIGMOD*, pp.103-114, 1996.

[10] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise, " In *Proceedings of KDD*, pp.226-231, 1996.

[11] B. Liu, N. De Souza Erico, Matwin Stan, Sydow Marcin, "Knowledge-Based Clustering of Ship Trajectories Using Density-Based Approach," in *Proceedings of 2014 IEEE International Conference on Big Data*, pp. 603-608, 2014.

[12] W. Chen, M.H. Ji, J.M. Wang, "T-DBSCAN: A Spatiotemporal Density Clustering for GPS Trajectory Segmentation," *International Journal of Online Engineering*, Vol.10, No.6, pp. 19-24, 2014.

[13] R. Agrawal, C. Faloutsos, A. Swami, "Efficient Similarity Search in Sequence Databases," *Lecture Notes in Computer Science*, Vol.730, pp.69-84, 1993.

[14] B D. J.erndt, J.Clifford, "Finding Patterns in Time Series: A Dynamic Programming Approach, " *Advances in Knowledge Discovery and Data Mining*, pp.229-248, 1996.

[15] L. Chen, M. Özsu, V. Oria, "Robust and Fast Similarity Search for Moving Object Trajectories," *in Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data*, pp.491-502, 2005.

[16] S. Elnekave et al., "Measuring Similarity between Trajectories of Mobile Objects," *Studies in Computational Intelligence*, Vol.91, pp.101-128, 2008.

[17] Y. Li, J. Han, J. Yang, "Clustering Moving Objects," in *Proceedings of the Tenth International Conference on Knowledge Discovery and Data Mining*, pp.617-622, 2004.

[18] M. D' Auria, M. Nanni, D. Pedreschi, "Time-focused Density-based Clustering of Trajectories of Moving Objects," *JIIS Special Issue on Mining Spatio-Temporal Data*, Vol.27, No.3, pp.267-268, 2006.

[19] X. L. Xie, G. Beni, "A Validity Measure for Fuzzy Clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol 13, No.8, pp.841-847, 1991.