論文 / 著書情報
Article / Book Information

Title	Mining User Similarity from GPS Trajectory Based on Spatial-temporal and Semantic Information
Author	Qiuhan Han, Atsushi Yoshikawa, Masayuki Yamamura
Journal/Book name	2022 3rd International Conference on Information Science, Parallel and Distributed Systems (ISPDS), , , pp. 174-180
Pub. date	2022, 7
DOI	https://doi.org/10.1109/ISPDS56360.2022.9874192
Copyright	(c)2022 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.
Note	This file is author (final) version.

2022 3rd International Conference on Information Science, Parallel and Distributed Systems (ISPDS)

Mining User Similarity from GPS Trajectory Based on Spatial-temporal and Semantic Information

Qiuhan Han
School of Computing
Tokyo Institute of Technology
Yokohama, Japan
qiuhhan.han@gmail.com

Atsushi Yoshikawa
School of Computing
Tokyo Institute of Technology
Yokohama, Japan

Masayuki Yamamura
School of Computing
Tokyo Institute of Technology
Yokohama, Japan

***Abstract*—In this study, we proposed a new framework to mine and analyze information from GPS trajectory data to find similar users from a spatial-temporal and semantic perspective. The framework combines spatial-temporal and semantic similarity techniques to achieve a system with low computational overhead and good similarity accuracy by using the characteristics of individual movements to identify similar users. It consists of three steps: first, spatial-temporal features are obtained by modeling and clustering stay points, and using them to calculate spatial-temporal similarities; next, using categories of points of interest within stay regions as semantic information, the semantic similarity can then be computed by frequent sequential pattern mining; finally, the spatial-temporal and semantic similarities can be combined to calculate the user similarity. We compared the results with those of related studies. The K-nearest neighbors experiments showed that the combination of spatial-temporal and semantic similarity methods exhibited excellent performance, being able to identify similar users more accurately. Consequently, our proposed method could be a useful identification framework in situations where large volumes of human spatial-temporal trajectory data exist, possibly due to the development of GPS devices and storage technology.**

Keywords-user similarity; GPS trajectory; data mining

I. INTRODUCTION

With the development of inexpensive data storage and more feasible data capture technologies, lifelogging—especially location logging—is becoming increasingly accessible to all. People can actively or passively share their location on smartphone apps that use integrated sensors to record their movements.

There has been an enormous amount of research on such logging data sets. Since human movements exhibit a high degree of temporal and spatial regularity [1], it should be possible to characterize an individual by travel distance with a significant probability of returning to a few frequented locations. Based on this concept, researchers have tried to glean some interesting information from trajectory data regarding statistical features and user profiles—what kind of people produce such trajectories? What are the user relationships that generate these trajectories? This has led, quite naturally, to the main aim of this paper: mining user similarity from their trajectory data.

II. RELATED WORKS

User similarity mining based on trajectory data can be categorized into two categories based on the type of information emphasized: spatial-temporal and semantic.

A. Spatial-temporal information

A natural concept in trajectory similarity mining is that the more similar the geographic information of the locations the trajectories pass through, the higher the similarity between the users who generated them. Reference [2] proposed a hierarchical measurement method and defined the stay point, which could be regarded as an aggregation of spatial-temporal information (discussed in the Preliminaries of Section III). To fully understand user behavior, some researchers clustered stay points to obtain significant places of users from both personal and public perspectives [3]. Some interpretable characteristics included information about users' home and visited locations combined with other features to infer user demographics [4,5,6].

However, the above studies focused primarily on spatial-temporal information and did not consider semantic information. The similarity between users should be related to such semantic information, especially when users with different homes, workplaces, or school locations share the similar behavior patterns.

B. Semantic information

Another perspective is that trajectories of people may be semantically similar even though they pass through different locations. For example, two movie lovers living in different areas would both tend to go to the cinema on weekends. Several researchers have proposed user similarity metrics based on the calculation of longest common subsequence or patterns [7,8]. The categories of locations visited by users have been used by researchers to reveal their interests [7]. They proposed maximal semantic trajectory pattern (MSTP), which uses an improved dynamic programming method to reduce the computational complexity. Reference [9] proposed relative-importance-based similarity (RIS) which considers the weights for longest common subsequence among different users and proposed

978-1-6654-8747-4/22/$31.00 ©2022 IEEE

common-patterns-distribution-based similarity (PDS) to distinguish frequent and occasional travelers, which identifies user similarities of different fineness. Reference [10] regarded the similarity measure as an assignment problem and used the Hungarian algorithm to solve it.

While previous studies have placed great emphasis on the semantic sequential patterns of trajectories, the temporal-spatial properties of trajectories which are critical for depicting user characteristics in geographic information science [3] have rarely been mentioned. Since both geographic and textual information is important in user similarity mining, we propose a framework that combines the two perspectives to fully extract both spatial-temporal and semantic information.

III. THEORY: PRELIMINARIES

In this section, we present definitions of the basic concepts of trajectory data mining.

Definition 1 (Trajectory). A trajectory of a user U is a sequence of continuous GPS points which can be represented as $T_U = \langle p_0, p_1, \dots, p_n \rangle$, where $p_i = (lat_i, lng_i, t_i)(0 \le i \le n)$ is a GPS point containing latitude, longitude, and time stamp information.

Figure 1 shows a sample of stay region and a corresponding stay point. The rectangle represented by a dashed line where a user stays over a time threshold but within a distance threshold is a stay region and its central point is the stay point. Their definitions can be expressed as follows:

Definition 2 (Stay region). A stay region is a geographical cluster where a user stays for longer than a period of δT but bounded with a distance of δD, which can be represented as $sr = \langle p_i, p_{i+1}, \dots, p_j \rangle$ where $distance(p_i, p_j) \le \delta D$, and $time(p_i, p_j) \ge \delta T$.

Definition 3 (Stay point). A stay point is the center of a stay region $sr = \langle p_i, p_{i+1}, \dots, p_j \rangle$, which can be represented as $s = (lat, lng, t_{arr}, t_{lea})$, where $s.lat = \sum_{k=i}^{j} p_k.lat / |sr|$ and $s.lng = \sum_{k=i}^{j} p_k.lng / |sr|$, stand for the average latitude and longitude of GPS points within stay region sr, respectively; $t_{arr} = p_i.t$ and $t_{lea} = p_j.t$ represent the arriving time and leaving time of the stay point, respectively.

In this way, we transform original trajectories into sequences of stay points and stay regions. The relatively more important geographical information is retained as much as possible while reducing the computation cost.

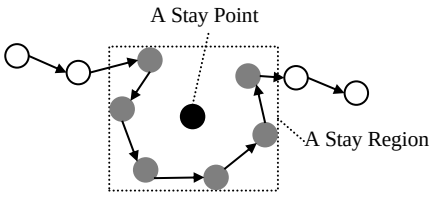

Figure 1. Example of a stay region and a stay point.

As for trajectory sequence pattern mining, the definitions of sequences and patterns can be expressed as follows:

Definition 4 (Sequence). A sequence is an ordered list of locations, which can be represented as $seq = \langle s_1, s_2, \dots, s_n \rangle$, where s_j is a semantic location and $1 <= j <= n$.

Definition 5 (Sequence Pattern). A trajectory pattern is a pair $< S, sup >$ where S is the sequence of semantic locations and sup is the frequency of S in all trajectories of the user.

IV. METHODOLOGY: PROPOSED FRAMEWORK

In this section, we present the proposed spatial-temporal and semantic similarity measurement, called the STS, framework, which follows the process shown in Figure 2.

A. Spatial-temporal similarity measure

Figure 3 shows our spatial-temporal similarity measurement method. We firstly cluster stay points to get spatial-temporal significant locations, then extract the spatial-temporal feature vectors, and finally compute the user similarity matrix.

1) Cluster stay points: The spatial-temporal information contained in stay point sequences can be further aggregated.

a) Significant Location: A location of geographic significance—such as a university—may contain several stay points. Such regions are defined as significant locations.

Definition 6 (Significant location). A significant location is a density-based cluster, containing several stay points. As a collection of stay points, a significant location is considered geographically realistic, which is a location that users have a high probability of visiting.

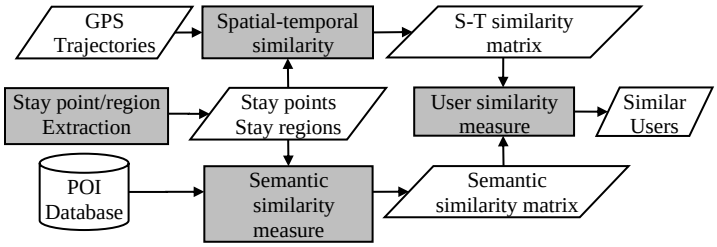

Figure 2. Proposed Framework

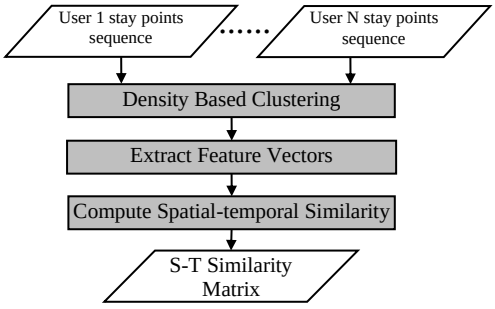

Figure 3. Spatial-temporal similarity measure.

b) Density-based clustering method: To discover such regions, we use a density-based clustering method. For all the stay points, the density-based spatial clustering of applications with noise (DBSCAN) algorithm can be applied to obtain the stay point clusters. These stay point clusters are considered as significant locations which have spatial-temporal significance.

2) Extract spatial-temporal feature vectors: There are infinite features that can be captured from trajectories [6], providing various perspectives to quantify and organize features in a hierarchical way. However, considering only the statistical characteristics of users may help categorize users with large variances in their daily routine, but not those who may have a consistent travel rhythm and belong to similar social groups.

For example, students always appear in dormitories and academic buildings, thus their home and workplace locations are always close to one other, which are hard to be distinguished if we only consider the most frequently visited locations as their characteristics. Consequently, apart from the obvious and high-frequency features, we should consider all the locations that users visit.

To fully mine the spatial-temporal information, we can use a feature vector to present a relationship of a user with the categories of the spatial-temporal clusters as follows:

Definition 7 (Spatial-temporal feature vector). The spatial-temporal feature vector of a user u is $F_u = \langle f_1, f_2, \dots, f_m \rangle$, where f_i is the frequency of stay point category i visited by user u and m is the number of unique spatial-temporal cluster categories.

$$f_i = \frac{n_i}{N} \times log \frac{|Users|}{|Users\ visiting\ i|} \quad (1)$$

where n_i is the number of clusters of category i; and N is the total number of clusters visited by user u.

3) *Computing spatial-temporal similarity:* For two spatial-temporal feature vectors of user U and user V, we proposed a spatial-temporal similarity measurement method as follows:

Definition 8 (Spatial-temporal similarity). The spatial-temporal similarity of two users can be written as $sim_{ST}(U, V)$, the value of which is negatively correlated with $dist(f_U, f_V)$.

$$sim_{ST}(U, V) = \frac{1}{1+dist(f_U, f_V)} \quad (2)$$

where $dist(f_U, f_V)$ is the cosine distance of spatial-temporal feature vectors of users U and user V.

B. Semantic similarity measure

Original trajectories are geographic sequences and do not contain semantic information. We transform the trajectories into semantic trajectories by combining user point of interest (POI) information, then mine the maximal frequent sequences to compute the semantic similarity score. Figure 4 shows the semantic similarity measurement method.

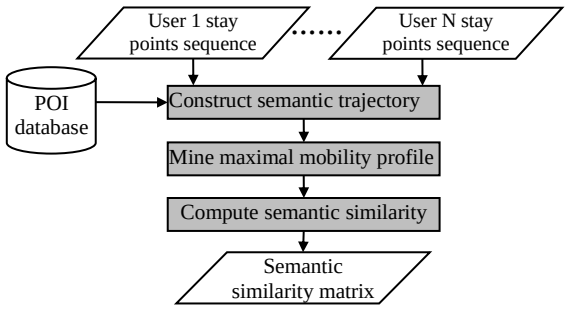

Figure 4. Semantic similarity measure.

1) Constructing semantic trajectory: To associate a given stay region with a POI dataset, the label of POIs within it can be used to represent its semantic meaning. One basic approach is to consider the label of the nearest POI of the stay point as the category of the corresponding stay region. Others use the most frequent POI category within the stay region as its semantic label. These methods do not show good results in our experiments because the actual stay region does not contain only one type of label and using a single label is not enough to represent the semantic complexity of the whole stay region. Based on the study by [11], the POI categories within a stay region can be regarded as short text, thus the feature vectors can be constructed based on the bag of words model. Following their study, we compute term frequency-inverse document frequency (TF-IDF) value of every stay region and use K-means to obtain the semantic locations and transform the GPS trajectory to semantic trajectory.

2) *Mining maximal mobility profile:* For semantic trajectories, the PrefixSpan algorithm [12] can be applied to extract a user frequent sequence pattern. In our experiments, we only used maximal frequent sequence patterns for our proposed method before comparing the results with those of other similarity measure methods which used the same pattern mining algorithm. The maximal mobility profile of a user is defined as follows:

Definition 9 (Maximal mobility profile). A user's maximal semantic mobility profile is a maximal frequent sequence pattern set containing the maximal sequence patterns of the user where none of its immediate super sets is frequent.

3) *Computing semantic similarity:* Given the maximal mobility profiles of user U and user V, we state the definition of their semantic similarity.

Definition 10 (Semantic similarity). For the maximal mobility profiles of two users U and V, the semantic similarity can be calculated as follows:

$$sim_{semc}(U, V) = \frac{\sum_{p_1 \in U, p_2 \in V} |lcs(p_1, p_2)| * sup(p_1)}{\sum_{p_1 \in U} |p_1| * sup(p_1)} \quad (3)$$

where $lcs(p_1, p_2)$ is the longest common sequence of pattern p_1 and p_2 ; $sup(p_1)$ means the support of pattern p_1 of user U.

C. User similarity definition

After computing the spatial-temporal similarity and semantic similarity of users U and V, we propose the STS similarity which measures these two kinds of similarity scores linearly.

Definition 11 (User similarity). The similarity of two users can be computed using the weighted spatial-temporal similarity and semantic similarity as follows:

$$sim(U,V) = w_1 * sim_{ST}(U,V) + w_2 * sim_{semc}(U,V) \quad (4)$$

V. EXPERIMENTAL RESULTS

A. Datasets

1) Trajectory dataset: The experiment was conducted using a dataset called Geo-life, collected by Microsoft Research Asia, consisting of 18,670 trajectories of 182 users over five years.

a) Splitting long trajectories: The trajectories of those users with more than 26 weeks (half a year) were split into two sub-trajectories and treated as two different users. This is to prevent some user trajectories from being too long and computationally overloaded when mining frequent sequences.

b) Selecting users: We followed the principle proposed by [9] to select only users having trajectories for more than 4 weeks with check-in locations on at least 4 days a week. These users were treated as active users while filtering out inactive users with few trajectories to make the experiment results more accurate.

c) Trajectory dataset description: TABLE I**.** shows the statistical characteristics of the processed trajectory dataset. The number of finally selected users is 98.

2) Ground truth: The original Geo-life paper used investigative questionnaires to obtain information from the trajectory owners as the ground truth dataset, but the questionnaires are not publicly available. However, a popular solution to create the ground truth dataset is to split the trajectory dataset into two parts [9,10]. For example, if we have an active user U whose trajectory is $T_U = \langle p_0, p_1, \dots, p_n \rangle$, we divide this trajectory into two parts, the former is $T_{U^*} = \langle p_0, p_1, \dots, p_{n/2} \rangle$ and the latter is $T_{U^\#} = \langle p_{n/2+1}, p_{n/2+2}, \dots, p_n \rangle$. As the sub-trajectories T_{U^*} and $T_{U^\#}$ are generated from the same user, the new generated user U^* and $U^\#$ can be considered as the nearest neighbor of each other.

3) *POI dataset:* We used the POI data set was crawled from Bai Du Map. TABLE II. shows the statistics for the different labels in the POI dataset. We removed POI labels that appear infrequently, and those that have no exact meaning, such as *Infrastructure*, *Address*, and *Others*.

TABLE I. THE TRAJECTORY DATASET

Active Users	***Weeks***	***Days***	***Days/Week***	***Days/user***
98	258	1409	5	14

TABLE II. THE POI DATASET

Label ID	***Label***	***Count***
1	Entertainment	64440
2	Food	165846
3	Hotel	34311
4	Education	61055
5	Health care	42437
6	Shopping	395117
7	Life service	166745
8	Construction	221471
9	Company	275904
10	Institution	79635
11	Bank	36501
12	Sport	20553
Total		***1564015***

B. Trajectory Segmentation

Before mining the user maximal mobility profile, the trajectory should be segmented into several sub-trajectories. We tried different ways; the comparison results are shown in the results section.

1) Segmenting by weeks: The basic concept is separating a trajectory according to the days. However, the extraction of stay points and stay regions greatly compresses trajectories, allowing us to obtain a sparse sequence of stay points and stay regions within one day, which are unfavorable for similarity calculation. In this way, the check-ins within one week are more appropriate to be connected as a sub-trajectory.

2) Segmenting by day of the week: Several researchers used a novel method to segment trajectory. For every user, they connected check-ins on a particular day of the week spanning over the entire trajectory to construct a sub-trajectory [9], such as the trajectory of Monday and so on. This segmentation method prevents the problem of over computation as users have too many sub-trajectories when mining their frequent patterns. Using this method, the number of sub-trajectories for every user can be limited to less than 7. This method was compared with segmentation by weeks in our experiments.

C. Parameter Selection

We used the same parameter values used in related works, as shown in TABLE III. Here, δD and δT are the distance threshold and time threshold for extracting the stay points or stay regions; ε, sup and k are the parameters of DBSCAN, PrefixSpan and K-means algorithms, respectively. As for STS, we set w_1 and w_2 to 0.5.

TABLE III. PARAMETER VALUES

Parameters	δD	δT	ε	sup	k	w_1	w_2
Value	200m	30min	200m	0.3	35	0.5	0.5

D. Evaluation Methods

The mean average precision (MAP) and normalized discounted cumulative gain (nDCG@K) metrics were used to evaluate the results. The definitions of MAP and nDCG@K can be simplified as proposed by [10], and can be computed as follows:

$$MAP = \frac{1}{N}\sum_{i=1}^{N}(1/Rank(Neighbor_i)) \quad (5)$$

$$nDCG@K = \frac{1}{N}\sum_{i=1}^{N}(1/log_2(Rank(Neighbor_i)+1)) \quad (6)$$

Where $Rank(Neighbor_i)$ means the rank of nearest neighbor of user i if it is within the retrieved neighbors other than positive infinite. N means all the users, including the ground truth. We also used the number of false predicted neighbors to evaluate the number of all the neighbors that are predicted incorrectly when we predicted different neighbors using different methods respectively.

E. Results

We compared the results of our method with those of MSTP in [7], and RIS and PDS in [9] to prove its effectiveness. Figure 5 and Figure 6 show the results of comparing our method with other methods in predicting k-nearest neighbors based on MAP and nDCG@K metrics. By introducing spatial-temporal information to extract features, our method achieved higher accuracy than other previous work in both MAP and nDCG@K metrics.

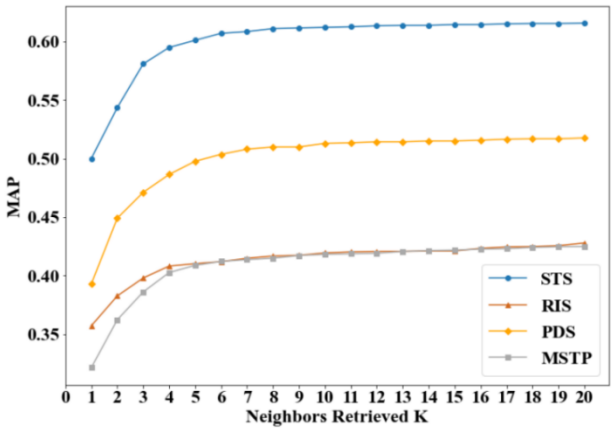

Figure 5. MAP of K Retrieved Neighbors

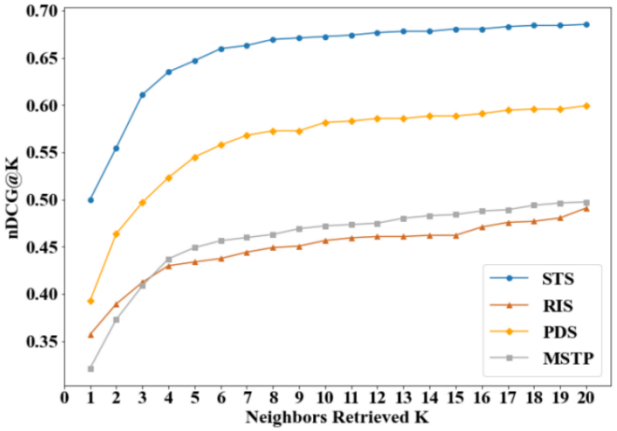

Figure 6. nDCG of K Retrieved Neighbors

Compared with PDS which has relatively better results, the MAP score of our STS method was improved from 0.52 to 0.62 and the nDCG@K score was improved from 0.6 to 0.68.

Figure 7 shows the comparison results of our method and other related methods on the number of falsely predicted neighbors. Compared with MSTP and RIS, our STS method encountered fewer prediction failures when predicting the same range of neighbors. The difference in the number of false predicted neighbors between our method and the PDS method is not so significant. However, considering that STS can obtain a higher MAP and nDCG@K accuracy, it outperforms the other methods overall as it takes geographical characteristics into account.

We also compared different approaches to trajectory segmentation in our STS method and the best-performing PDS method. Figure 8 and Figure 9 show that segmentation by day of the week results in slightly higher MAP and nDCG@K values than segmentation by week. It may be because the maximum frequent sequence patterns mined after the day of the week segmentation represent some frequent possible travel patterns of the user in a certain week. It is more logical than the sequence patterns which contain different weeks produced by sequence pattern mining based on the week segmentation. However, it has less impact on STS compared with PDS because the spatial-temporal features extracted in STS do not depend on trajectory segmentation; therefore, only the part of the semantic similarity measure is affected.

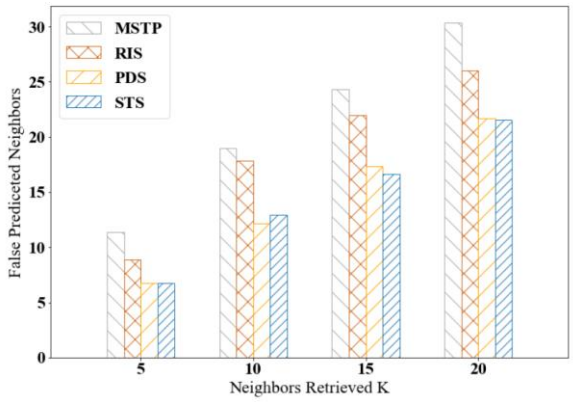

Figure 7. Number of false predicted neighbors.

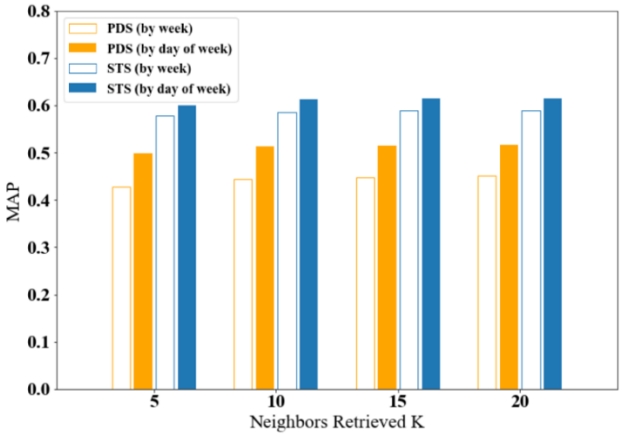

Figure 8. MAP of Different Trajectory Segmentation Methods

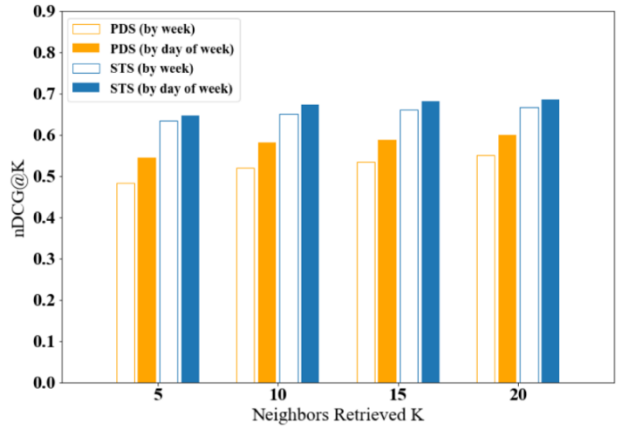

Figure 9. nDCG of Different Trajectory Segmentation Methods

F. Discussions

1) Spatial-temporal features: Compared with methods [7,9] that only considered semantic information, our experiment demonstrated that introducing spatial-temporal information increases the accuracy of user similarity identification, similar to [3]. However, in contrast to [3,4,5], we used user visited locations to construct the spatial-temporal feature vector, allowing for maximum preservation of the spatial-temporal information to achieve higher accuracy.

In our study problem, both kinds of similarities need to be considered. We hope that it was feasible for the proposed method to consider two trajectories of the output of a user as nearest neighbors. In contrast, in real life, the geographic location users may change: moving houses and switching jobs can make their before and after trajectory models different; semantic patterns can likewise change since their interests change over time. Therefore, both spatial-temporal and semantic information has the same importance in mining user similarity and should be well studied.

2) Maximal sequence patterns: Some studies [9,10] considered and calculated all frequent patterns. Although this improved the accuracy of their results, it is worth stating that mining all frequent patterns takes too much time. For the Geo-life dataset, the number of frequent patterns mined using PrefixSpan increases exponentially with the number of extracted stay regions, making computation based on all frequent patterns unfeasible. Since most of the frequent patterns are subsets of the maximum frequent patterns, considering only the maximum frequent patterns can greatly reduce the computational run time while still achieving good accuracy. How to balance the two to obtain more accurate results with less computational overhead should be further investigated.

VI. CONCLUSIONS

In this study, we obtained similarity scores of different levels of trajectories and calculated their spatial-temporal similarity by modeling the trajectories and calculating their semantic similarity by mining their semantic information. By combining semantic and spatial-temporal similarities, we achieved higher accuracy than that of previous studies with acceptable time constraints and identified similar users more accurately on large public datasets. The MAP score of our method was improved by about 19.2% compared to PDS and the nDCG@K was about 13.3%. This study lays the foundation for further research on trajectory-based user similarity mining, which could be used in applications such as friend and POI recommendations. However, only classical data mining methods are used to calculate user similarity manually in this study. In other words, user embeddings are constructed based on classical approaches, while using some classification methods or neural networks can mine features automatically and even more effectively. Furthermore, this problem can be transformed into a *trajectory user link* problem, which is a classification problem and can be solved using machine learning or deep learning methods. In the future, deep learning methods are expected to solve this problem and improve accuracy.

REFERENCES

[1] M.C. Gonzalez, C.A. Hidalgo, and A.L. Barabasi, “Understanding individual human mobility patterns,” Nature, 453(7196): 779–782, 2008.

[2] Y. Zheng, L. Zhang, Z. Ma, X. Xie, and W.Y. Ma, “Recommending friends and locations based on individual location history,” ACM Transactions on the Web (TWEB), 5(1): 1–44, 2011.

[3] M. Yang, C. Cheng, and B. Chen, “Mining individual similarity by assessing interactions with personally significant places from GPS trajectories,” ISPRS International Journal of Geo-Information, 7(3): 126, 2018

[4] J.J.C Ying, Y.J. Chang, C.M. Huang, and V.S. Tseng, “Demographic prediction based on users mobile behaviors,” Mobile Data Challenge, 2012: 1–4, 2012.

[5] A. Almaatouq, F. Prieto-Castrillo, and A. Pentland, “Mobile communication signatures of unemployment,” in: International conference on social informatics. Cham. pp. 407–418, 2016.

[6] L. Wu, L. Yang, Z. Huang, Y. Wang, Y. Chai, X. Peng, and Y. Liu, “Inferring demographics from human trajectories and geographical context,” Computers, Environment and Urban Systems, 77: 101368, 2019.

[7] J.J.C Ying, E.H.C. Lu, W.C Lee, T.C. Weng, and V.S. Tseng, “Mining user similarity from semantic trajectories,” in: Proceedings of the 2nd acm sigspatial international workshop on location based social networks. San Jose. pp. 19–26, 2010.

[8] X. Chen, J. Pang, and R. Xue, “Constructing and comparing user mobility profiles for location-based services,” in: Proceedings of the 28th Annual ACM Symposium on Applied Computing. Coimbra. pp. 261–266, 2013.

[9] P. Mazumdar, B.K. Patra, R. Lock, and S.B. Korra, “An approach to compute user similarity for GPS applications,” Knowledge-Based Systems, 113:125–142, 2016.

[10] Z. Lin, Q. Zeng, H. Duan, and F. Lu, “Finding similar users from GPS data based on assignment problem,” in: Proceedings of the 4th International Conference on Communication and Information Processing. Qingdao. pp. 283–288, 2018.

[11] X. Xiao, Y. Zheng, Q. Luo, and X. Xie, “Inferring social ties between users with human location history,” Journal of Ambient Intelligence and Humanized Computing, 5(1): 3–19, 2014.

[12] J. Han, J. Pei, B. Mortazavi-Asl, H. Pinto, Q. Chen, U. Dayal, and M. Hsu, “Prefixspan: Mining sequential patterns efficiently by prefix-projected pattern growth,” in: Proceedings of the 17th international conference on data engineering. Heidelberg. pp. 215–224, 2001.