Measuring User Similarity with Trajectory Patterns: Principles and New Metrics

Xihui Chen[1], Ruipeng Lu[2], Xiaoxing Ma[3], and Jun Pang[1,2]

[1] Interdisciplinary Centre for Security, Reliability and Trust, University of Luxembourg
[2] Faculty of Science, Technology and Communication, University of Luxembourg
[3] State Key Laboratory for Novel Software Technology,
Department of Computer Science and Technology, Nanjing University

Abstract. The accumulation of users' whereabouts in location-based applications has made it possible to construct user mobility profiles. Trajectory patterns, i.e., traces of places of interest that a user frequently visits, are among the most popular models of mobility profiles. In this paper, we revisit measuring user similarity using trajectory patterns, which is an important supplement for friend recommendation in on-line social networks. Specifically, we identify and formalise a number of basic principles that should hold when quantifying user similarity with trajectory patterns. These principles allow us to evaluate existing metrics in the literature and demonstrate their insufficiencies. Then we propose for the first time a new metric that respects all the identified principles. The metric is extended to deal with location semantics. Through experiments on a real-life trajectory dataset, we show the effectiveness of our new metrics.

1 Introduction

Nowadays, most people are equipped with mobile devices that are able to acquire their real-time positions. This technical progress leads to the emergence and popularity of *geo-social networks* (GSN) such as Bikely and Foursquare. What is attractive in GSNs is that people can share their locations with their friends. For example, photos and videos can be tagged by their shooting places. Even the traditional on-line social networks, e.g., Google+ and Facebook, have also upgraded to support location sharing. With GSNs becoming popular, an enormous number of locations have been posted and accumulated into large datasets of users' movements. This access to users' mobility history offers an opportunity to improve the friend recommendation service of GSNs because a user's historical movements significantly reveal his personal interests [1, 2]. Thus, recommending friends with similar interests can be supplemented by finding people with similar movements.

One method to identify users with similar movements is to construct and compare their *mobility profiles* which are composed of their *trajectory patterns* [3]. Intuitively, a trajectory pattern is a sequence of places of interest which a user frequently visits. The frequency by which the pattern is followed is called its *support value*. For instance, every morning Pierre, a student in Oxford travels by train from his home to Oxford from which he walks to Trinity College. This daily routine can be described as a trajectory pattern: $Home \rightarrow Oxford\ station \rightarrow Trinity\ College$. Typical transition time between two successive visits can also be extracted and annotated on trajectory patterns.

L. Chen et al. (Eds.): APWeb 2014, LNCS 8709, pp. 437–448, 2014.

Profiling user mobility has attracted a lot of research in recent years, and different models have been proposed, e.g., Lévy-walk [4–6] and Markov chains [7]. Compared with these mobility profile models, trajectory patterns provide a more concise representation of users' *typical* movements as only the places which are *meaningful* to users are taken into account. This subsequently results in a more efficient comparison due to the elimination of the positions during transition between places, especially compared to those methods based on users' raw trajectories [8, 2]. Thus, in this paper, we concentrate on measuring user similarity using trajectory patterns.

Related Work. Measuring user similarity with trajectory patterns has been studied in a few papers. Ying et al. [9] propose a metric based on *maximal trajectory pattern* (MTP) similarity. A maximal trajectory pattern is a pattern that is not contained in any other patterns. For any two trajectory patterns from two users respectively, a similarity value is calculated by referring to the length of their longest common sequences. Two users' similarity is then calculated as the weighted average of the similarity values of all pattern pairs. The weight assigned to two patterns is the average support values. Later, Chen et al. [10] identify and fix a weakness in the metric of Ying et al. [9] that the maximum similarity value (1.0) cannot be achieved even for two identical users.

Our Motivations. In the literature, the effectiveness of a metric is assessed by the difference between the calculated similarity values and the ground-truth similarity obtained in other ways, e.g., by questionnaires. However, there are no formal principles to capture the basic properties that a valid user similarity metric on trajectory patterns should respect. Take the relation *equality* as an example. Intuitively, two users are equal or their similarity is maximum if and only if their mobility profiles are exactly the same. However, even the metric proposed by Chen et al. [10] fails to satisfy this property as it ignores the differences between support values. In other words, two users are considered identical when they have the same trajectory patterns even if they visit these patterns with significantly distinctive frequencies. Without identifying design principles first, we cannot propose a meaningful user similarity metric to capture the real similarity between users based on their trajectory patterns.

Our Contributions. In this paper, we identify and define the basic principles that hold when measuring user similarity based on trajectory patterns. These principles enable us to re-evaluate existing metrics and discuss their insufficiency in capturing user similarity. Instead of fixing them, we propose for the first time a new metric which respects all the basic principles. Due to the importance of location semantics in identifying users' hobbies, we extend our new metric to measure user similarity to take into account the semantics of visited locations. Last but not least, we perform extensive experiments on real-life trajectory datasets and demonstrate the effectiveness of our metrics.

2 Preliminaries

In this section, we briefly introduce the basic concepts related to profiling user mobility and describe existing user similarity metrics based on trajectory patterns.

Basic Concepts. A trajectory is the path followed by a user through space in a certain time period. It can be considered as a trace of chronologically ordered spatio-temporal

points which record the user's geographical positions at different time points. Let $\mathcal{L}$ be the set of possible positions and $\mathcal{T}$ the totally ordered set of time points. A trajectory can be denoted as the sequence $(\langle \ell_1, t_1 \rangle, \ldots, \langle \ell_n, t_n \rangle)$ where $\ell_i \in \mathcal{L}$ and $t_i \in \mathcal{T}$ $(1 \leq i \leq n)$.

We use *regions of interest* (RoI) to represent the places which are meaningful to users, e.g., Trinity College in the example of Section 1. In fact, an RoI R can be seen as a set of adjacent geographic positions. Thus we have $R \subset \mathcal{L}$. As we previously mentioned, a trajectory pattern indicates one of a user's regular traces of RoIs [3]. Thus we represent a trajectory pattern P as a sequence of RoIs, i.e., $P = (R_1, \ldots, R_n)$ $(n \geq 1)$. It is also denoted as $R_1 \rightarrow \ldots \rightarrow R_n$ in this paper. We use $len(P)$ to denote its length, i.e., $len(P) = n$. If a user sequentially travels all the RoIs of a trajectory pattern in a trajectory, then we say that the trajectory *spatially contains* the trajectory pattern or the trajectory pattern has an *occurrence* in the trajectory.

Definition 1 (Spatial Containment). *For a trajectory T and a trajectory pattern $P = R_1 \rightarrow \ldots \rightarrow R_n$, we say that P is spatially contained in T if and only if there exists a subsequence of T, i.e., $T' = (\langle \ell'_1, t'_1 \rangle, \ldots, \langle \ell'_n, t'_n \rangle)$ such that $\forall 1 \leq i \leq n, \ell'_i \in R_i$.*

The movements of a user u in a time period can be stored as a dataset of trajectories and one trajectory pattern may have multiple occurrences in this dataset. We use *support value* (denoted as $sup_u(P)$) to quantify the frequency of its occurrence. Its value is calculated as the percentage of the trajectories containing pattern P among all his trajectories. A trajectory pattern is *frequent* if its support value is larger than a threshold σ. Let $\mathcal{P}_u^\sigma$ be user u's set of frequent trajectory patterns. Then $\mathcal{P}_u^\sigma = \{P \,|\, sup_u(P) \geq \sigma\}$.

As we discussed above, trajectory patterns captures users' regular movements and their support values quantify their visiting frequencies. These two aspects actually cover the regularity of user mobility. Thus we model user u's mobility profile $\mathcal{M}_u$ as the pair $\langle \mathcal{P}_u^\sigma, sup_u \rangle$. In the sequel, we use $\mathcal{P}_u$ for short by assuming that σ is given implicitly.

In some works (e.g., see [10]), transition time between successive RoIs is also considered when comparing two users. It is usually used as a discounter to the calculated user similarity. In this paper we only focus on the key step of user similarity calculation and users' regularity on transition time can be added similarly as in [10, 11].

MTP-Based Metrics. We briefly describe the metric proposed by Ying et al. [9] and its revision by Chen et al. [10]. Both methods use the set of *maximum trajectory patterns* (MTP) to represent a user mobility profile so as to avoid duplicate comparison between trajectory patterns. Given two patterns $P = (R_1, \ldots, R_n)$ and $Q = (R'_1, \ldots, R'_m)$, we say that Q is a subsequence of P (denoted by $Q \sqsubseteq P$) if there exists $j_1, \ldots, j_m$, such that $R_{j_i} = R'_i$ $(1 \leq i \leq m)$. Given user u's trajectory pattern set $\mathcal{P}_u$, the maximal trajectory pattern set is defined as $M(\mathcal{P}_u) = \{P \in \mathcal{P}_u \,|\, \nexists P' \in \mathcal{P}_u \; s.t. \; P \sqsubseteq P'\}$.

The main idea of the two MTP-based metrics is to compute the similarity between maximal trajectory patterns and then combine the similarity values. The two metrics calculate the similarity between maximal patterns in the same way, which is based on the length of their *longest common sequences*. For two patterns P and Q, the set of their longest common sequences is $\{S \,|\, S \sqsubseteq P \wedge S \sqsubseteq Q \wedge (\forall S' \sqsubseteq P \wedge S' \sqsubseteq Q, len(S) \geq len(S'))\}$. Let $lenLCS(P, Q)$ be the length of their longest common sequences. Then

the similarity between P and Q is $sim(P,Q) = \frac{2 \cdot lenLCS(P,Q)}{len(P)+len(Q)}$. Furthermore, a weight is calculated for the pair of maximal patterns, i.e., $w(P,Q) = \frac{1}{2}(sup_u(P) + sup_{u'}(Q))$.

The difference between the two MTP-based metrics is the way to combine the similarity values between maximal trajectory patterns. Ying et al. [9] calculate the average weighted similarity as the final user similarity:

$$sim(u,u') = \frac{\sum_{P_i \in M(\mathcal{P}_u)} \sum_{Q_j \in M(\mathcal{P}_{u'})} w(P_i,Q_j) \cdot sim(P_i,Q_j)}{\sum_{P_i \in M(\mathcal{P}_u)} \sum_{Q_j \in M(\mathcal{P}_{u'})} w(P_i,Q_j)}.$$

Chen et al. [10] find that the average similarity cannot guarantee the maximum similarity value (1.0) for identical mobility profiles. For example, suppose mobility profile $\mathcal{M}_u$ with pattern set $\{P_1, \ldots, P_n\}$ where any two patterns have the same support value but share no common parts, i.e., $lenLCS(P_i,P_j) = 0$ and $sup_u(P_i) = sup_u(P_j)$ for any $1 \leq i \neq j \leq n$, Thus, for $1 \leq j \leq n$ we have $sim(P_i,P_i) = 1$ and $sim(P_i,P_j) = 0$ if $i \neq j$. The similarity of $\mathcal{M}_u$ to itself, i.e., $sim(u,u)$, will be calculated as $\frac{1}{n}$, instead of the intuitive value 1.0. Thus, Chen et al. [10] propose a different combination method. First, for each maximal pattern P_i of user u, the method finds the most similar maximal pattern of u', denoted as $\psi_{u,u'}(P_i)$. Then they compute his *relative similarity* to u' as

$$sim(u \,|\, u') = \frac{\sum_{P_i \in M(\mathcal{P}_u)} sim(P_i, \psi_{u,u'}(P_i)) \cdot w(P_i, \psi_{u,u'}(P_i))}{\sum_{P_i \in M(\mathcal{P}_u)} w(P_i, \psi_{u,u'}(P_i))}.$$

In the end, the user similarity between u and u' is defined as the average of the two relative similarities: $sim(u,u') = \frac{1}{2}(sim(u \,|\, u') + sim(u' \,|\, u))$. In the above example, the two relative similarities are both 1.0, hence $sim(u,u) = 1.0$.

In the rest of paper, we use MSTP to refer to the measurement of Ying et al. [9] as it is originally designed to measure user similarity with location semantics and MTP for the metric of Chen et al. [10].

3 Principles

In this section, we present the basic principles that should hold when comparing users based on their trajectory patterns. Then we demonstrate by examples the insufficiencies of existing metrics with respect to the principles.

As we reduce the calculation of user similarity to the comparison of their mobility profiles, we investigate the basic principles that a valid similarity metric for two mobility profiles should satisfy. To begin with, we introduce two concepts about users' mobility profiles. First, given two users u_1 and u_2, we say that u_1's mobility profile is *contained* in u_2's mobility profile, denoted by $\mathcal{M}_{u_1} \prec \mathcal{M}_{u_2}$, if $\mathcal{P}_{u_1} \subseteq \mathcal{P}_{u_2}$ and $\forall P \in \mathcal{P}_{u_1}, sup_{u_1}(P) \leq sup_{u_2}(P)$ and $\mathcal{M}_{u_1} \neq \mathcal{M}_{u_2}$. Intuitively, this means that user u_1's regular movements are only part of those of u_2. Second, we use $\mathcal{M}_{u_1 \lhd u_2}$ to represent the mobility profile whose pattern set consists of all the common patterns shared

by u_1 and u_2, i.e., $\mathcal{P}_{u_1 \triangleleft u_2} = \mathcal{P}_{u_1} \cap \mathcal{P}_{u_2}$ and for any $P \in \mathcal{P}_{u_1 \triangleleft u_2}$, its support value equals to that of user u_1, i.e., $sup_{u_1 \triangleleft u_2}(P) = sup_{u_1}(P)$. It is obvious that the mobility profile $\mathcal{M}_{u_1 \triangleleft u_2}$ is contained in the mobility profile of u_1, i.e., $\mathcal{M}_{u_1 \triangleleft u_2} \prec \mathcal{M}_{u_1}$.

Example 1. Suppose four users whose pattern sets are

$$\mathcal{M}_{u_1} = \{A(0.1), C(0.2)\}; \qquad \mathcal{M}_{u_2} = \{A(0.1), C(0.3)\};$$
$$\mathcal{M}_{u_3} = \{A(0.1), B(0.2), C(0.4)\}; \quad \mathcal{M}_{u_4} = \{A(0.3), B(0.1), D(0.2)\}.$$

For the sake of simplicity, we put the support value in the parentheses for each pattern. Then $\mathcal{M}_{u_1} \prec \mathcal{M}_{u_2} \prec \mathcal{M}_{u_3}$. Furthermore, $\mathcal{M}_{u_3 \triangleleft u_4} = \{A(0.1), B(0.2)\}$ and $\mathcal{M}_{u_4 \triangleleft u_3} = \{A(0.3), B(0.1)\}$

Definition 2 (Principles). *A valid similarity metric based on user mobility profiles should satisfy all the principles described below:*

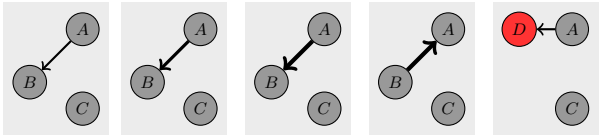
1. $sim(\mathcal{M}_{u_1}, \mathcal{M}_{u_2}) \geq 0$;
2. $sim(\mathcal{M}_{u_1}, \mathcal{M}_{u_2}) \leq 1$;
3. $sim(\mathcal{M}_{u_1}, \mathcal{M}_{u_2}) = sim(\mathcal{M}_{u_2}, \mathcal{M}_{u_1})$;
4. $sim(\mathcal{M}_{u_1}, \mathcal{M}_{u_2}) = 0$ *if and only if* $\mathcal{P}_{u_1 \triangleleft u_2} = \emptyset$;
5. $sim(\mathcal{M}_u, \mathcal{M}_u) = 1$;
6. $sim(\mathcal{M}_{u_1}, \mathcal{M}_{u_2}) > sim(\mathcal{M}_{u_1}, \mathcal{M}_{u_3})$ *if* $\mathcal{M}_{u_3} \prec \mathcal{M}_{u_2} \prec \mathcal{M}_{u_1}$;
7. $sim(\mathcal{M}_{u_1}, \mathcal{M}_{u_2}) > sim(\mathcal{M}_{u_1}, \mathcal{M}_{u_3})$ *if* $sim(\mathcal{M}_{u_1}, \mathcal{M}_{u_2 \triangleleft u_1})) > sim(\mathcal{M}_{u_1}, \mathcal{M}_{u_3 \triangleleft u_1}))$ *and* $sim(\mathcal{M}_{u_2}, \mathcal{M}_{u_2 \triangleleft u_1})) > sim(\mathcal{M}_{u_3}, \mathcal{M}_{u_3 \triangleleft u_1})$.

The first two principles regulate the range of the similarity value between two users. Principle 3 says that user similarity is *symmetric* and principle 4 states that two users have the minimum similarity value, i.e., 0 if and only if they have no common regular movements. Principle 5 indicates that user similarity should be maximum, i.e., 1.0, when a user is compared to himself. The last two principles are about comparing the similarity of a user to different users. The intuition of principle 6 is that users sharing more regular movements with a user should be more similar to him than users sharing less common behaviours. For instance, in Example 1 user u_2 is more similar to u_3 than u_1 as u_2 travels pattern C more regularly than u_1. Principle 7 says that a user is more similar to users who share more movements and have less different movements than those sharing less but having more different movements. The similarity values calculated with a valid metric should be consistent with this reasoning.

With these principles, we re-evaluate the existing metrics MSTP and MTP and find that they cannot satisfy all the principles. We use the following example to demonstrate their weaknesses.

Example 2. Suppose the following five users:

$$\mathcal{M}_{u_1} = \{A\ (0.4), B\ (0.4), C\ (0.4), A \rightarrow B\ (0.1)\};$$
$$\mathcal{M}_{u_2} = \{A\ (0.4), B\ (0.4), C\ (0.4), A \rightarrow B\ (0.2)\};$$
$$\mathcal{M}_{u_3} = \{A\ (0.4), B\ (0.4), C\ (0.4), A \rightarrow B\ (0.3)\};$$
$$\mathcal{M}_{u_4} = \{A\ (0.4), B\ (0.4), C\ (0.4), B \rightarrow A\ (0.3)\};$$
$$\mathcal{M}_{u_5} = \{A\ (0.4), C\ (0.4), D\ (0.4), A \rightarrow D\ (0.3)\}.$$



(a) u_1 (b) u_2 (c) u_3 (d) u_4 (e) u_5

Fig. 1. Mobility profiles in Example 2

Table 1. Pairwise user similarity

	MSTP					MTP				
	u_1	u_2	u_3	u_4	u_5	u_1	u_2	u_3	u_4	u_5
u_1	0.5	0.5	0.5	0.42	0.42	1.0	1.0	1.0	0.83	0.83
u_2	0.5	0.5	0.5	0.42	0.42	1.0	1.0	1.0	0.58	0.58
u_3	0.5	0.5	0.5	0.39	0.39	1.0	1.0	1.0	0.79	0.79
u_4	0.42	0.42	0.39	0.5	0.39	0.83	0.58	0.79	1.0	0.79
u_5	0.42	0.42	0.39	0.39	0.5	0.83	0.58	0.79	0.79	1.0

Table 2. User similarity by our method

	u_1	u_2	u_3	u_4	u_5
u_1	1.0	0.96	0.93	0.76	0.50
u_2	0.96	1.0	0.97	0.71	0.47
u_3	0.93	0.97	1.0	0.67	0.44
u_4	0.76	0.71	0.67	1.0	0.44
u_5	0.50	0.47	0.44	0.44	1.0

Figure 1 depicts the mobility profiles in a rectangle region. We use grey circles to indicate RoIs and arrows between RoIs to represent the transition direction whose thickness implies support values. Table 1 shows the results given by the two metrics.

From Table 1, it is clear that both metrics satisfy principles 1, 2, 3 and 4. Principle 5 is violated by metric MSTP as the similarity of any user to himself is not 1.0, which has been pointed out by Chen et al. [10]. Principle 6 is violated by both of them. Since $\mathcal{M}_{u_1} \prec \mathcal{M}_{u_2} \prec \mathcal{M}_{u_3}$, according to principle 6, we have $sim(\mathcal{M}_{u_3}, \mathcal{M}_{u_1}) < sim(\mathcal{M}_{u_3}, \mathcal{M}_{u_2})$. However, both metrics compute the same similarity values for them, i.e., 0.5 and 0.1, respectively. Principle 7 does not hold for both of the metrics either. Take the MTP metric as an example. According to its definition,

$$sim(\mathcal{M}_{u_2}, \mathcal{M}_{u_4 \triangleleft u_2}) = 0.82; \quad sim(\mathcal{M}_{u_2}, \mathcal{M}_{u_5 \triangleleft u_2}) = 0.86$$
$$sim(\mathcal{M}_{u_4}, \mathcal{M}_{u_4 \triangleleft u_2}) = 0.82; \quad sim(\mathcal{M}_{u_5}, \mathcal{M}_{u_5 \triangleleft u_2}) = 0.86.$$

As $sim(\mathcal{M}_{u_2}, \mathcal{M}_{u_5 \triangleleft u_2}) > sim(\mathcal{M}_{u_2}, \mathcal{M}_{u_4 \triangleleft u_2})$ and furthermore $sim(\mathcal{M}_{u_5}, \mathcal{M}_{u_5 \triangleleft u_2}) > sim(\mathcal{M}_{u_4}, \mathcal{M}_{u_4 \triangleleft u_2})$, if principle 7 holds we will have the relation $sim(\mathcal{M}_{u_2}, \mathcal{M}_{u_5}) > sim(\mathcal{M}_{u_2}, \mathcal{M}_{u_4})$. However, the metric cannot distinguish u_2's similarity to u_4 and u_5 and outputs the same similarity value (0.58) in both cases.

Neither of the metrics can give a precise evaluation of similarity for all users. From Figure 1, it is clear that the similarity values should decrease when comparing u_1 with the other users (from u_2 to u_5) – u_2 should be the most similar one to u_1 as they share a same set of trajectory patterns while u_5 is the least.

4 New Metrics

In this section, we propose for the first time user similarity metrics that satisfy all the basic principles discussed above. We first present a metric to measure user similarity

based on their movement called *mobility similarity* and then extend the metric to handle location semantics, called *location-semantic similarity*.

4.1 Mobility Similarity

The MTP-based metrics [9, 10] are problematic in comparing user similarity due to the inappropriate comparison between maximal trajectory patterns. In this section, we propose a new metric which does not compare maximal patterns but directly compare users' original mobility profiles. Moreover, instead of longest common patterns, we consider all common patterns of two users. Our main idea is to (1) compare two users based on the *relative importance* of their common patterns to each user's mobility profile and (2) take into account the difference of two users' frequencies by which they follow the common regular movements. Intuitively, if a user shares more common patterns with another user and their support values are also closer, then he is more similar to this user. This idea is consistent with the principles identified in Section 3.

We start with the calculation of the relative importance of the common patterns to a user mobility profile. A trajectory pattern can be interpreted as a description of users' movement *regularity*. The more RoIs it contains and the more frequent it occurs, the more regularity of the user it can represent. With the regularity of trajectory patterns, we can then quantify the regularity of a users' mobility profile. Given a user u, the regularity that his mobility profile represents can be calculated as follows:

$$\Gamma_u = \sum_{P \in \mathcal{P}_u} len(P) \cdot sup_u(P).$$

Given two users u and u', the relative importance of their common patterns with respect to u's mobility profile can be assessed by the ratio between the regularity of the common patterns and the whole pattern set of u. Recall that $\mathcal{M}_{u \lhd u'}$ is the mobility profile whose pattern set is composed of u' and u's common patterns and the support value of any trajectory pattern is equal to that of user u. Thus, $\Gamma_{u \lhd u'}$ is the movement regularity of user u expressed by the common patterns. Let $\Phi_{u,u'|u}$ be the relative importance of the common movements of u and u' to u's mobility profile, then it can be calculated as $\Phi_{u,u'|u} = \frac{\Gamma_{u \lhd u'}}{\Gamma_u}$.

We proceed to quantify the difference between the support values of two users' common trajectory patterns. The *Bray-Curtis similarity* [12] delivers reliable similarity measurements, especially in ecology. It can also be adopted as a metric of the similarity between two vectors. Given a user u, his support values of the trajectory patterns shared with u' can be modelled as a vector of real numbers each of which corresponds to the support value of a common pattern. Due to its popularity and simplicity, we make use of Bray-Curtis similarity to assess the closeness of two users' support values of their common patterns as the following:

$$\Psi_{u,u'} = 1 - \frac{\sum_{P \in \mathcal{P}_u \cap \mathcal{P}_{u'}} |\, sup_u(P) - sup_{u'}(P) \,|}{\sum_{P \in \mathcal{P}_u \cap \mathcal{P}_{u'}} (sup_u(P) + sup_{u'}(P))}.$$

Finally, the similarity between users u and u' can be calculated as

$$sim(u, u') = \sqrt{\Phi_{u,u'|u} \cdot \Phi_{u,u'|u'}} \cdot \Psi_{u,u'}.$$

It is easy to verify that our metric satisfies all the principles discussed in Section 3. We apply our metric to Example 2, and the results are shown in Table 2. We can see that our metric can give more precise similarity values which reflect the different similarities among users. Especially, the similarities of u_1 to the other users decrease from u_1 to u_5. We use CPS to refer to our metric, as it mainly utilises common pattern sets of users.

4.2 Location-Semantic Similarity

It has been addressed that the consideration of the functionalities of places can reveal more about users' similar hobbies. For instance, two users who live in different cities both like reading. According to their mobility, we cannot find their similarity because they go to different book stores. However, when considering the functionalities of places, e.g., 'book store', we will be able to discover their common interest. We call the functionalities of places *location semantics*. People always stay at a place for the service provided by the place. Given a trajectory, we can learn the trace of places where the user stayed for a certain amount of time [10, 11]. By labelling each of such places with its functionality, we can obtain a trace of location semantics, called a *semantic trajectory*. Similar to mining trajectory patterns from geographic trajectories, we can also mine *semantic trajectory patterns* from a user's semantic trajectories.

Let $\mathcal{LS}$ be the set of location semantics. Then a semantic trajectory pattern can be defined as a sequence of location semantic, i.e., $(\mu_1, \ldots, \mu_n)$ where for each $1 \leq i \leq n$, $\mu_i \in \mathcal{LS}$. It can also be represented as $\mu_1 \rightarrow \ldots \rightarrow \mu_n$. However, in practice a place usually corresponds to multiple location semantics. For instance, some shopping malls contain both shops and restaurants. For a visit to a place, its functionality that a user really uses is thus not certain. This uncertainty can be modelled as a probability distribution over all possible location semantics of the place, indicating the likelihood of how users use a functionality during their visits. A location semantic trajectory can thus be in the form of a sequence of sets of location semantics each of which corresponds to a probability distribution. Mining semantic trajectory patterns from such probabilistic semantic trajectories has been studied and termed as *probabilistic pattern mining* [13]. However, due to its underlying complexity, we propose in this paper a different method to obtain the set of semantic trajectory patterns by exploring user mobility profiles.

Although the metric such as the MSTP metric [9] can calculate user similarity with location semantics, it ignores the uncertainty of the real purposes of users' visits. Each location is assigned with a set of semantic tags, instead of a probability distribution on the tags. Furthermore, due to its dependence on maximal patterns, the MSTP metric with location semantics suffers the same problems as discussed in Section 3. In this paper, the calculation of user similarity with location semantics consists of two steps:

1. Transform trajectory patterns into semantic trajectory patterns and calculate their corresponding support values;
2. Calculate user similarity based on the obtained semantic trajectory patterns.

Once semantic trajectory patterns and their support values are available, the metric given in the previous section can be used. Thus we focus on the first step. Associating a location with its location semantics a user uses have been recognised as the problem of labelling locations with *semantic tags* [14]. A semantic tag corresponds to a type of

location semantics. Given an RoI R and $\mu \in \mathcal{LS}$, we use $\Pr(tag(R) = \mu)$ to denote the probability that a user stays at R for its functionality μ. For a trajectory pattern $P = (R_1, \ldots, R_n)$, we represent its induced semantic pattern as $lsp(P)$. We assume the tag labelling of RoIs in a trajectory is independent from each other. The likelihood that $lsp(P)$ is $Q = (\mu_1, \ldots, \mu_n)$, i.e., $Pr(lsp(P) = Q)$, can be calculated as follows:

$$Pr(lsp(P) = Q) = \prod_{1 \le i \le n} Pr(tag(R_i) = \mu_i).$$

For a semantic trajectory Q of a fixed length, any trajectory patterns of the same length in a user's profile may have a (positive) probability to induce Q. Thus, the support value of Q can be calculated as:

$$sup_u^{LS}(Q) = \sum_{P \in \mathcal{P}_u} Pr(lsp(P) = Q) \cdot sup_u(P).$$

Similar to trajectory patterns, we should choose the representative location-semantic patterns to compare users' similarity. A proper threshold of support values is thus required. From the calculation of the support values of semantic patterns, we can see that they depend on their length and the number of trajectory patterns of the same length. Therefore, the threshold for semantic patterns cannot be uniform, which is different from frequent trajectory patterns. Let $minPro$ be the minimum probability that a semantic tag is non-negligible to be the real semantic tag of an RoI. Then the threshold for a semantic pattern of length n can be calculated as the following:

$$\sigma_{LS}(n, \mathcal{P}_u) = minPro^n \cdot \sigma \cdot |\{P \in \mathcal{P} \,|\, len(P) = n\}|.$$

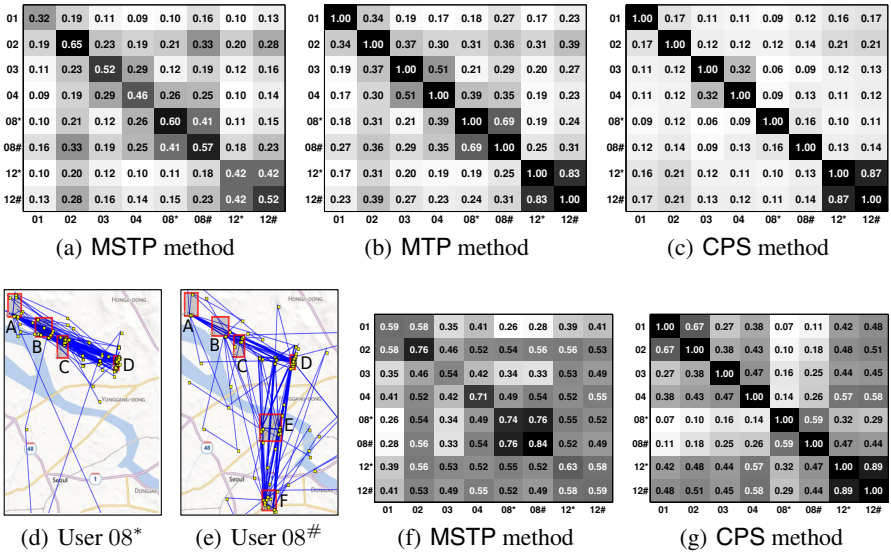
Intuitively, it equals to the support value of a semantic pattern, each of whose semantic tags has a larger probability than $minPro$ in all the trajectory patterns of the same length. In the end, the semantic trajectory pattern set of user u is obtained as

$$\mathcal{P}_u^{LS} = \{Q \,|\, (\exists P \in \mathcal{P}_u, len(P) = len(Q)) \wedge sup_u^{LS}(Q) \ge \sigma_{LS}(len(Q), \mathcal{P}_u)\}.$$

Example 3. Consider $\mathcal{M}_{u_3}$ from Example. 2 and suppose that $\mathcal{LS}$ consists of only two location semantic tags, e.g., $\mu_1 = \mathsf{hotel}$ and $\mu_2 = \mathsf{restaurant}$. For the sake of simplicity, we denote the distribution of an RoI R as a pair $d_R = \langle p_1, p_2 \rangle$ with $p_1 = Pr(tag(R) = \mu_1)$ and $p_2 = Pr(tag(R) = \mu_2)$. Suppose $d_A = \langle 0.4, 0.6 \rangle$, $d_B = \langle 0.2, 0.8 \rangle$ and $d_C = \langle 0.5, 0.5 \rangle$. For the semantic trajectory pattern μ_1, as patterns A, B and C can all induce it, its support value is calculated as: $sup_{u_3}^{LS}(\mu_1) = 0.4 \times 0.4 + 0.2 \times 0.4 + 0.5 \times 0.4 = 0.44$. If $\sigma = 0.2$ and $minPro = 0.2$, since we have 3 trajectory patterns with length 1 in $\mathcal{M}_{u_3}$, the support value threshold for semantic pattern μ_1, i.e., $\sigma_{LS}(1, \mathcal{P}_{u_3})$ is $0.2 \times 0.2 \times 3 = 0.12$. As the semantic pattern μ_2 has the same length as μ_1, its support value threshold is also $\sigma_{LS}(1, \mathcal{P}_{u_3})$. The calculation for other patterns is similar. Finally, we compute the set $\mathcal{P}_{u_3}^{LS}$ as follows

$$\{\mu_1(0.44), \mu_2(0.76), \mu_1 \to \mu_2(0.096), \mu_1 \to \mu_1(0.024), \mu_2 \to \mu_1(0.036), \mu_2 \to \mu_2(0.144)\}.$$

If the distribution for RoI D is also $\langle 0.2, 0.8 \rangle$ which is the same for RoI C, then u_3 and u_5 will have identical semantic pattern sets, i.e., $\mathcal{P}_{u_3}^{LS} = \mathcal{P}_{u_5}^{LS}$. Thus, u_3 and u_5 become much more similar when considering location semantics.



(a) MSTP method (b) MTP method (c) CPS method

(d) User 08^* (e) User $08^\#$ (f) MSTP method (g) CPS method

Fig. 2. User mobility similarity by three metrics (a, b, c) & the mobility profiles of users 08^* and $08^\#$ (d, e) & user location-semantic similarity (f, g)

5 Evaluation

Our aim of the evaluation is to check whether our metric can accurately capture the real similarity between users in practice as well as its consistency with the basic principles.

The Dataset. We explore a real-life dataset of GPS trajectories collected by Yonsei University in Korea to evaluate the effectiveness of our metrics. It consists of 1,865 daily trajectories from 12 users, which cover a total length of 32,626 km. Although users moved in different cities or even countries, we focus on their local movements in Seoul. We select six users in terms of the number of their trajectories. We construct another two additional users based on the dataset so as to help compare the performance of different metrics. One is based on the user with the identity 08 in the dataset by dividing the user into two (08^* and $08^\#$) since the user has different movement patterns in two different periods. The other two users (12^* and $12^\#$) are derived from user 12 by evenly dividing his trajectories into two parts.

When constructing users' mobility profiles, we adopt the approach of Chen et al. [10]. In their approach, stay points are first detected from trajectories, which represent the places a user are likely to have stayed in the trajectories. Then all stay points are clustered into RoIs by a hierarchical clustering algorithm. Trajectories of stay points are then transformed into traces of RoIs from which user's mobility profiles are constructed using a trajectory pattern mining tool [3]. The minimum support value is set to 0.1 in the experiments. Due to the page limit, we omit the values of other related parameters.

In Figure 2, we show the mobility similarities and location-semantic similarities between all pairs of selected users by the three metrics – MSTP, MTP, and CPS. We use different grey levels to distinguish the similarity values between users. A darker cell indicates that the corresponding pair of users are more similar.

User Mobility Similarity. We have mentioned that the MTP-based metrics have been validated in the literature as effective ways to quantify users' similarity using ground-truth data. Thus, by comparing the similarity values of our metric to the values of these two metrics, we can verify whether our metric correctly assesses user similarity.

In general, we have two main observations. First, if users are ordered according to their similarity values to a same user, our metric will output a similar order to the other two metrics. We take the similarity value of a user to another user computed with a given metric as a variable. Then we can calculate the covariance of two variables of a user with regard to different metrics. A positive covariance will indicate the user's similarity values calculated by the two metrics are consistent. In other words, the similarity of a user to a given user has a similar ranking when calculated with the metrics. On average, the covariances of our metric with respect to MSTP and MTP are 0.09 and 0.04, respectively, which validates our observation that CPS is also consistent with the ground-truth. Second, when comparing Figure 2(b) and Figure 2(c), we observe that for some pairs of users, the similarity values calculated with our new metric have significant differences from the other two metrics. However, after projecting users' original GPS trajectories on the map, we see that the similarity calculated with our metric is more precise. For example, the similarity values between users 08^* and $08^\#$ have a rather large difference among the three metrics MSTP and MTP output 0.41 and 0.69 respectively, while CPS only gives 0.16. We plot their trajectories on the map and present them in Figure 2(d) and Figure 2(e). RoIs are labelled by red rectangles and users' stay points are tagged by yellow dots. Blue lines represent the transition between stay points in a trajectory. We also name the RoIs and put their identities beside them. User $08^\#$ has two more RoIs (E, F) than 08^*. Furthermore, more than 57% of user $08^\#$'s trajectories go through these two RoIs and only about 15% of his trajectories contain RoIs A, B and C. However, about 78% of 08^*'s trajectories contain A, B and C. Therefore, the reasonable similarity value between $08^\#$ and 08^* should be around 0.20 after considering the small proportion of common patterns and the large difference between their support values. By this example, we show that our metric indeed gives rise to a more precise similarity measurement.

User Location-Semantic Similarity. We proceed to illustrate the effectiveness of our metric when adding location semantics in user similarity calculation. Our major purpose is to check whether our metric can capture users' similarity when location semantics are added, but not to learn the real similarity between the users. Thus, in our experiments we select five location semantic tags, and for each RoI we assign to it a probability distribution over the tags. Since only the metric proposed by Ying et al. [9] can handle location semantics, we show and compare the similarity values calculated by their method and our new CPS-based metric in Figure 2(f) and 2(g). Since the MSTP metric only considers the location semantic tags that an RoI may associate with non-negligible likelihoods, in the implementation of the metric, we set the minimum probability as 0.2 and for each RoI we only consider the subset of location semantic tags with probabilities larger than 0.2. From Figure 2(f) and 2(g), we can see that our metric calculates similar similarity values to MSTP. This means our metric keeps the right ranking between the similarity values of different pair of users as the effectiveness of the MSTP metric has been evaluated in [9]. Compared to the mobility similarity, an interesting

observation is that the similarity between users 08^* and $08^\#$ increases to 0.59 from 0.16. This is mainly because the RoIs E and F have similar distributions to B and C. From the above discussion, we can conclude that our metric not only satisfies all the basic principles but also outputs more precise measurements for users' similarity based on trajectory patterns and location semantics.

6 Conclusion

We have identified a number of principles and proposed new metrics (with/without location semantics), when quantifying users' similarity based on their trajectory patterns. The effectiveness of our metics is illustrated through extensive experiements. In the future, we want to evaluate our metrics on more real-life datasets, especially when considering location semantics. This might lead to more efficient and effective ways to treat semantics in mobility data.

References

1. Crandall, D.J., Backstrom, L., Cosley, D., Suri, S., Huttenlocher, D., Kleinberg, J.: Inferring social ties from geographic coincidences. PNAS 107(52), 22436–22441 (2010)
2. Zheng, Y., Zhang, L., Ma, Z., Xie, X., Ma, W.Y.: Recommending friends and locations based on individual location history. ACM Transactions on the Web 5(1), 1–44 (2011)
3. Giannotti, F., Nanni, M., Pedreschi, D., Pinelli, F., Axiak, M.: Trajectory pattern mining. In: Proc. KDD, pp. 330–339. ACM Press (2007)
4. Rhee, I., Shin, M., Hong, S., Lee, K., Kim, S.J., Chong, S.: On the levy-walk nature of human mobility. IEEE/ACM Transaction on Networking 19(3), 630–643 (2011)
5. Song, C., Koren, T., Wang, P., Barabási, A.-L.: Modelling the scaling properties of human mobility. Nature Physics 6, 818–823 (2010)
6. Brockman, D., Hufnagel, L., Geisel, T.: The scaling laws of human travel. Nature 439(26), 462–465 (2006)
7. Shokri, R., Theodorakopoulos, G., Troncoso, C., Hubaux, J.P., Boudec, J.Y.L.: Protecting location privacy: optimal strategy against localization attacks. In: Proc. CCS, pp. 617–627. ACM Press (2012)
8. Xiao, X., Zheng, Y., Luo, Q., Xie, X.: Finding similar users using category-based location history. In: Proc. GIS, pp. 442–445. ACM Press (2010)
9. Ying, J.C., Lu, H.C., Lee, W.C., Weng, T.C., Tseng, S.: Mining user similarity from semantic trajectories. In: Proc. SIGSPATIAL, pp. 19–26. ACM Press (2010)
10. Chen, X., Pang, J., Xue, R.: Constructing and comparing user mobility profiles for location-based services. In: Proc. SAC, pp. 261–266. ACM Press (2013)
11. Chen, X., Pang, J., Xue, R.: Constructing and comparing user mobility profiles. TWEB (accepted, 2014)
12. Bray, J.R., Curtis, J.T.: An ordination of upland forest communities of southern Wisconsin. Ecological Monographs 27, 325–349 (1957)
13. Zhao, Z., Yan, D., Ng, W.: Mining probabilistically frequent sequential patterns in large uncertain databases. IEEE Transaction on Knowledge and Data Engineering (2013) (preprint)
14. Ye, M., Shou, D., Lee, W.C., Yin, P., Janowicz, K.: On the semantic annotation of places in location-based social networks. In: Proc. KDD, pp. 520–528. ACM Press (2011)