Exploiting Spatiotemporal User Behaviours for User Linkage

Wei Chen[1], Hongzhi Yin[2], Weiqing Wang[2]
Lei Zhao[1], Wen Hua[2], Xiaofang Zhou[2]

[1] School of Computer Science and Technology, Soochow University, China
[3] School of ITEE, The University of Queensland, Brisbane, Australia

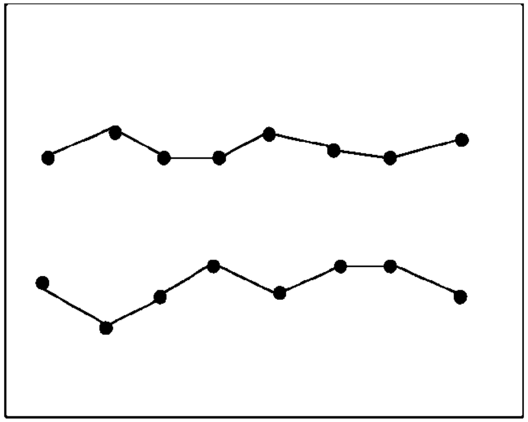
Soochow Advanced Data Analytics Lab
苏州大学先进数据分析研究中心

Outline

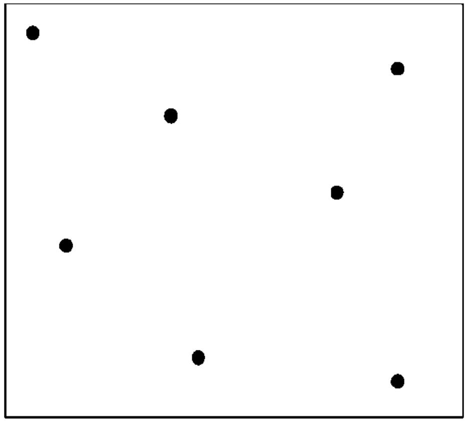
- Introduction
- Problem Statement
- Feature Extraction
- Similarity Measure
- Experiments
- Conclusion

Introduction

- The proliferation of GPS-enabled devices and mobile techniques has led to the emergence of large amount of spatiotemporal information.
  - Trajectory data: adjacent points of a trajectory are sampled in a short time period.
  - Discrete check-in data in social network: the time between two check-ins is usually large.



Trajectory



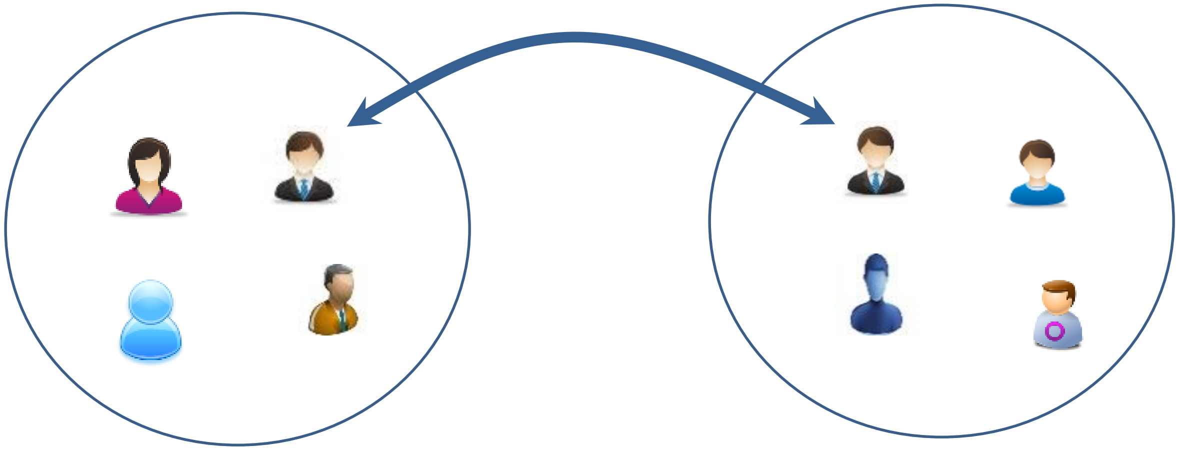
Discrete check-in record

Introduction

- Spatiotemporal data based studies:
  - Route planning in road networks
  - Activity trajectory recommendation
  - Understand human mobility pattern
  - ……
  - Cross-domain user linkage with spatiotemporal data [1]
  - [1] C. Riederer, Y. Kim, A. Chaintreau, N. Korula, and S. Lattanzi. Linking users across domains with location data: Theory and validation. In WWW, 2016, pp. 707–719.

Introduction

- Cross-domain user linkage: link the same user across different domains



Domain A   Domain B

- Example: Facebook---Twitter

Problem Statement

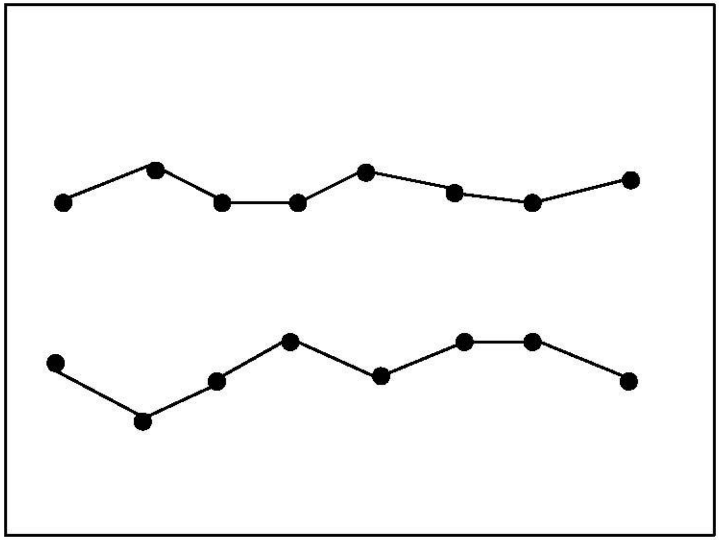
- Spatiotemporal record
  - A spatiotemporal record on both trajectory data and check-in data is defined as: $d = (u, lat, lng, t)$
  - u: the unique id of a user
  - lat: latitude of the record
  - lng: longitude of the record
  - t: timestamp of the record
- Example

adnys.txt *

adnys,	34.0553261066,	118.246986866,	07:50:22
adnys,	34.0314662928,	118.462771922,	08:15:33
adnys,	38.5494890477,	121.740014302,	12:29:09
adnys,	33.1258538651,	117.311940422,	15:41:15
adnys,	34.259089,	116.867585778,	06:27:12
adnys,	34.1838325292,	118.275179542,	11:34:44
adnys,	21.2819466313,	157.836713791,	16:26:08
adnys,	33.9847729546,	118.449375629,	20:11:25
adnys,	37.7896414689,	122.394288182,	21:43:17

Problem Statement

- Two kinds of important data
  - Check-in data, which can be used to extract features directly.
  - Trajectory data, which needs preprocessing before extracting features.



Trajectory

Problem Statement

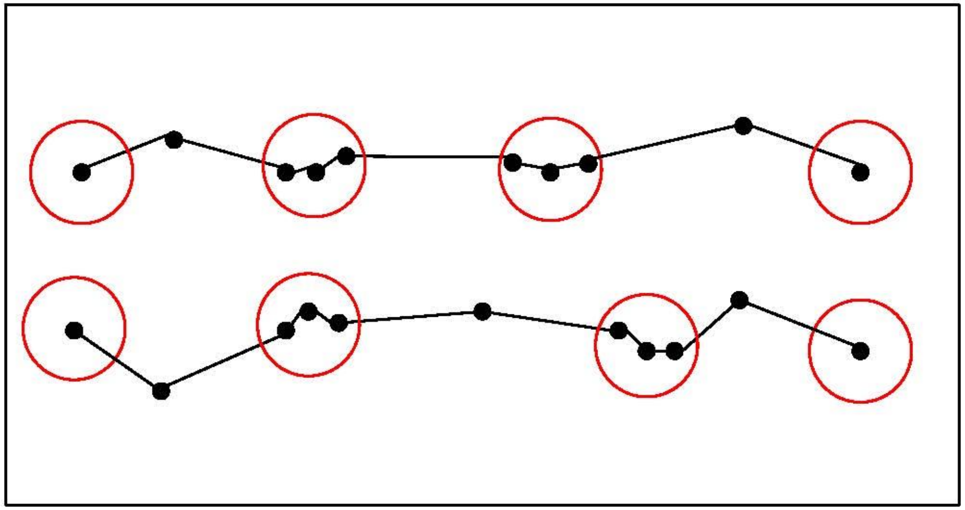
- Stay point [2]: a stay point s stands for a geographic region where a user stayed over a certain time interval.
  - Given a trajectory $\tau = (p_1, p_{2,} \cdots, p_n)$, if there exists a group of consecutive points $P = (p_i, p_{i+1}, \cdots, p_j)$ of τ such that $\forall i < k \leq j$, $Distance(p_i, p_k) \leq \delta_d$ and $|p_j.t - p_k.t| \geq \delta_d$ then we have a stay point s in the form of

$$(s.lat, s.lng) = (\frac{\sum_{k=i}^{j} p_k.lat}{|P|}, \frac{\sum_{k=i}^{j} p_k.lng}{|P|})$$

  - [2] Y. Zheng, L. Zhang, X. Xie, and W. Y. Ma. Mining interesting locations and travel sequences from GPS trajectories. In WWW, 2009, pp. 791-800.

Problem Statement

- Stay region candidate point
  - Given a trajectory $\tau = (p_1, p_{2,} \cdots, p_n)$, the start point p_1, the end point p_n, each point of P is defined as stay region candidate point, denoted as r_c.
- Example



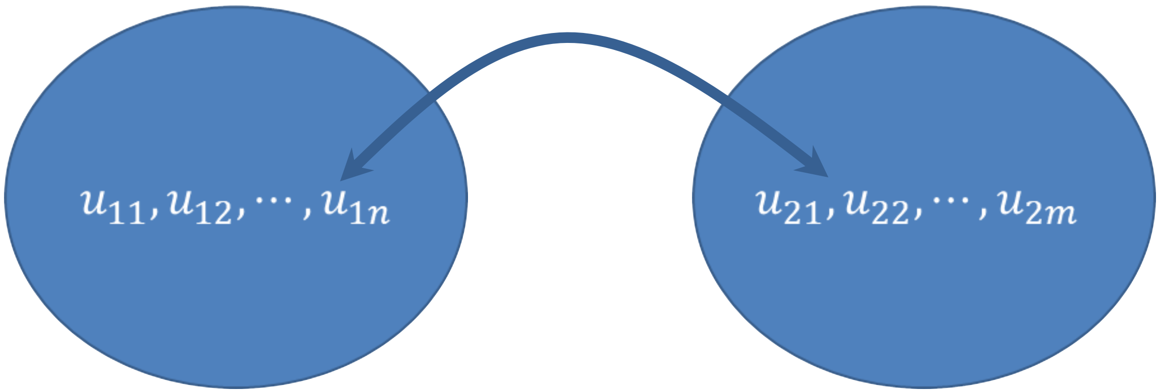
Trajectory

Problem Statement

- Semantics behind the check-ins and stay region candidate points:
  - Shopping mall
  - Home region
  - Work region
  - Bus station
  - …… …

Problem Statement

- Formulation: Given user sets $U_1 = \{u_{11}, u_{12}, \cdots, u_{1n}\}$ and $U_2 = \{u_{21}, u_{22}, \cdots, u_{2m}\}$, where each user is associated with a set of spatiotemporal records, we aim at finding linked user pairs across these two domains.



domain A domain B

User Linkage

- Extract features
- Measure user similarity

Feature Extraction

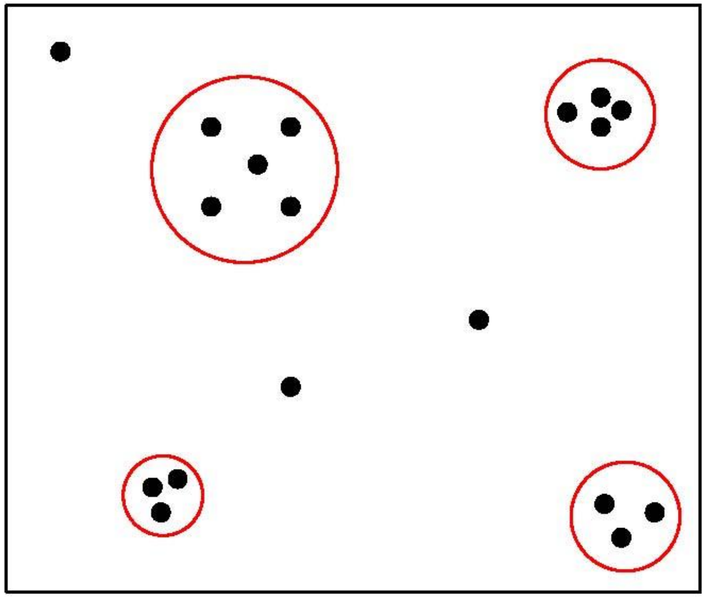
- Features
  - Stay region distribution
  - Global time distribution
  - Local time distribution

Feature Extraction

- Stay region distribution [3]
  - $p = \sum_j \chi\left(d_{r_c^i, r_c^j} - d_c\right), \begin{cases} \chi(x) = 1, & if\ x < 0 \\ \chi(x) = 0, & otherwise \end{cases}$
  - $\delta = \begin{cases} \min\limits_{p_{r_c^j} > p_{r_c^i}} (d_{r_c^i, r_c^j}), & if\ p_{r_c^j} > p_{r_c^i} \\ \max\limits_j (d_{r_c^i, r_c^j}), & otherwise \end{cases}$
  - [3] A. Rodriguez and A. Laio. Clustering by fast search and find of density peaks. Science, vol. 344, no. 6191, pp. 1492-1496, 2014.

Feature Extraction

- Example



Feature Extraction

- Region weight calculation.
  - In real life, many people tend to visit popular areas, such as the downtown of a city, a large bus station, and a popular cinema. Obviously, the importance of the extracted stay regions are diverse.
  - Highlight the individual region.
  - Lighten the popular region.

Feature Extraction

- Region weight calculation.

(a) User Region

User	Region
u_1	$(R_1^1, \cdots, R_1^l)$
u_2	$(R_2^1, \cdots, R_2^k)$
$\cdots$	$\cdots$
u_n	$(R_n^1, \cdots, R_n^m)$

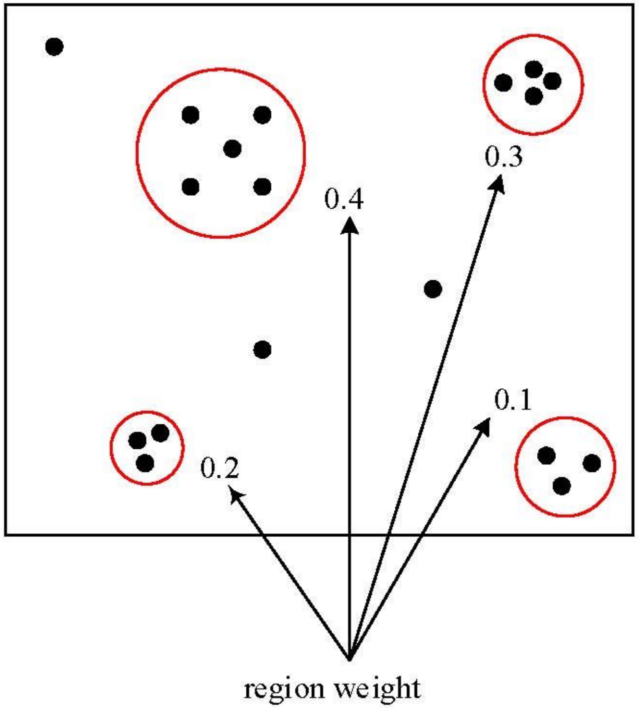
(b) Region Weight

Weight
$\{\omega(R_1^1), \cdots, \omega(R_1^l))\}$
$\{(\omega(R_2^1), \cdots, \omega(R_2^k))\}$
$\cdots$
$\{\omega(R_n^1), \cdots, \omega(R_n^m))\}$

$$\omega(R_1^i) = \frac{\dfrac{N}{1 + \sum S(R_1^i, R_o)}}{\sum \dfrac{N}{1 + \sum S(R_1^i, R_o)}}$$

Feature Extraction

- Example



- Note: the points outside the region are omitted.

Feature Extraction

- Spatiotemporal features
  - Stay region distribution
  - Global time distribution
  - Local time distribution

Feature Extraction

- Global time distribution
  - We extract the temporal features from the global perspective, where the stay region factor is omitted.
  - Given a set of stay region candidate points $(r_c^1, r_c^2, \cdots, r_c^n)$ of a user u, the Expectation Maximization (EM) algorithm is used to find optimal parameters with timestamp set $(r_c^1.t, r_c^2.t, \cdots, r_c^n.t)$.
- Example

adnys.txt *

adnys,	34.0553261066,	118.246986866,	07:50:22
adnys,	34.0314662928,	118.462771922,	08:15:33
adnys,	38.5494890477,	121.740014302,	12:29:09
adnys,	33.1258538651,	117.311940422,	15:41:15
adnys,	34.259089,	116.867585778,	06:27:12
adnys,	34.1838325292,	118.275179542,	11:34:44
adnys,	21.2819466313,	157.836713791,	16:26:08
adnys,	33.9847729546,	118.449375629,	20:11:25
adnys,	37.7896414689,	122.394288182,	21:43:17

Feature Extraction

- Global time distribution
  - E-step: the probability of the sample $r_c^i.t$ generated by the cluster $(\mu_k, \sum_k)$ is:

$$\gamma_{ik} = \frac{\alpha_k N(r_c^i.t|\mu_k, \sum_k)}{\sum_{j=1}^{K} \alpha_k N(r_c^i.t|\mu_j, \sum_k)}$$

  - M-step: the maximum likelihood method is used to update model parameters as follows:

$$\alpha_k = \frac{1}{n}\sum_{i=1}^{n} \gamma_{ik}$$

$$\mu_k = \frac{\sum_{i=1}^{n} \gamma_{ik} r_c^i.t}{\sum_{i=1}^{n} \gamma_{ik}}$$

$$\sum_k = \frac{\sum_{i=1}^{n} \gamma_{ik}(r_c^i.t - \mu_k)^2}{\sum_{i=1}^{n} \gamma_{ik}}$$

Feature Extraction

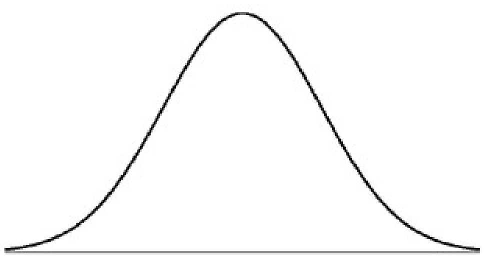
- Global time distribution

(a) Time Cluster

User	Time Cluster
u_1	$(T_1^1, \cdots, T_1^l)$
u_2	$(T_2^1, \cdots, T_2^k)$
$\cdots$	$\cdots$
u_n	$(T_n^1, \cdots, T_n^m)$

(b) Time Cluster Weight

Weight
$\{\omega(T_1^1), \cdots, \omega(T_1^l))\}$
$\{(\omega(T_2^1), \cdots, \omega(T_2^k))\}$
$\cdots$
$\{\omega(T_n^1), \cdots, \omega(T_n^m))\}$



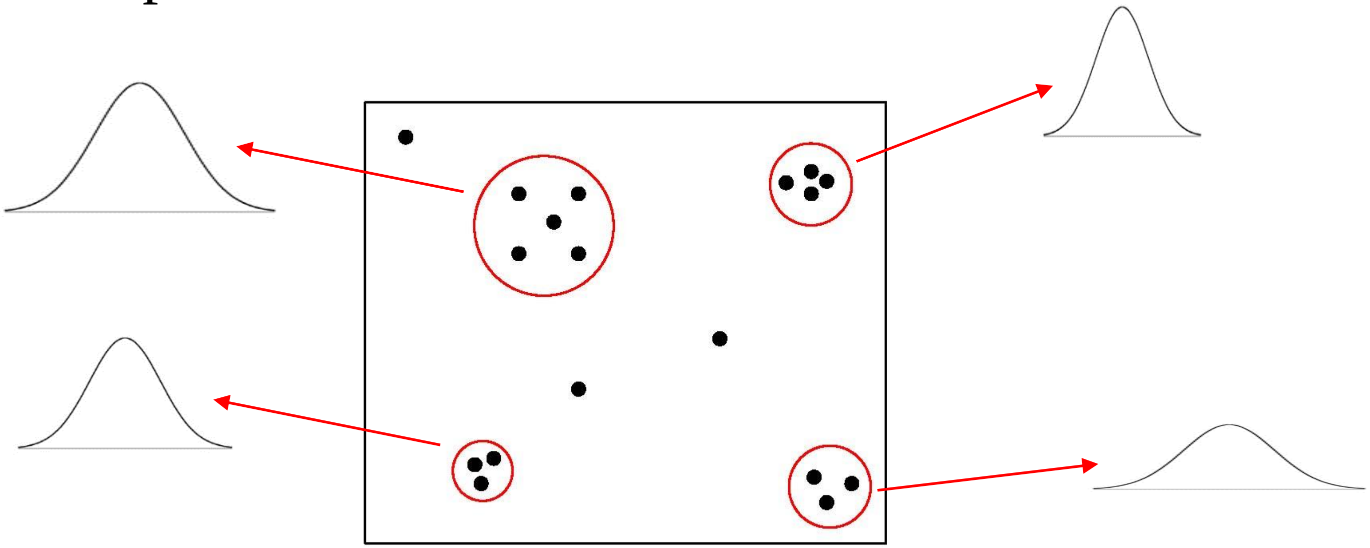
$$\omega(T_1^i) = \frac{\dfrac{N}{1 + \sum S(T_1^i, T_o)}}{\sum \dfrac{N}{1 + \sum S(T_1^i, T_o)}}$$

Feature Extraction

- Spatiotemporal features
  - Stay region distribution
  - Global time distribution
  - Local time distribution

Feature Extraction

- Local time distribution
  - We use the same method to extract time clusters and calculate corresponding weights in each stay region.
- Example



User Linkage

- Extract feature
- Measure similarity

Similarity Measure

- Stay region similarity
  - Assume the stay regions of u_1 and u_2 are:

    $\{(R_1^1, \omega(R_1^1)), (R_1^2, \omega(R_1^2)), \cdots, (R_1^m, \omega(R_1^m))\}$

    $\{(R_2^1, \omega(R_2^1)), (R_2^2, \omega(R_2^2)), \cdots, (R_2^n, \omega(R_2^n))\}$

  - The stay region similarity $S(u_1, u_2)_r$ is defined as:

    $$S(u_1, u_2)_r = \sum_{i=1}^{m} \sum_{j=1}^{n} S(R_1^i, R_2^j)\omega(R_1^i)\omega(R_2^j)$$

Similarity Measure

- Global time similarity.
  - Assume the global time clusters of u_1 and u_2 are:

    $\{(T_1^1, \omega(T_1^1)), (T_1^2, \omega(T_1^2)) \cdots, (T_1^k, \omega(T_1^k))\}$

    $\{(T_2^1, \omega(T_2^1)), (T_2^2, \omega(T_2^2)) \cdots, (T_2^l, \omega(T_2^l))\}$

  - The global time similarity $S(u_1, u_2)_t$ is defined as:

    $S(u_1, u_2)_t = \sum_{i=1}^{k} \sum_{j=1}^{l} S(T_1^i, T_2^j)\omega(T_1^i)\omega(T_2^j)$

Similarity Measure

- Local time similarity.
  - Assume the time distribution in a stay region $(R_1^i, \omega(R_1^i))$ of u_1 is $\{(T_1^1, \omega(T_1^1)), (T_1^2, \omega(T_1^2)) \cdots, (T_1^k, \omega(T_1^k))\}$, in a stay region $(R_2^j, \omega(R_2^j))$ of u_2 is $\{(T_2^1, \omega(T_2^1)), (T_2^2, \omega(T_2^2)) \cdots, (T_2^l, \omega(T_2^l))\}$, the local time similarity in these two regions is defined as:

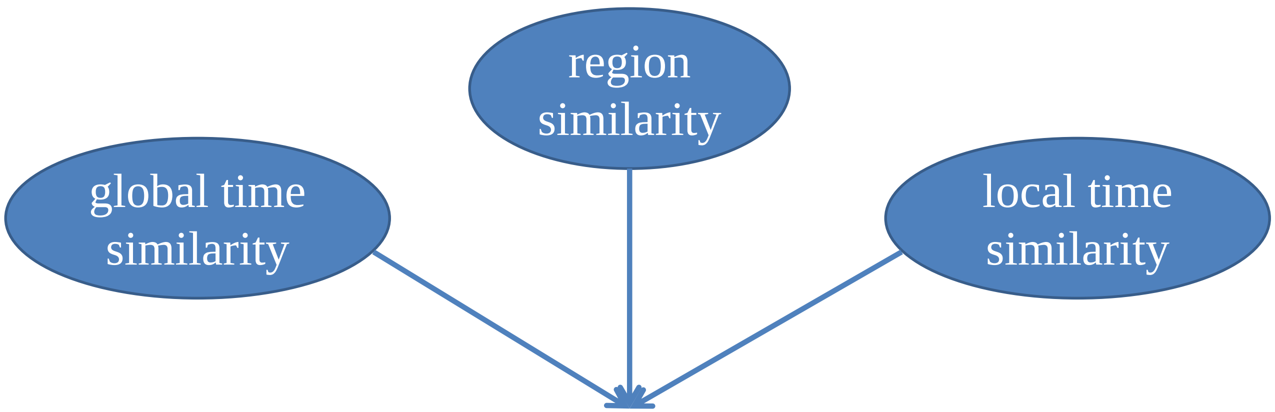
$$S(R_1^i, (R_2^j)\omega(R_1^i)\omega(R_2^j) \sum_{i=1}^{k} \sum_{j=1}^{l} S(T_1^i, T_2^j)\omega(T_1^i)\omega(T_2^j)$$

  - The local time similarity between u_1 and u_2 is defined as:

$$S(u_1, u_2)_{rt} = \sum_{i=1}^{m} \sum_{j=1}^{n} (S(R_1^i, R_2^j)\omega(R_1^i)\omega(R_2^j) \sum_{i=1}^{k} \sum_{j=1}^{l} \cdot S(T_1^i, T_2^j)\omega(T_1^i)\omega(T_2^j))$$

Similarity Measure

- Finally, the similarity between u_1 and u_2 is defined as:



$$S(u_1, u_2) = S(u_1, u_2)_r + S(u_1, u_2)_t + S(u_1, u_2)_{rt}$$

Experiments

- Dataset
  - Beijing Walk Trajectories (BJW) -- Beijing Car Trajectories (BJC)
  - Foursquare (FS) -- Twitter (TW)
  - Instagram (IT) -- Twitter (TW)

Dataset	Domain	Users	Trajectories	Locations/Check-ins
BJW-BJC	Walk	182	14337	2190957
	Car	182	5475	925380
FS-TW	Foursquare	282	-	7832
	Twitter	282	-	88820
IT-TW	Instagram	1066	-	283740
	Twitter	1066	-	284051

Experiments

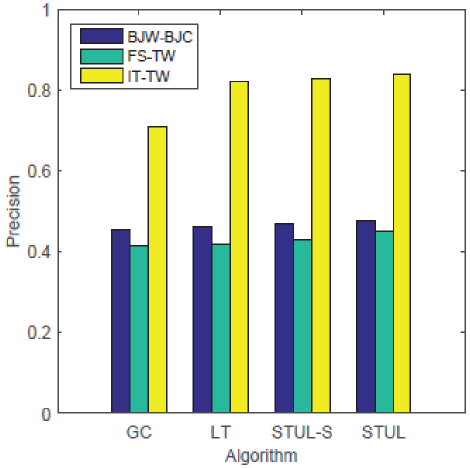
- Compared methods:
  - **GC**: Each user is denoted by a set of grid cells.
    - Z. Li, B. Ding, J. H, R. Kays, P. Nye. Mining Periodic Behaviors for Moving objects. In KDD, 2010, pp. 1099-1108.
  - **LT**: Each user is presented by a set of bins.
    - C. Riederer, Y. Kim, A. Chaintreau, N. Korula, and S. Lattanzi. Linking users across domains with location data: Theory and validation. In WWW, 2016, pp. 707–719.
  - **STUL-S**: A simplified version of STUL, where the extracted features are directly used to measure the user similarity.
- Our approach:
  - **STUL**

Experiments

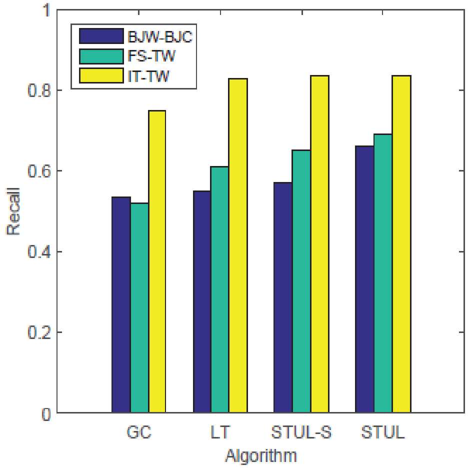
- Evaluation metrics:
  - $precision = \frac{k}{n}$
  - $recall = \frac{k}{m}$
  - $F1 = \frac{2*Recall*Precision}{Recall+Precision}$

Experiments

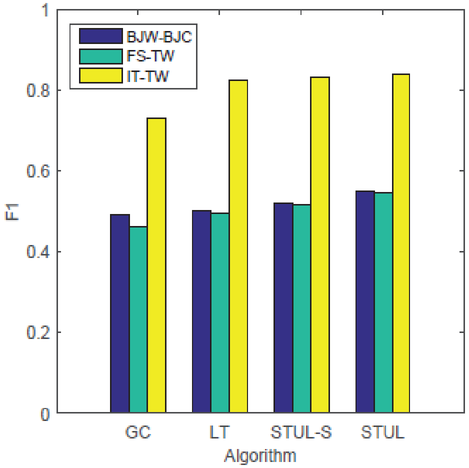
- Performance of the proposed algorithms in different datasets



(a) Precision



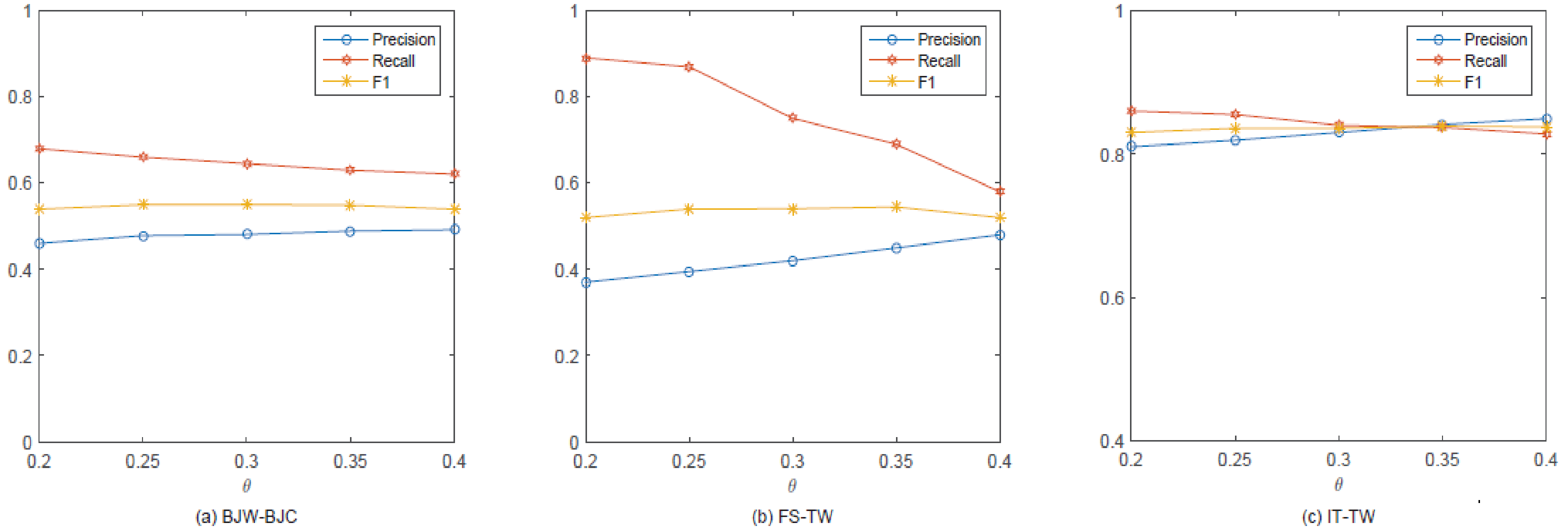
(b) Recall



(c) F1

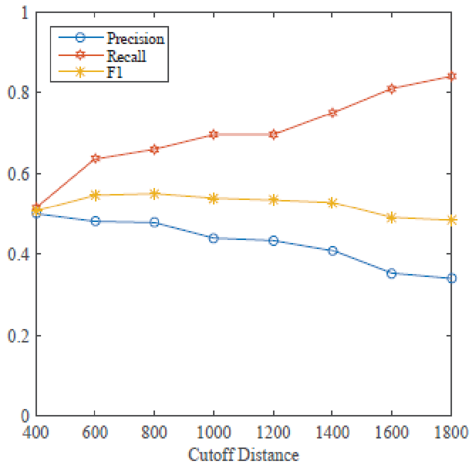
Experiments

- Performance of STUL w.r.t varied θ

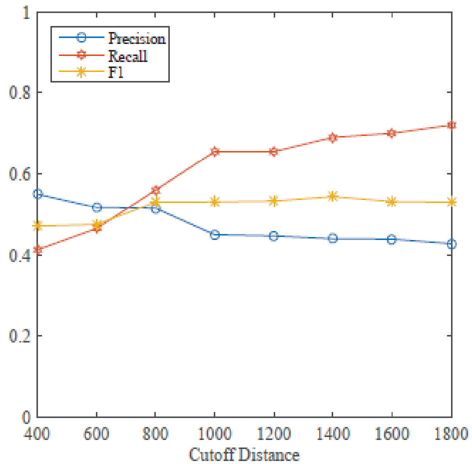


Experiments

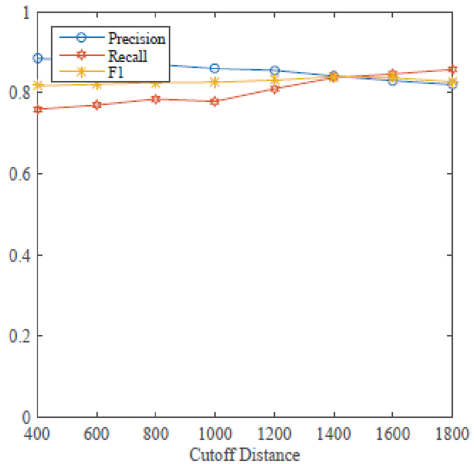
- Performance of STUL w.r.t. varied cutoff distance



(a) BJW-BJC



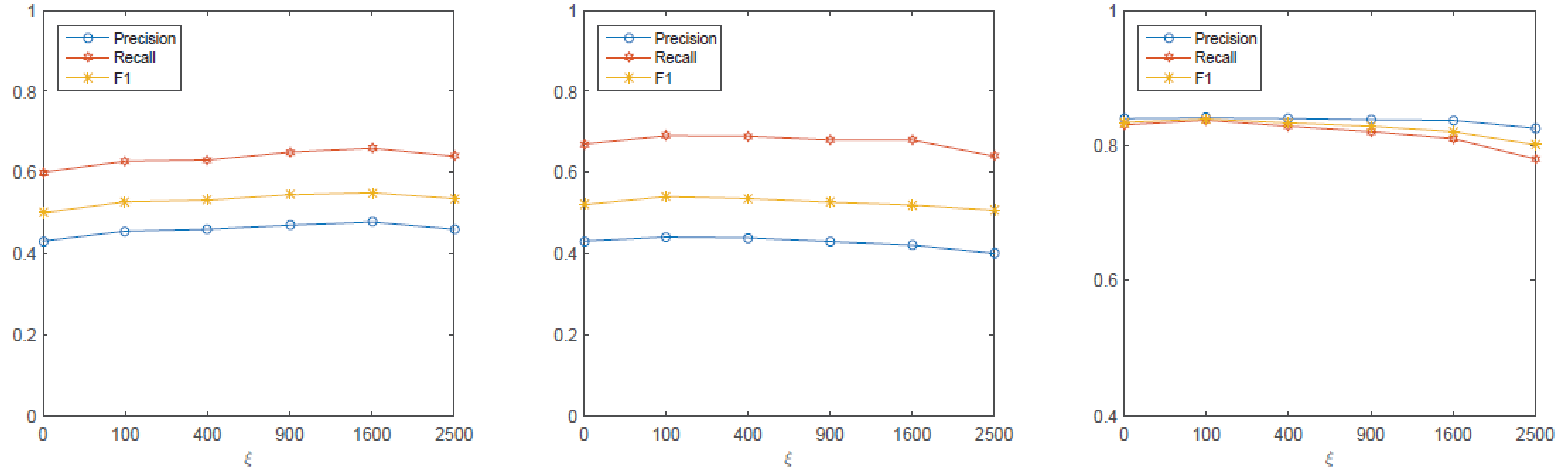
(b) FS-TW



(c) IT-TW

Experiments

- Performance of STUL w.r.t. varied ξ



(a) BJW-BJC (b) FS-TW (c) IT-TW

Conclusion

- To connect the actually linked users from different domains with spatiotemporal data, we propose the novel model STUL.
  - From spatial perspective, a density-based method is developed to extract stay regions that a user will visit repeatedly.
  - From temporal perspective, we use GMM to extract the time distribution. Based on these features, we measure the similarity between users. The real-world dataset based experiments demonstrate the high performance of STUL.

Thank You

Q & A