D-017

移動軌跡データを対象とした効率的な類似検索手法

石塚 淳† 鈴木 優‡ 川越 恭二‡

立命館大学大学院理工学研究科† 立命館大学情報理工学部‡

1 はじめに

近年，GPS (Global Positioning System) などの位置計測機器の発達により，移動時の位置や時刻等の情報を容易に得られるようになった．このため，観光分野において位置情報を用いた様々なサービスを観光客に提供することができ，観光ルートを決定する際の支援を行うことが可能になると考えられる．例えば，観光客にとって初めて訪れた場所では，過去に訪れた観光客の行動履歴は重要である．そのため，過去の訪問者の移動軌跡データから，利用者に対しての行動支援をするためのサービスが提供可能となると考える．つまり，利用者の移動軌跡と類似した移動軌跡データが存在した場合，その移動軌跡データは利用者にとって有用なものであると考えられる．

しかし，利用者の移動軌跡と類似した移動軌跡データの検索を行う場合，検索対象データの量が多い場合には，時間のかかる処理となる．また，検索の高速化のためにデータの間引きを行った場合には，検索精度が低下するという問題点がある．このため，移動軌跡データを検索するためには，検索精度が低下することなく，計算時間を短縮する必要がある．

そこで本稿では，観光客向けの行動支援サービスを想定し，大量の移動軌跡データから利用者の観光ルートと類似したデータを高速に検索するために，スルーエリアと呼ぶ範囲を提案する．さらに，スルーエリアを用いた類似検索手法を提案する．観光地には観光客の辿るルートの多くに含まれ，観光客がたくさん集まるポイントが存在する．そのポイントを含む範囲をスルーエリアとして検索対象データの絞り込みに利用し，検索対象データの削減を行う．また，スルーエリアの大きさを変化することによって，検索精度の低下を抑える．提案手法により，移動軌跡データへの検索精度を低下させることなく検索速度を高速化することができる．

2 提案手法

2.1 提案手法での類似度の定義法

本稿で扱う移動軌跡データとは，利用者の位置情報を要素とした時系列データである．また，時系列データは同じ要素数を持つと仮定する．従来の研究 [1] では，時系列データ間の類似度の定義にユークリッド距離が用いられることが多い．そこで本研究でも，二つの行動履歴間のユークリッド距離を基にして時系列データ間の類似度とする．

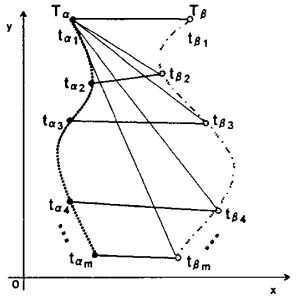

図 1:本手法での類似度定義法

任意の時系列データ $T_i(i=1,2,\cdots,I)$ は次のように定義する．時系列データを $T_i=[t_{i1},t_{i2},\cdots,t_{im}]$ とする．ここで，$t_{ij}\,(j=1,2,\cdots,m)$ とは 2 次元上の軌跡データの通る点であるため，2 次元ベクトル $t_{ij}=(x_{ij},y_{ij})$ と表現できる．二つの移動軌跡データ $T_\alpha,T_\beta\ (\alpha,\beta=1,2,\cdots,I\,|\,\alpha\neq\beta)$ 間の類似度は以下のとおりである．

$$D(T_\alpha,T_\beta)=\sum_{j=1}^{m}\sqrt{(x_{\alpha j}-x_{\beta j})^2+(y_{\alpha j}-y_{\beta j})^2} \qquad (1)$$

しかし，(1) 式では，ある行動履歴と逆の順序を持った行動履歴の類似度が低下するという問題が発生する．そこで，この問題に対応するため (1) 式を拡張し，以下のように類似度を新たに定義する．移動軌跡データ間の類似度は，移動軌跡データに含まれる全ての要素間のユークリッド距離を算出し，その最小の値を利用して定義する．拡張方式による類似度算出の手順を図 1 を用いて説明する．

1. T_α と T_β の要素間の全ての組み合わせにおけるユークリッド距離を求める．

$$D(t_{\alpha k},t_{\beta j})=\sqrt{(x_{\alpha k}-x_{\beta j})^2+(y_{\alpha k}-y_{\beta j})^2} \qquad (2)$$

ただし，k は j と同様に移動軌跡データの要素番号とする．(2) 式で求めた D の値を k 行 j 列の行列 $\boldsymbol{D}$ で表現する．ここで行列 $\boldsymbol{D}$ の k 行 j 列の要素 $\boldsymbol{D}_{kj}$ は次式となる．

$$\boldsymbol{D}_{kj}=D(t_{\alpha k},t_{\beta j}) \qquad (3)$$

2. $\boldsymbol{D}$の要素 $D(t_{\alpha k},t_{\beta j})$ の中の最小値を D'_{min} とする．また，その最小値となる行列$\boldsymbol{D}$の k,j の要素に関して以下の処理を行う．

$$D(t_{\alpha p},t_{\beta j})=\infty\ (p=1,2,\cdots,m) \qquad (4)$$

$$D(t_{\alpha k},t_{\beta q})=\infty\ (q=1,2,\cdots,m) \qquad (5)$$

Efficient Similarity Search for Trajectory Data
Jun ISHIZUKA, Yu SUZUKI and Kyoji KAWAGOE
†Graduate School of Science and Engineering, Ritsumeikan Univ.
‡Faculty of Information Science and Engineering, Ritsumeikan Univ.

3. Dの全ての要素の値が ∞ になるまで 2. を繰り返す．ここで i 回目の繰り返しで求めた D'_{min} を D'_i とする．
4. 3. で求めた D' の総和を求め，$T_\alpha T_\beta$ 間の類似度 $S(T_\alpha, T_\beta)$ とする．

$$S(T_\alpha, T_\beta) = \sum_{i=1}^{m} D'_i \quad (6)$$

$S(T_\alpha, T_\beta)$ の値が小さいほど $T_\alpha T_\beta$ 間のデータの類似度が高いと考えることができる．

2.2 問題点

2.1 節で説明を行った拡張方式を用いて類似度を計算した場合，多くの計算量が必要となる．移動軌跡データ T_α について，大量の移動軌跡データの中から，T_α と最も類似度の高い移動軌跡データ T_β を検索する場合，全ての移動軌跡データと T_α について類似度を計算する必要がある．T_α の要素数が m，移動軌跡データ数が n であれば計算量は $O(mn)$ となり，移動軌跡データ数に比例して計算量が増加する．そのため，検索対象の移動軌跡データ数の削減は効率的に検索を行うために重要な問題となる．そこで，本研究ではこの問題点を解決するため，スルーエリアを用いた類似検索手法を提案する．

3 スルーエリアを用いた類似検索手法

3.1 スルーエリア

スルーエリアとは実際の観光客の行動において多く訪れるポイントを範囲として設けたものである．京都を例にした観光地では，京都駅などの主要な駅，たくさんの観光客の訪れる金閣寺，清水寺等の有名な観光スポットがあるが，その観光スポットをスルーエリアとして定義する．図 2 の場合，$\{1, 2, \cdots, 5\}$ の四角形がスルーエリアである．観光客の移動軌跡データを線で表し，実線を問合せデータ，点線を検索対象データとする．図 2 の例では問合せデータが三つのスルーエリアを通過している．そのため，この三つのスルーエリアを通過する検索対象データだけに対して検索処理を行うことにより，データ数を削減し，計算量の増加を抑える．

スルーエリアの位置は観光での利用を想定し，システム利用者によってあらかじめ与えられているものとする．スルーエリアの形には，一辺の長さが w である正方形を用いる． w の値により検索処理時間，精度が変化する．つまり，w の値を小さくすることにより検索精度が低下する場合がある．一方，w の値を大きくすると，計算量が増加することが考えられる．したがって，計算量と精度のトレードオフとなる．すなわち，精度の低下を抑え，検索時間を短縮させるための，最適な w の値を設定することが重要である．そこで，検索対象となる移動軌跡データの削減を行い，検索時間を短縮するため，初めに精度を維持した最小の w の値とする．

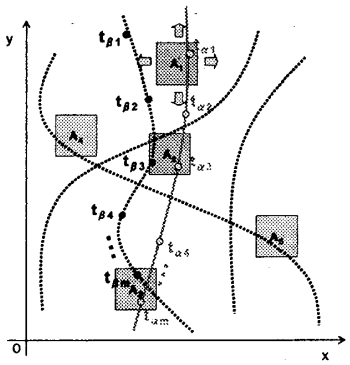

図 2:スルーエリアを用いた類似検索手法

3.2 スルーエリアを用いた検索手法の手順

スルーエリアを用いた検索手法の手順を図 2 を用いて説明を行う．サイズが w ，p 個のスルーエリアを $A_k\,(k = 1, 2, \cdots, p)$ とする．ここで実線の T_α を問合せデータ，点線の T_β を検索対象データの一つとする．また，その他の点線のデータも検索対象データとしている．すなわち，実線の問合せ移動軌跡データと類似度の高い検索対象データを点線の中から検索することとなる．検索の手順は以下のようになる．

1. T_α が含まれているスルーエリアを求める．
   ここで T_α は A_1, A_2, A_3 を通過している．
2. 1. で求めたスルーエリアが含まれている移動軌跡データのみを検索対象へと絞り込む．ここでは検索対象データの各要素が一つでも 1. で検索されたスルーエリアを通過していれば検索対象へと絞り込まれることとなる．
3. 2. により絞り込まれたデータから 2.1 節の拡張した方法で距離を求め類似度を求める．

4 まとめと今後の課題

本研究では利用者の観光ルートにおける移動軌跡データと特徴が類似している移動軌跡データを，大量のデータの中から効率的に検索するために，スルーエリアを用いた類似検索手法を提案した．本手法により，検索精度の低下を抑えた状態で検索時間を短縮することが可能となる．今後は，本稿で述べた類似度の定義の方法と他の研究での時間を考慮した類似度定義の方法 [2] との比較を行う．さらに，スルーエリアのための索引構造の手法を考案し，既存の 2 次元時系列データの索引を用いた検索手法との検索精度，検索時間の比較を行う．評価実験はサンプルデータではなく実際のデータを使って行い，本手法の有効性を実証する予定である．

参考文献

[1] E.J.Keogh, et al. “Dimensionality Reduction for Fast Similarity Search in Large Time Series Databases”. *Knowledge and Information Systems 3(3)*, pp. 263–286, 2001.

[2] 柳沢豊ほか. “移動軌跡データに対する類似検索手法”. *FIT2002*, pp. 37–38, 2002.