D-19

移動軌跡データに対する類似度検索手法
Shape-based Similarity Query for Trajectory Data

柳沢 豊† 赤埴 淳一† 佐藤 哲司†
Yutaka Yanagisawa Junichi Akahani Tetsuji Satoh

1. はじめに

近年，GPSなどのデバイスにより継続的に取得されたユーザの移動軌跡データを分析することで，ユーザの行動の特徴を抽出しようとする研究が進んでいる．たとえば，オフィス内を移動する多くのユーザの移動軌跡を分析すれば，オフィスの構造を最適化するための情報を得られると考えられる．同様に，車の移動軌跡データから交通渋滞の予測や，道路の新設の必要性などを検討する手がかりが得られる．

従来より，こうした時間の経過とともに値の変化する各種のセンサデータや信号を管理するための，時系列データベースの研究が行われている [1]．その中で，時系列に沿った値の変化の様子が似ているデータを探す，類似検索の技術が提案されている．しかし，従来の時系列データベースでは移動軌跡の空間上での形状に基づく類似検索を行うことができない．例えば，あるユーザが広場を歩いた軌跡とよく似た軌跡をもつ別のユーザを探す，というような検索は行えない．

そこで本論文では，まず空間上の二つの離散的な移動軌跡データの類似度を定義し，サンプリング間隔の異なる移動軌跡のデータの粒度をそろえる方法について述べる．さらに，この類似度に基づいてる移動軌跡を効率的に探すために，時系列データベースで用いられている手法を拡張した新しいインデックス作成方法を提案する．また，このインデックスの有無による検索速度の比較実験についても述べる．

2. 移動軌跡データ

空間 $\boldsymbol{R}^2$ 上の質点の移動軌跡は，時間 t の連続関数 $\lambda(t) = (x, y) = \boldsymbol{x}$ と表せる．また一般に移動軌跡は始点と終点があるので，関数の値域を時区間 $\boldsymbol{T}_\lambda = [t_S, t_E]$ と定義する $(t_S < t_E)$．t_S, t_E の値は移動軌跡によって異なる．一方，実際にGPSなどの位置取得デバイスから得られる移動軌跡データは，離散的な時刻列 $\boldsymbol{T}_{\dot\lambda} = \langle t_1, t_2, \ldots, t_m \rangle$ にそれぞれ測定されたベクトル列 $\dot\lambda = \langle \boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m \rangle$ になる（図1左）．一般に，移動軌跡データのサンプリング間隔 $\Delta t = t_{i+1} - t_i$ は，センサのスペックにもよるが一定にならないことが多い．しかし，移動軌跡データ間の距離を考える上では，このサンプリング間隔を揃える必要性が生じる．そのため，離散的な移動軌跡データ λ の不連続部分を折れ線近似を用いて線形に補完し，連続的な移動軌跡 $\tilde\lambda$ を計算する方法を与える．

$$\tilde\lambda(t) = \begin{cases} \dot\lambda(t) & (t \in \boldsymbol{T}_{\dot\lambda}) \\ \frac{t-t_i}{t_{i+1}-t_i}\dot\lambda(t_i) + \frac{t_{i+1}-t}{t_{i+1}-t_i}\dot\lambda(t_{i+1}) & (t \notin \boldsymbol{T}_{\dot\lambda}) \end{cases}$$

ただし $t_i < t < t_{i+1}$

†日本電信電話株式会社 NTT コミュニケーション科学基礎研究所

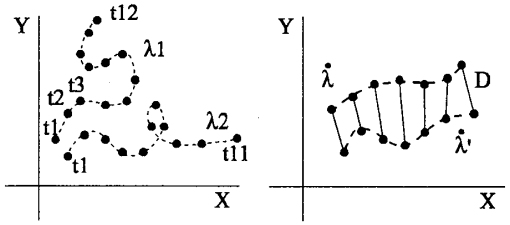

図 1: 移動軌跡データと類似度

なお，$\dot\lambda(t_i)$ は，時刻 t_i に測定された位置座標ベクトル $\boldsymbol{x}_i$ を返す関数である．

3. 移動軌跡データの類似度

時系列データベースでは，m 個の要素をもつ2つの数列 $\boldsymbol{c} = \langle w_1, w_2, \ldots, w_m \rangle$ と $\boldsymbol{c}' = \langle w'_1, w'_2, \ldots, w'_m \rangle$ の間の類似度として，m 次元空間上でのユークリッド距離 $D(\boldsymbol{c}, \boldsymbol{c}') = \sqrt{(w_1 - w'_1)^2 + \ldots + (w_m - w'_m)^2}$ が用いられることが多い [1]．これを拡張し，時系列データの各要素 w_i を $\boldsymbol{R}^2$ 上のベクトル $\boldsymbol{x}_i$ に置き換えると，離散的な移動軌跡データ間の類似度を定義できる．すなわち，$\dot\lambda$ と $\dot\lambda'$ の類似度 $D(\dot\lambda, \dot\lambda')$ は，$\boldsymbol{R}^2$ 上のベクトル $\boldsymbol{x}, \boldsymbol{x}'$ の距離 $D(\boldsymbol{x}, \boldsymbol{x}') = \sqrt{(x - x')^2 + (y - y')^2}$ を用いて，

$$D(\dot\lambda, \dot\lambda') = \sqrt{\sum_{i=1}^{m} D(\boldsymbol{x}_i, \boldsymbol{x}'_i)^2}$$

と定義することができる（図1右）．ただし実際には，2章で述べたようにセンサのサンプリングレートは不定なので，$\dot\lambda$ と $\dot\lambda'$ のサンプリングレートが揃うように，再サンプリングしてからこの式を適用する必要がある．

そこで本論文では，各移動軌跡データ $\dot\lambda$ をまず2章で述べた方法によって補完して $\tilde\lambda$ を作成し，これを一定の時間間隔 Δt で再サンプリングした移動軌跡データ $\dot\lambda_{\Delta t} = \left\langle \tilde\lambda(t_1), \tilde\lambda(t_1 + \Delta t), \ldots, \tilde\lambda(t_1 + k\Delta t) \right\rangle$ を予め用意しておく方法を使う．つまり，データベースでは，位置取得センサから得た座標データをリアルタイムに線形補完し，一定時間間隔 Δt で再サンプリングした状態で蓄積する．こうすることで，時間的な条件を揃えた状態で移動軌跡データの形状の類似度を計算できる．

4. 類似度検索のためのインデックス

ある時系列データ $\boldsymbol{c}$ について，データベースに蓄積されているすべての時系列データの中から，$\boldsymbol{c}$ ともっと

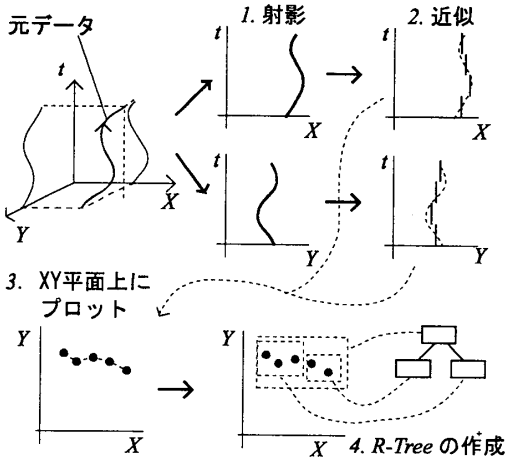

図 2: 2D-PAA を用いたインデックス

も類似度の高いデータ c' を探す場合，すべてのデータについて $D(c,c')$ を計算する必要がある．c の要素数が m，移動軌跡の数が n であれば計算量は $O(nm)$ となる．

ここで，文献 [1] で述べられている次元圧縮手法 PAA (Piecewise Aggregate Approximation) によれば，c の各要素の平均 $\bar{w}$ と c' の各要素の平均 $\bar{w}'$ に関して $D(c,c') > |\bar{w} - \bar{w}'|$ となることが証明されている．この事実を利用すると，あるデータ c との距離が θ 以下のデータを探すには，それぞれの平均値の差が θ 以下になるようなデータ群をまず探し，それらのデータ群から実際に距離が θ 以下のデータを探せばよい．つまり，平均値を用いることで探索するデータ数を絞ることができる．

本論文ではさらにこの技術を拡張した 2D-PAA を用いて，移動軌跡データ λ の類似度検索に適用できるようにした．まず図 2 に示すように，移動軌跡データを X, Y それぞれの平面上に射影し，その平面上で時間軸に沿って N 個ずつのデータの平均値を計算する．次にこれを再度 XY 平面上にプロットし，各点に対して R-Tree を作成する．ここで，あるクエリ λ_q との距離が θ 以下のデータを探すときは，まず λ_q の各要素について N 個単位で平均を計算しておき，R-Tree を用いて差が θ 以下になるような点群を探し出す．その後に，それらの点群に対応する移動軌跡データと λ_q について距離計算を行い，距離が θ 以下のデータを見つけ出せばよい．

5. 評価

実験のために，グラウンドのような道路のない平面上を自由に歩く人間の移動軌跡に近い移動軌跡データを，シミュレーションによって作成した．このデータに対してインデックス付けおよび検索の実験を行った[‡]．データは各々1000 個の座標データを含む．変化させたパラメータは，クエリの移動履歴の長さ (要素数/Length) と，データベース内の座標データ数 (=移動軌跡のデータ数 × 各移動軌跡データの要素数/Points) の二つである．これらを変化させながら，クエリにもっとも類似しているデータを探すときの検索速度を計測した．インデックスを使う場合の結果を図 3 に，使わない場合の結果を図 4 に示す．その結果，インデックスを用いると平均で約 7 倍高速に検索できることが分かった．またデータ数が増加するよりも，クエリの長さが長くなった時のほうがインデックスの効果は高く現れた．

[‡]使用 PC のスペックは Pentium III 700MHz/Memory 512MB, OS は Windows 2000 である

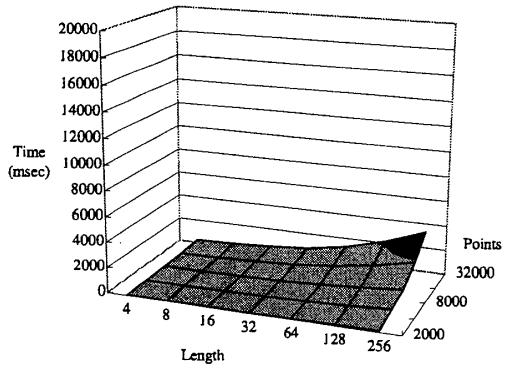

図 3: インデックスありの場合

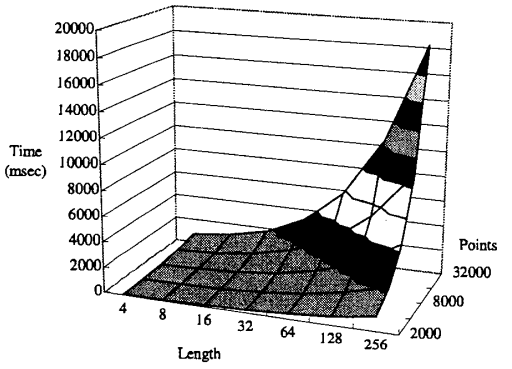

図 4: インデックス無しの場合

6. まとめ

本論文では，移動軌跡データを空間上を移動する質点の座標データを時系列多次元ベクトルとして表し，ベクトル間のユークリッド距離を使って軌跡の形状の近似度を定義した．また，与えられた移動軌跡に似た形状をもつ移動軌跡をデータベース内から効率よく見つけ出す方法についても述べた．

参考文献

[1] E. J. Keogh, K. Chakrabarti, M.J. Pazzani and S. Mehrotra: Dimensionality Reduction for Fast Similarity Search in Large Time Series Databases, Knowledge and Information Systems 3(3): pp.263–286(2001).

[2] H.D. Chon, D. Agrawal, and A.E. Abbadi: Query Processing for Moving Objects with Space-Time Grid Storage Model, In Proc. of the 3rd Intl. Conf, on Mobile Data Management, pp.121–129 (2002).