報告番号	※甲　　第　　号

主　論　文　の　要　旨

論文題目　Indexing, Retrieval, and Compression of Moving Objects in Networks: A String Processing Approach
（ネットワーク上の移動オブジェクトに対する索引，検索および圧縮：文字列アルゴリズムによるアプローチ）

氏　　名　　小出 智士

論 文 内 容 の 要 旨

自動車などの移動オブジェクトから収集される位置情報の時系列（軌跡）データは近年の通信コストの低下に伴い大規模化してきている．このような軌跡データは，従来よりも高度なナビゲーションシステム，自動運転開発，交通行動分析など，データ駆動型のアプリケーションにとって重要なコンポーネントである．これらのアプリケーションは，目的地までの所要時間の短縮，自動車の安全性の向上，渋滞の解消といった社会的な課題の解決に寄与し得る，という点で重要なものであり，現在も活発に研究開発が行われている．

大規模な軌跡データを用いることによって機械学習技術や統計モデルの信頼性が向上するが，一方でそのような大規模軌跡データをどのように格納し，また必要となったデータをどのように検索するのか，というデータマネジメントの方法の確立は重要な工学的課題である．本論文では，上述したようなアプリケーションを広く想定した上で，軌跡データに対する検索やデータ格納などの，基盤的な操作に関する情報技術を提案し，実データを用いた評価を通してそれらの有効性を確認する．

本論文におけるアプローチの特徴として，通常は緯度経度（二次元空間上の点）の列として表現される軌跡データをグラフの道路ネットワークのエッジ（道路エッジ）の列として表現する，という点がある．このような表現を用いることには，以下で述べる二つの利点がある．

第一の利点は，道路エッジの列として表現することで，ユーザが検索する際に，経路を指定することが可能になる，という点である．従来，軌跡データの索引化技術は多くのものが緯度経度の単位でデータを管理しており，ユーザに許された典型的な問合せは「ある矩形状の範囲に存在した軌跡を検索する」というものであった．一方で本論文のようなエッジ列表現を用いると，「ある経路を走行した軌跡を検索する」などの，より実用的かつ高度な問合せが可能になる．

第二に，エッジの列（すなわち記号列）として表現し直すことで文字列アルゴリ

ズムの適用が可能になる．本論文全体を通して，文書検索やゲノム情報処理などで開発されてきた文字列アルゴリズムを,移動オブジェクトの軌跡データに適用する,というアプローチをとるが，既存アルゴリズムの単なる適用にとどまらず，移動軌跡特有の問題設定やデータの特徴を考慮した新しい手法の提案を行った．本論文では大きく分けて二つの技術的貢献を行った.以下ではそれらについて個別に説明し,それらの間の関連性についても述べる.

第一の貢献は，論文中でSNT-index（Suffix-array-based Network-constrained Trajectory index）と呼ばれる，大規模な移動軌跡データに対する索引データ構造の提案である．これを用いることで「ある経路パターンで」「ある時間幅の中で」走行したような移動軌跡をデータベース中から高速かつ正確に検索することが可能になる（この問合せはStrict Path Queryと呼ばれる)． 手法の鍵となるのは FM-indexと呼ばれる文字列索引である．これは，文字列(移動軌跡)に対する高速なパタンマッチングを可能とする，コンパクトなデータ構造である．FM-indexは文字列索引であるため，時刻情報を取り扱うことができない．そこで，時刻情報についてはB+treeを用いて索引化を行い，両者を「逆接尾辞配列」で結びつけることによって，従来法に対して大幅な性能改善を実現した.

SNT-indexを用いることで，Strict Path Query 以外の以下の二つの問合せも実現可能であることを示した．(a) ある経路を，ある時刻に走行した車両がその後走行した経路を列挙する（Trajectory Extraction Query)，(b) ある二地点間を，ある時刻に移動した軌跡が通過した経路をすべて列挙する（Time-period-based All Path Enumeration Query).

提案したデータ構造（SNT-index）および提案したアルゴリズムを,実データを用いて評価を行った．いずれの問合せアルゴリズムも従来法の数十倍から数百倍の高速化を実現できる，という結果となり，提案アプローチの有効性を確認することができた.

一方で,FM-indexは静的な索引であり,頻繁なデータ更新を行うことができない,というデータ構造である．したがって，SNT-indexも頻繁なデータ更新を行うことは苦手である．一方で新しいデータに対して，索引をすべて構築し直す，というアプローチは大規模データに対しては非効率的である.本論文では一定の時間間隔で,すでにある索引構造を破壊することなく，新しいデータの追加を行うための方法を検討し，上述した問題の緩和を行った.

本論文の第二の貢献は，軌跡データの圧縮に関連するものである．上述したFM-indexはインメモリ索引であるため，大規模データを格納する際にはメモリ使用量の問題が生じる．特に，アルファベット集合（文字種類数）が大きなときには効率(圧縮率・検索速度)が低下する，という問題が知られている．移動軌跡では文字種類数は道路エッジの種類数であり，これは典型的には数万～数十万というサイズとなる.これはゲノムや英文テキストなどのFM-indexが想定する典型的なターゲットと比較して非常に大きな数である.

この問題を回避するため，道路ネットワークのスパース性を考慮した新しい手法「相対移動ラベリング」を提案した．相対移動ラベリングは以下のような考え方に基づくネットワーク上の軌跡データの変換である．上述したように，軌跡データは

大きなアルファベット集合をもつような文字列であるが，ある道路エッジを走行している際に，次に移動できる道路エッジの候補は少ない．例えば，十字路交差点を考えると，次に移動可能な道路エッジは物理的に接続した道路エッジであり，その数は典型的には 3 つである．すなわち，一つ前に走行した道路エッジを記憶しておくことで，軌跡データを小さなアルファベット集合上の文字列に変換することができる．小さなアルファベット集合に変換することで，一文字あたりに必要なビット数を大幅に削減することができる．

一方で相対移動ラベリングはオリジナルの文字列を変換してしまうため，従来の FM-index で可能であった高速パタンマッチングができなくなるのでは，という懸念がある．この問題に関しても本論文では肯定的な解を示す．FM-index は本質的には文字列中の各文字をその接尾辞の順序でソートする，というものである．提案手法では接尾辞の順序を破壊しないように FM-index を構築したあとに相対移動ラベル化を適用し，それを PseudoRank と呼ばれる提案アルゴリズムと組み合わせることで，従来の FM-index と全く変わらない機能を持つアルゴリズムを構築できることを示した．

提案手法の有効性を示すために，本論文では（1）情報理論的な解析（2）実データによる評価実験，を行った．第一に，情報理論的な解析では，相対移動ラベリングが，ある条件のもとで文字列のエントロピーを最小化するような変換になっていることを示した．文字列のエントロピーは圧縮率および PsuedoRank の計算時間に比例する量であるため，相対移動ラベリングが理論的にも妥当性の高いものである，ということが明らかになった．第二に，実データによる実験においても，提案手法は圧縮率・検索アルゴリズムの処理速度の両面で，従来手法を大幅に上回る性能となることを確認することができた．特に，提案手法の検索速度は，圧縮しているにも関わらず，圧縮していない FM-index（従来法）よりも高速化する，ということが確認できた．この結果は上述した理論解析の結果と整合するものである．

以上で述べた，本論文の貢献（検索および圧縮）をまとめると以下のようになる．

- ネットワーク上の軌跡データの索引化手法として，従来用いられてこなかった文字列索引（FM-index）を用いた方法を提案し，軌跡が通過した経路に基づく種々のクエリを効率的に処理するためのアルゴリズムについても提案した．
- 軌跡データを FM-index に格納した際の圧縮率およびパタンマッチングの処理速度の低下を抑制するために，道路ネットワークの構造（スパース性）に着目した新しい手法を提案し，情報理論的解析，および実データを用いた評価の両面で既存手法を大きく上回る圧縮率，および処理速度を実現可能である，ということを示した．

最後に，将来の研究の方向性について述べる．提案手法を拡張することで，本論文で考慮しなかった問合せへの対応も可能であると期待できる．また，本論文では FM-index を用いたが，本論文で考慮していない文字列アルゴリズムを用いることで，新しい軌跡問合せが可能になるかもしれない．また SNT-index と同様の問合せをサポートしつつ，動的更新が可能な索引構造の研究も興味深い問題である．別の方向性として，本論文で提案した手法を軌跡データ以外へ応用する，ということも可能かもしれない．例えば，相対移動ラベリングは道路ネットワーク構造に動機づけられた方法ではあるが，シンボル間の遷移がスパースであるものであれば常に適用可能である．そのようなアプリケーションの拡大も興味深い研究の方向性である．