

GPS データからの重要位置の高速検出アルゴリズム

Algorithm for Detecting Significant Locations from Raw GPS Data

加美伸治^{*1,*2}
Nobuharu Kami

榎本敦之^{*3}
Nobuyuki Enomoto

吉川隆士^{*1}
Takashi Yoshikawa

馬場輝幸^{*1}
Teruyuki Baba

森川 博之^{*2}
Hiroyuki Morikawa

^{*1} NEC システムプラットフォーム研究所
System Platforms Research Laboratories, NEC Corporation

^{*2} 東京大学先端科学技術研究センター
RCAST, The University of Tokyo

^{*3} NEC ビッググローブアプライアンス事業開発本部
Appliance Business Development Div., NEC BIGLOBE Ltd.,

1. はじめに

ユーザが長い時間滞留した点は、高い確率で名所やレストランといった GPS データの中で重要な意味をもつ参照点 (以後 waypoint と呼ぶ) となりうる。そのため、GPS データの軌跡の中で一定時間以上滞留していた点を自動抽出する処理は、様々なロケーションウェアなアプリケーションにおいて基本的かつ重要な処理となる。本稿では、GPS データから重要な位置を高い検出精度で高速に自動抽出できるアルゴリズムを示す。

2. 従来手法と課題

[1]では、GPS データから、一定時間 t_h 以上の間、最大移動距離が l_h 以内にとどまっているデータセグメントを抽出し、その重心から一番近い点を代表点として選ぶことで waypoint を抽出している。しかしながら、GPS データの各データポイント間の距離と l_h との比較を行う必要があり、 N 個のデータに対して $O(N^2)$ アルゴリズムになってしまう。また、GPS データには測位誤差が含まれるが、[1]では測位誤差に対する耐性が考慮されていない。すなわち、 l_h の決定は、測位誤差レベルに応じてなされなければならないが、測位誤差は場所ごとに異なり、適切に l_h を設定することが難しい。

3. アルゴリズム

上記の課題を解決するために、筆者らはノイズ耐性が強く、 $O(N)$ 時間で計算可能な waypoint 検出アルゴリズムを開発した。本手法の基本的アイデアは、「真の滞留点ではデータポイントが密集している」という密度情報に着目し、Locality Sensitive Hashing (LSH) [2]を利用して GPS データのうち密集領域のみからデータポイントをサンプリングすることである。LSH は任意の 2 点のデータポイントを距離が近いほど高い確率で衝突させるハッシュ関数である。密集領域のデータポイントは互いに距離が近いため、LSH を利用するとこれらのデータポイントは高い確率で同じハッシュ値に衝突する。そこで、このハッシュ値をビンラベル、衝突データポイント数を頻度とするヒストグラムを作成し、頻度の多いビンに登録されたデータポイントを取り出すことで、密集領域のみからデータポイントをサンプリングする。このデータポイントの集合は真の滞留点近傍に分布しているため、その重心を waypoint とすることで真の滞留点を推定できる。この操作にかかる計算量はデータポイントのハッシュ値を計算するだけなので $O(N)$ である。また、真の滞留点から遠く離れたデータポイントは除外されるため、隣接滞留点間クロストークを抑制し検出精度を向上する効果もある。

次に、こうしてサンプリングされたデータポイントの集合を、あらかじめ指定されたクラスター半径 ρ を終了条件に Ward 法を用いてクラスタリングする。クラスタリングを行うのは、真の滞留点から遠いデータポイントの検出精度への影響を取り除くためである。本手法は確率的な手法であるため、真の滞留点から離れたデータポイントも密集領域のデータポイントと衝突するエラーが一定確率 P_{res} で存在する。 P_{res} はパラメータ設定によって望むだけ小さくすることができるが、その分計算時間がかかることになるため、設計者は計算時間と検出精度のトレードオフを考慮してパラメータ設定を行なう必要がある。今回筆者らは、 P_{res} を極限まで小さくする代わりにクラスタリング操作を導入することで、真の滞留点から離れたデータポイントの検出精度への影響を除外し、高い検出精度と高速計算の実現を図っている。クラスター半径 ρ はノイズレベルから決まるユーザの設定パラメータであり従来手法の l_h に対応するが、 l_h に比べてパラメータ設定におけるトランス幅が広いのが特徴である。これは、上記のように既にノイズ分布の”エッジ”部分のデータポイントが除外されているためクラスタリン

グ結果が ρ に大きく依存しないことに起因している。すなわち広範なノイズレベルに対して同じパラメータ値で高い検出精度を実現可能であり、高いノイズ耐性を有しているといえる。

最後に、こうして得られた各クラスターの中で重心が一番近い場所に位置するデータポイントを waypoint として抽出し、真の滞留点の推定点とする。抽出された waypoint の集合を重要度に応じて序列化して表示するために、各滞留点の密集度をスコア値として利用する。各クラスターのデータポイント数と waypoint からの距離の分散はその密集度を反映しているため、データポイント数が多いほど、また分散が小さいほど高順位として抽出された waypoint をソートする。

4. 評価

今回、本アルゴリズムを実際の GPS ログデータに対して適用した。用いたデータは宮古島の旅行時のデータで総データ点数は 1617 点であった。図 1 にその GPS 軌跡と実際に抽出された上位 14 個の waypoint を示す (GoogleMaps[3]を用いて表示)。すべての waypoint のスコア値は、その waypoint が示す場所でのデータポイントの密度を正しく反映していた。また、すべての抽出された場所は実際に立ち寄った観光名所、レストラン、空港、ホテルなどに対応しており、提案アルゴリズムが GPS データの重要な場所を正しく抽出することが確認できた。さらに、それらの抽出にかかった処理時間は 1 秒程度であり、オンラインアプリケーションに適用する上で問題ない応答性を示した。

5. まとめ

GPS データから重要な位置を高速に検出する手法として、ノイズ耐力に優れた $O(N)$ 時間で計算可能な手法を提案した。本手法は密集部分から選択的にデータサンプリングを行うことで実現され、実際の GPS ログデータへの適用により高い検出性能・応答性を確認した。

謝辞: 本研究の一部は、独立行政法人情報通信機構 (NICT) の委託研究「ダイナミックネットワーク技術の研究開発」の成果である。

参考文献

- [1] Hariharan, R., Toyama, K. "Project Lachesis: Parsing and Modeling Location Histories", Geographic Information Science 2004, pp.106-124, (2004)
[2] Indyk, P., Motwani, R. "Approximate nearest neighbors: towards removing the curse of dimensionality". Proceedings of the thirtieth annual ACM symposium on Theory of computing, pp.604-613, (1998)
[3] Google Maps, <http://maps.google.com/>

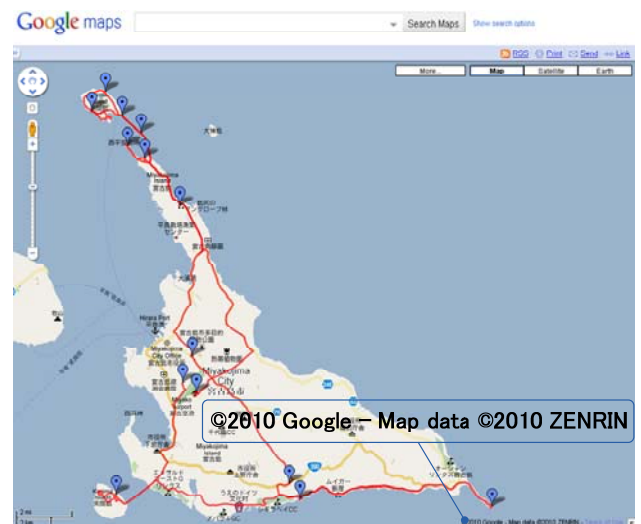


図 1 実際の waypoint 抽出例