5L-08

機械学習による空間索引の性能評価

鈴木 駿也 † 杉浦 健人 †† 石川 佳治 ††

† 名古屋大学情報学部コンピュータ科学科 †† 名古屋大学大学院情報学研究科

1 はじめに

地理情報サービスの普及により，空間問合せ処理の需要が高まっている．例えば，グーグルマップの “ここで検索” は典型的な空間ウィンドウ問合せとして挙げられる．このような空間問合せを効率的に処理するために R 木や kd 木などの空間索引が用いられるが，空間ウィンドウ問合せでは一般に複数のリーフノードをまたいだ検索が必要となり，実行時間の遅延の原因となっている．つまり，B+ 木上での範囲問合せなどのようにリーフノード間を直接行き来できず，対象ノード間の遷移に時間がかかってしまう．

一方，機械学習を用いて索引を構築する RMI（recursive model index）[1] が近年提案され B 木など従来索引を上回る結果を残したことで，幅広い種類の索引への機械学習の適用が行われている．地理情報サービスを扱う際に必須となる空間索引もその一つであり，機械学習を用いた空間索引である RSMI（recursive spatial model index）[2] は，R 木や kd 木を上回る性能を達成した．RSMI は多次元データを空間充填曲線に基づき一次元へ射影しており，B+ 木のようなリーフノード間の遷移も可能とするため，空間ウィンドウ問合せにおけるノード遷移の効率化が期待できる．

しかし，RSMI の性能評価は十分であるとはいえない．そのため本研究では実験を通して RSMI の再検証を行い，その性能特性などについて議論する．

2 関連研究

RSMI は，機械学習で B 木を再現した RMI と，空間データを RMI に適応した ZM（Z-order model）[3] を基に考案された．本章では，それらの 2 つの学習索引と，RSMI について紹介する．

2.1 Recursive Model Index

RMI は B 木などの索引をキーという入力に対しそのデータ位置を返す関数として捉え，その学習を行う．ただし，B 木は全てのキーを索引付けするわけではなく，ページ毎に索引付けを行う．つまり，索引をたどった後にリーフノードの走査が必要となり，B 木はページサイズ分の誤差を含むと考えられる．これに対し，RMI では予測誤差の下限と上限を記憶することで，データ位置の誤差を保証している．

RMI の学習では，一般的な索引中のデータがキーで整列されていることを利用し，データの累積分布関数を学習しデータ位置を予測する．しかし，一つの学習モデルで全範囲の累積分布関数を高精度に予測することは困難である．そのため，RMI は上位から下位へと階層的に学習範囲の狭い学習モデルを構築することで予測精度を向上させた．つまり，RMI は上位の学習モデルから次に使用する下位の学習モデルを再帰的に予測し，リーフモデルでの予測結果を最終的な予測値とする．

2.2 Z-order Model

ZM は，RMI を 2 次元データに適用したものである．RMI は 1 次元のキーの累積分布関数を学習するため，2 次元データの 1 次元への変換が必要である．多次元データを 1 次元に変換する方法として，多次元データの座標から空間充填曲線の値を求める手法が挙げられる．ZM は，空間充填曲線に Z-order 曲線を用いたものであり，キーを Z-order 曲線の値とした累積分布関数を学習する．

ZM は 2 次元データに対する機械学習ベースの索引を実現するが，Z-order 曲線の性質により，累積分布関数が線形から大きく外れるという問題がある．Z-order 曲線によってデータを 1 次元化する際には空間をグリッド分割するためのセル長が必要である．このセル長が小さいとき，つまりより詳細な空間分割を行うと全体のグリッド数が増大し，データの分布がスパースとなり累積分布関数の学習効率が悪化する．一方でセル長が大きい場合一つのセルに複数のデータが存在しうる，つまり複数のデータに同じキーが割り当てられ，同様に索引の効率が悪化する．

3 Recursive Spatial Model Index

RSMI は，ZM の学習をより効率的に行えるよう拡張したものである．以下では，拡張点，学習方法，問合せの実行方法のそれぞれについて述べる．

3.1 ZM からの拡張点

■データの順序づけ 累積分布関数を線形に近付けるためにランク空間への射影を行う．前述のとおり，Z-order 曲線をじかに使用した際の累積分布関数は非線形となり，その学習コスト及び誤差は共に大きい．そこで，全データ数を N として，各データを $N \times N$ グリッドのランク空間へ射影したあとに Z-order 曲線への変換を行う．具体的には，データを x 軸・y 軸それぞれでソートし，その順序を $N \times N$ のランク空間の座標とする．これにより，ランク空間のグリッドには高々 1 点だけ存在し，かつ変換後のキーの最大値は N^2 に抑えられる．結果，直接 Z-order 曲線に変換する場合と比べランク空間上の点の Z-order 曲線値をキーとした累積分布関数は線形に近付き，効率的な学習が可能となる．

Performance evaluation of spatial index with machine learning
Syunya Suzuki†, Kento Sugiura††, and Yoshiharu Ishikawa††
†Department of Computer Science, School of Informatics, Nagoya University
††Graduate School of Informatics, Nagoya University

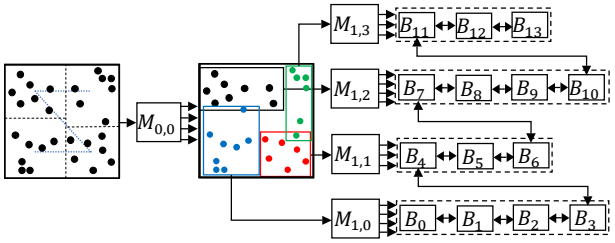

図 1　RSMI の構造

■大規模データへの対応　RMI の紹介で述べたように一つの学習モデルでの大規模データの予測は難しいため，上位から下位へと再帰的にモデルないしデータを予測する．しかし，ZM のように Z-order 曲線上で直接階層的にモデルを構築すると，Z-order 曲線の性質により空間的に離れたデータが同じモデルに割り当てられ予測性能を悪化させてしまう．そこで，RSMI では Z-order 曲線への変換前に学習データセットを空間的に分割する．図 1 にデータセットの分割及び構築された索引の例を示す．まず，データセット中の点の数が均一になるよう空間を分割し，対応する Z-order 曲線の値を教師データとしてモデル（図中 $M_{0,0}$）を構築する．その後，構築したモデルにより各点を分類し，パーティション群として保持する．この空間の分割は再帰的に行われ，パーティション内の点の数が十分小さくなった時点でブロック分類用のモデル（図中 $M_{1,0}$ など）を構築する．なお，ブロックはデータ格納の最小単位であり，全てのブロックは一続きのポインタで繋がれている．最後に，後の問合せ処理のために，学習に用いたデータで再度予測を行い予測誤差の下限と上限を保存する．

3.2　モデル学習

RSMI はデータ数 N 以下の分割に対しサイズ B のブロックへ格納する．つまり，ある点 p の座標から，その点が格納されているブロックを予測する．パーティションのデータ数を N，点 p のランク空間上の曲線値を $p.rank$，ブロックのサイズを B とすると，点 p が格納されているブロック $p.blk$ は以下の式で表される．

$$p.blk = \left\lfloor p.rank \cdot \frac{N}{B} \right\rfloor \tag{1}$$

モデル M は p の座標から格納されているブロックを予測する．p の座標を $p.cord$ とすると，p のブロックは以下のように予測できる．

$$p.blk \approx M(p.cord) \tag{2}$$

RSMI は多層パーセプトロンの勾配降下法を用い，以下の誤差関数 L_M を最小化するように学習する．

$$L_M = \sum_{p \in P} (M(p.cord) - p.bld)^2 \tag{3}$$

3.3　問合せ処理

RSMI は機械学習による空間索引であり，点問合せ，ウィンドウ問合せを処理する．以下ではそれぞれの処理方法を概説する．

■点問合せ　点問合せは，与えらえた点が存在するか検証する問合せである．点問合せは RSMI の上位モデルから下位モデルへ予測を行っていく．非リーフモデルでは与えられた点の座標から下位のパーティションを予測する．リーフモデルでは，座標から格納されているブロックを予測する．非リーフモデルでは予測の下限・上限誤差を記憶しているためその誤差間のブロックを走査し，与えらえた点が見つかれば存在するという結果を，誤差間に見つからなければ存在しないという結果を返す．

■ウィンドウ問合せ　ウィンドウ問合せは矩形の問合せ範囲が与えられ，その範囲内に存在する点を返す問合せである．RSMI は基本的にはデータを Z-order 曲線値の順番に従ってブロックに格納する．Z-order 曲線は左下から右上へと空間を充填する曲線であるため，ウィンドウ問合せの解となる点は問合せ範囲の左下・右上の境界の点の間に格納されている．そのため，まず問合せ範囲の境界の左下・右上の座標で点問合せを行い，それぞれのブロック番号を得る．その後，左下点のブロックから右上点のブロックへ走査し，ブロック中の各点が問合せ範囲内であれば解にその点を追加する．ただし，RSMI は厳密には点座標の Z-order 曲線値順にデータが格納されないため，解は近似的なものとなる．

4　評価分析

2 次元点のデータセットで RSMI 及び ZM の索引構築，問合せを行う．データセットとして uniform，normal，skewed という人工データセットを用いる．uniform は一様分布，normal は正規分布に従うデータセットである．skewed は，y 座標は一様分布に従い，x 座標は y^4 となるデータセットである．各分布のデータセットの点の数を 10 万から 1,000 万まで変化させ，索引構築時間や問い合わせ実行時間について検討する．

5　まとめと今後の課題

本稿では，機械学習により B 木を再現した RMI やそれを 2 次元データへ適用した ZM，そして ZM を拡張し学習を効率化した RSMI の概要について述べた．今後は，4 章で述べたように RSMI の性能評価のための実験を行っていく予定である．

謝辞

本研究は JSPS 科研費（16H01722, 19K21530，20K19804）の助成，及び国立研究開発法人新エネルギー・産業技術総合開発機構（NEDO）の委託業務（JPNP16007）から得られた結果による．

参考文献

[1] T. Kraska, A. Beutel, E. H. Chi, J. Dean, and N. Polyzotis, “The case for learned index structures,” in *Proc. SIGMOD*, pp. 489–504, 2018.
[2] J. Qi, G. Liu, C. S. Jensen, and L. Kulik, “Effectively learning spatial indices,” *PVLDB*, vol. 13, no. 12, pp. 2341–2354, 2020.
[3] H. Wang, X. Fu, J. Xu, and H. Lu, “Learned index for spatial queries,” in *Proc. MDM*, pp. 569–574, 2019.