1W-1

非一様データによる高次元空間での最近傍検索手法の性能評価

金森竜太　秋山裕信　三好涼介　三浦 孝夫
法政大学工学部電気電子工学科

1　前書き

高度データ構造の利用は，特にマルチメディア，意思決定支援システム，データマイニング等で重要であり，このような分野では近似検索 (近傍検索) は不可欠な機能である。対象とするデータは 50 個程度の属性 (次元) を有しており，天文学では 900 にも達する。これまで高次元ベクトル空間での近似検索のための手法が数多く提案されている [1]。例えば空間分割手法 (グリッドファイル，四分木) やデータ分割手法 (R 木，X 木，SR 木) など既に多方面で利用されている。

しかし，これらはすべて，対象データの次元が増加するにつれて性能劣化が報告されており，いわゆる“次元の呪い”現象を呈する [3]。特に，最近傍検索は高次元 (10 次元程度でさえも) では全件探索 (線形走査) より劣化する。

本稿では，次元の呪いの解析が，どの次元にも一様に分布するデータを仮定して解析されていることに注目し，実際の (一様ではない分布の) データでは主張されている程に悲観的に考えなくても良いことを実験で示す。

2　次元の呪い

本研究で対象とするデータ V はベクトル形式 $(v_1, \cdots, v_N)$ であるとし，何らかの方法で V は，そのまま又はこれと同等な表現により格納・管理されると仮定する。格納されたデータは，各次元の値をすべて指定して検索する完全一致検索や一部だけを指定する部分一致検索，一部または全部の次元で区間を与える範囲検索などで操作される。各次元の値をすべて指定して，それに最も近い位置にあるデータを求める検索を”最近傍”検索という。この中でも，最近傍検索はもっとも重要でありながら効率よい実現が難しいものとして知られる。以下では，最近傍検索に限って考察する。

高次元データを何らかの手法で表現・管理するとき，空間分割手法やデータ分割手法の性能を理論的・実験的に解析することができる [3]。データ分布は一様で各次元は独立という仮定の下では，分割手法，クラスタリング手法のいずれにおいても平均性能が 10 次元以上で線形走査に劣ることを示すものであり，最近傍探索では次元数が数十程度の値を超えると全件走査に劣る，という。つまり，どのような多次元空間管理手法であっても，線形走査より性能劣化する次元が存在し，次元の増加によりその入出力はデータ数に比例する値に漸近し，かつ平均的に全てのブロックにアクセスする次元が存在する，ことを意味する。この結果は衝撃的であり，”次元の呪い”と呼ばれる。

しかし，ここで想定している仮定はかなり非現実的である。現実に生じるデータで完全に一様分布するケース (あるいは正規分布に従うランダム性を有して分布するケース) はほとんどない。更にどの次元にも同様に分布する高次元データが生じることはほとんど無い，と考えられる。さらに何も工夫しないデータ構造はありえず，”次元の呪い” 現象は (100 次元以下の) ”通常扱う程度の高次元データ”では現実には問題にならないのではないか，という疑問がある。本研究ではこれを実験により確認する。

3　データ分布

実験のためのデータを生成する。本研究では一様分布および非一様分布に従う 100 次元データ 8 万件からなるファイルをそれぞれ作成する。

本稿で用いる一様分布データは 100 次元空間内に等間隔に格子状に配置した 8 万個のデータである。このため Mersenne Twister により発生させた乱数で 100 次元の属性値を生成し 8 万件のデータを作る。

一方，本稿で使用する非一様 (現実) データは 2 次元の点データ 119,898 点の分布に基づく。これは，各レコードが 14 バイト長 (2 つの 7 バイト long integer 型項目からなる) のニューヨーク，ボストン，フィラデルフィア周辺の郵便アドレス 119,989 件である*。非一様データは，地図情報を縦・横 100 分割した 1 万個のグリッドに分割し，そのデータの偏りを求める。次に，Mersenne Twisiter を用いて 2 つの乱数を発生させ，それに対応する地図情報のグリッドの偏りから数値を得る。この操作を 50 回繰り返し，数値の平均を 1 つの 100 次元データの偏りとする。この操作を 10 万回繰り返し 100 次元データの全体の偏りを求める。この求めた偏りより必要なデータ数 8 万件を生成する。

この結果，一様分布データではすべてのグリッドがデータを 1 件含み，非一様分布データでは 2643 グリッドに 80000 点全てが格納される。

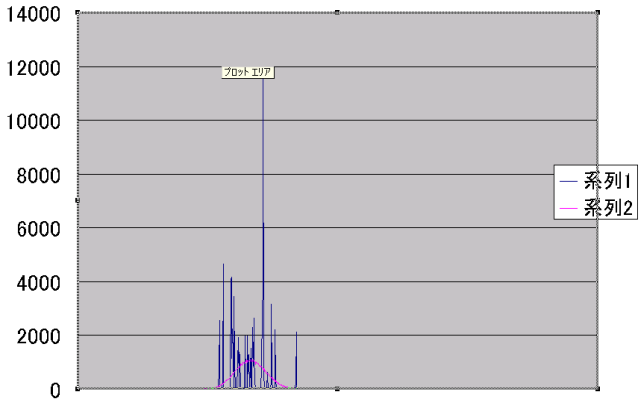
分布の偏り状況を示すため，非一様分布データ (系列 1) と一様データ (系列 2) について，原点からの距離とその件数を示す。これからわかる様に非一様分布データは数箇所に集中して発生しており，異なる性質を有することが確認できる。

4　実験

本稿で指摘する問題を確認するため，FreeBSD 4.6.2 上で前述のデータ 2 種類を用いる。ここでは各点の各次元を 4 分割して番号を割り当て，これを各次元毎に準備する。さらに，各々に振り分けた番号の組み合わせをその点のグリッド番号とする。これらのグリッドデータに最近傍検索を行い，参照されたグリッド数の割合を比較する。実験データとグリッド位置データはメモリ上に展開する。

*) これは `www.rtreeportal.org` で公開されている地図情報を用いている。

”Analyzing Nearest Neighbor Queries on Non-Uniform Disribution Data” : Ryuta Kanamori, Hironobu Akiyama, Ryosuke Miyoshi, Takao Miura : Hosei University, Dept.of Elec. and Elec. Eng.
Kajino-cho 3-7-2, Koganei, Tokyo, JAPAN



最近傍検索は特定のデータ構造を用いることなく行う。すなわち，任意の点をランダムに指定し，その点を含むグリッドとそれに隣接するグリッドを調べる。解が見つからないときは検索範囲を隣接グリッドに広げ，対象が見つかるまでこれを続ける。ここでは，”データを含まないグリッドは検索しないと”いう工夫を加える。この処理を数百回程度繰り返し，その平均割合をそのデータにおける検索性能とする。

実験の結果を表とグラフで示す。一様データは 1000 回の，非一様データは 739 回実行した平均である。

分布	平均参照グリッド数
非一様データ	1463.59
一様データ	1559.54
全件探索非一様分布	2643
全件探索一様分布	80000

実際に参照したグリッドの分布を示す。上段に一様分布のケースを，下段に非一様分布のケースを示す。

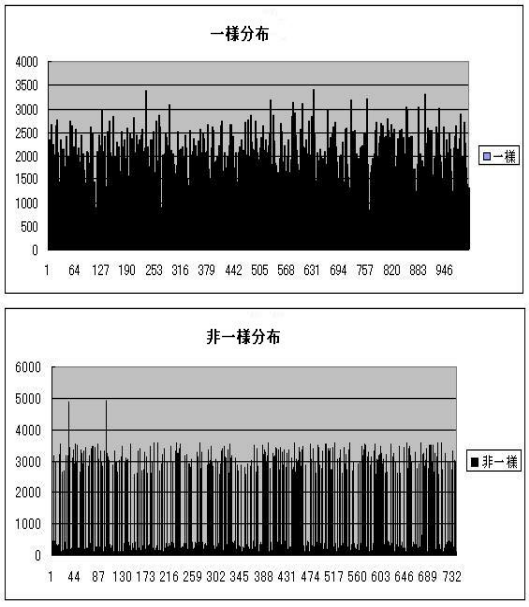


図 1: 一様分布データ・非一様分布データへのアクセス

非一様分布データは一様分布データに比べ 93.8 %の参照回数で検索できている。グリッドデータへの参照回数を比較すると，非一様分布データではアクセスパターンが大きく異なり，一部のグリッドへの負荷の集中が見られ，平均すると一様分布データのものより小さくなる。

この結果は想定する程には大きいものではないが，全件探索と比較するとその差は歴然としている。つまり，一様分布データではすべてのグリッドがデータを 1 件含み，非一様分布データでは 2643 グリッドに 80000 件全てが格納される。データを含まないグリッドの探索を回避した工夫が，全件探索と比べて大きく性能向上させる原因であろう。従って，非一様分布データについては 2643/1463.59 = 1.805 倍の，一様分布データについては 80000/1559.54 = 51.30 倍の性能向上が見られる。いずれにせよ次元の呪い問題は，データの分布および探索上の工夫で改善できることがわかる。

5 結論

本稿では，現実の地図データと同じ分布に基づく高次元データを生成し，次元の呪い現象が報告されている状況ほどには発生しないことを実験により確認した。加えて，さまざまな次元縮小技法が知られている [2]。これらと組み合わせることで更に現実的な処理を実現できる可能性がある。

参考文献

[1] Gaede, V. and Gunther, O.: Multidimensional Access Methods, *ACM Comp. Surveys* 30-2, pp.170-231, 1998

[2] 三好 涼介, 三浦 孝夫, 塩谷 勇”拡張可能グリッドファイルにおける最近傍検索の改善” , 電子情報通信学会論文誌 (D1), 2005 3 月予定

[3] Weber, R., Schek, H.J., Blott,S.: A Quantitative Analysis and Performance Study for Similarity-Search Methods in High-Dimensional Spaces, VLDB, pp.194-205, 1998