2 値量子化に基づく球面分割を用いた類似検索のための縮小構造 Sketch

樋口 直哉 * 篠原 武 * 今村 安伸 *
(*九州工業大学大学院情報工学府先端情報工学専攻)

1 はじめに

計算機の演算能力や記憶容量の向上により，大量のマルチメディアデータを用いたシステムが多く作られている．そのため，膨大なデータの中から必要なデータだけを探し出す情報検索技術が重要である．マルチメディアデータは多くの場合でかなり高次元であり，また劣化や加工も多く，完全一致検索により目的のデータを探すことは難しい．そこで本研究では，多次元データの近似検索として縮小構造 Sketch について検証する．

2 Sketch と近似検索

Sketch[2] は基礎分割関数を用いて作成され，実空間での類似性をある程度保持するようにオブジェクトをバイナリ文字列で表したものである．Sketch 間の相違度はハミング距離を用い，ビット演算による高速な検索が可能である．Sketch は実空間上での類似性を完全には保持しないため，Sketch 上での最近傍解が実空間上での最近傍解と等しいとは限らない．Sketch を用いた k 近傍検索では，まず Sketch をフィルタリングとして用いて K 近傍解を得る．[3] 次にその K 近傍解に対して実距離計算を行うことで k 近傍解を得る．ここで，$K > k$ である．K を大きくすると検索精度は高くなるが，実距離計算コストが増加する．

Sketch の検索精度は基礎分割関数の性能によって決まる．データを均等に分割することで良い基礎分割関数が作成できることが知られている [4] ため，本研究では，データを均等に分割する基準としてデータ中央値を用いた手法を提案し，その効果を検証する．

3 基礎分割関数

基礎分割関数の一つとして，一般化超平面分割（GHP）がある．GHP は，2 点の中心点対の垂直二等分超平面を用いて空間を分割する．バイナリ文字列で表現する際には，どちらの中心点に近いかによって『0』と『1』に分け，長さ m の Sketch を生成するのには m 回，空間を分割する．中心点対の評価関数として，Sketch の衝突を最小にする手法を用いた．図 1 は中心点対 $(P1, P2)$ を用いて，$S = \{A, B, C, D, E, F, G, H, I, J\}$ を分割する例である．

3.1 M-GHP

M-GHP[1] は GHP の中心点選択において，一つの中心点に対してデータ中央値を基準に点対称となる点を二つ目の中心点とし，この二つの中心点との距離によって空間を分割する手法である．

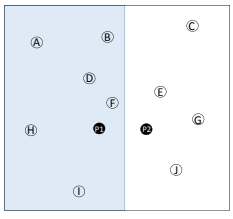

図 1: GHP

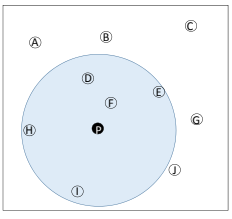

図 2: BP

4 2 値量子化に基づく球面分割 (QBP)

球面分割 (BP) はランダムに選んだオブジェクトを中心点とし、中心点と各データとの距離の中央値を半径として空間を分割する手法である．図 2 は中心点 P を用いて，$S = \{A, B, C, D, E, F, G, H, I, J\}$ を分割する例である．

4.1 QBP

QBP はランダムに選んだオブジェクトの座標値を，中央値を閾値として空間の最小値または最大値に 2 値量子化した点を中心点とし，中心点と中央値との距離を半径として空間を分割する手法である．

5 実験

約 2,900 本の動画から抽出した約 700 万件の画像フレームデータに約 9 万件の質問データを用いて実験を行った．m=64，K=500，k=1 とし，64 ビットの Sketch を作成した．評価関数のスコアである衝突確率と検索における正答率を表 1 に示す．ここで，正答率とは実空間での最近傍解が Sketch を用いて得られた K 近傍解に存在する確率のことである．

表 1: 画像データに対する精度 (正答率)

手法	衝突確率	正答率 [%]
GHP	5.53×10^{-6}	86.13
M-GHP	5.65×10^{-6}	88.56
BP	1.46×10^{-4}	77.76
QBP	5.02×10^{-6}	87.92

6 まとめと今後の課題

BP は衝突確率が高く，正答率も低い結果となった．これは中央値付近のデータの多くが半径の内側となってしまうため，衝突が多くなったのだと考えられる．QBP の衝突確率は他の手法よりも小さくなったものの，正答率では M-GHP に劣る結果となった．本実験には評価関数に最小衝突法を用い，多くの場合では衝突が少ないほど正答率が高くなるが，QBP と M-GHP のようにこの関係が成り立たない場合が存在する．よって別の評価基準の検討が必要である．

参考文献

[1] 村田 敏輔，今村 安伸，篠原 武: 縮小構造 Sketch のためのデータ中央値を基準とする基礎分割関数に関する研究，火の国シンポジウム (2016).

[2] Qin Lv，Moses Charikar, and Kai Li: Image similarity search with compact data structures, In CIKM '04, pp.208–217, New York, NY,USA, 2004. ACM.

[3] Qin Lv, William Josephson, Zhe Wang, Moses Charikar, and Kai Li: Efficient filtering with sketches in the ferret toolkit In MIR '06, , pages279–288, New York, NY,USA, 2006. ACM.

[4] A.J. Muller, T. Shinohara: Efficient Similarity Search by Reducing I/O with Compressed Sketches, *International Workshop on Similarity Search and Applications*, pp.30–38, 2009.