空間索引を用いた近傍点検索に対する近似アルゴリズムによる高速化

青木　隆明*　篠原　武
(九州工業大学大学院情報工学府情報科学専攻)

1　はじめに

マルチメディアデータの検索においては，完全一致検索よりも，似ているものを検索する近似検索の方が需要が多い．本論文では，近似検索を高速化させる近似アルゴリズムを四つ提案し，その解の精度と検索コストの改善率に対する有効性を検証する．索引構造としては，多次元の特徴データを扱えるようにB-treeを拡張したR-tree [1] を採用する．

2　質問方法

近似検索は，質問オブジェクトからの非近似度（距離）により特徴空間内からオブジェクトを取得する過程と捉えることができる．本論文では，質問方法として，質問点から最も距離が小さいオブジェクトを取得する最近傍質問を使用する．

以下に，特徴空間を複数の部分空間に分割し索引付けを行う，一般的な索引構造 (R-treeを含む) における最近傍質問を紹介する．まず初期検索範囲を無限大とし，検索範囲と重複するすべての部分空間を訪問し，その中に含まれるオブジェクトと質問点の距離計算を行う．検索範囲内にあるオブジェクトが見つかれば，質問点とそのオブジェクト間の距離に検索範囲を収縮する．さらに，そのオブジェクトを暫定解に登録する．収縮した検索範囲内に他のオブジェクトが存在しないことが確認できると，暫定解を正式な解として返す．本論文では，検索範囲内に暫定解以外のオブジェクトが存在しないことを確認する作業のことを検索の後始末と呼ぶ．この段階で，既に最適解は発見されている．

3　近似アルゴリズム

近似アルゴリズムは，解の精度をある程度犠牲にして空間索引を用いた近似検索をさらに高速化するものである [2]．本論文では，近似アルゴリズムを適用しない検索手法のことを通常手法，通常手法で得られる解のことを最適解と呼ぶ．ユーザが求める解は必ずしも最適解でなくてよく，検索の途中に得られている解でも許容範囲内である可能性が高いため，近似アルゴリズムは有効である場合が多い．

(a) 距離計算回数限定質問

本手法は，多くの近似アルゴリズムの中で最も素朴な手法の一つである [3, 4]．本手法は，距離計算の回数に制限 f を設け，検索中に距離計算回数が制限回数 f を上回ると，検索を途中で打ち切る．

(b) 検索範囲収縮度打ち切り法

本手法は一定以上の距離計算をしても検索範囲が収縮しない場合は検索の後始末に陥っていると考え，検索を途中で打ち切る．具体的には，検索範囲が収縮した際にそれまでに費やされた距離計算回数を保存しておく．検索時には，現時点の距離計算回数と保存してある距離計算回数の差が事前に設定した閾値 q を超えたら検索を打ち切る．

(c) MBR重複度打ち切り法

R-treeを用いた空間索引では，検索範囲と重複する部分空間をすべて訪問し，その中に含まれるオブジェクトとの距離計算を行う．しかし，重複領域にはオブジェクトが含まれていない可能性もある．本手法は $\alpha(0 < \alpha \leq 1)$ 倍した検索半径において重複する部分空間のみを訪問する．

(d) 適用型距離計算回数限定質問

図1に，最適解との距離に対する検索全体の距離計算回数と解発見時の距離計算回数の関係を示す．検索全体の距離計算の内，解発見時までの距離計算以外は検索の後始末に費やされている．

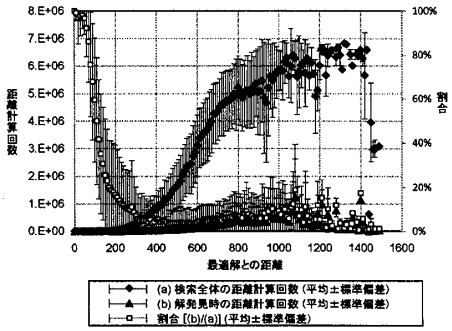

図 1: 検索全体のコストと解発見時のコストの関係

図1より，解との距離が比較的近い場合は検索全体の距離計算も少なく，検索の後始末の割合も少ないが，距離が比較的遠くなると検索の後始末の割合が増加することが分かる．本手法は，この経験則を基に，検索の後始末に陥ったと思われる時点で検索を打ち切る．

*• Accelerating Similarity Search using Spatial Indexes by Approximate Algorithms
• Takaaki Aoki and Takeshi Shinohara
• Department of Artificial Intelligence Kyushu Institute of Technology

具体的には，事前の調査により，図1に示すような解との距離に対する解発見時の距離計算回数を保存しておく．近似検索システムに質問点を連続的に与える場合を考え，質問点の列を $q_0, \ldots, q_i, \ldots, q_n$ とする．質問 q_0 は通常手法の検索を行い，その解を保存する．次回以降の質問 $q_i (i = 1, \ldots, n)$ の際は，検索に先だって検索範囲を q_i と q_0 の解との距離に収縮する．さらに，収縮した検索範囲に対する平均的な解発見時の距離計算回数を学習データより求め，その回数を検索時の制限回数 f とする．また，検索時の検索範囲の収縮に対しても，同様に学習データより解発見時の距離計算回数を求め，制限回数 f を書き換える．

4 実験

データベースには約2,800本の動画から切り出した，約700万件の画像フレーム [5] を登録し，通常手法と四つの近似アルゴリズムを用いて質問を行う．質問用のデータとして，データベースの動画とよく似ている動画 (近似データ)，似ている動画 (準近似データ)，データベースに似ている動画が存在しない動画 (非近似データ) の3種類，約3万フレームずつを用意し，近似アルゴリズムの効果を検証する．

検索の高速化率に対する評価方法として，Improvement in Efficiency(IE) [3] を用いる．$IE = \frac{cost}{cost^A}$ であり，$cost$ は通常手法の距離計算回数，$cost^A$ は近似アルゴリズム A の距離計算回数を表す．一方，解の精度に対する評価方法としては Error on the Position(EP) [3] を用いる．ここで，OX はデータベース中のすべての要素を質問点からの距離の昇順でソートしたリストとする．近似アルゴリズムにより，あるオブジェクト O が解として得られたとすると，$EP = OX(O) - 1$ と定義される．つまり EP は，近似アルゴリズムが通常手法と比較して，順序がどれほど相違する解を返したかということを表す．

4.1 実験結果

各手法の性能を比較するために，同程度の IE が得られるパラメータ値を2箇所選択し，そのときの EP の値を比較する．表1-2に準近似・非近似データの実験結果を示す．近似データは通常手法でも高速に検索でき，近似アルゴリズムの効果は比較的少ないため，結果の掲載は省略する．適用型距離計算回数限定質問では，学習データから制限回数を求める際に，学習データの平均値に標準偏差 (σ) の何倍を加えるかということをパラメータとして与える．

準近似データに対しては，適用型距離計算回数限定質問が他の手法よりも非常に少ない EP を記録した．非近似データに対しては，MBR重複度打ち切り法の EP が適用型距離計算回数限定質問よりもやや低かった．しかし，非近似データの解は近似度が低く，近似検索システムにとってその解の精度が要求される場合は少ないので，EP の僅かな差を懸念する必要はないと考える．以上をまとめると，適用型距離計算回数限定質問の性能が最も良いといえる．

表1: 実験結果・準近似データ

手法	パラメータ値	IE	EP
(a)	$f = 210,000$	5.59	0.17
	$f = 380,000$	3.56	0.067
(b)	$q = 180,000$	5.54	0.12
	$q = 330,000$	3.57	0.056
(c)	$\alpha = 0.40$	5.66	0.11
	$\alpha = 0.49$	3.50	0.029
(d)	$+1\sigma$	5.55	0.053
	$+2\sigma$	3.58	0.023

表2: 実験結果・非近似データ

手法	パラメータ値	IE	EP
(a)	$f = 430,000$	6.61	0.16
	$f = 710,000$	4.07	0.070
(b)	$q = 360,000$	6.58	0.14
	$q = 610,000$	4.09	0.063
(c)	$\alpha = 0.40$	6.53	0.081
	$\alpha = 0.47$	4.09	0.028
(d)	$+1\sigma$	6.65	0.11
	$+2\sigma$	4.09	0.046

5 まとめと今後の課題

本論文では，空間索引を用いた近似検索をより高速化する四つの近似アルゴリズムの提案を行い，それらの有効性を検証した．我々は近傍質問の最後に必要となる検索の後始末という作業に着目し，それに陥ったと思われる時点で検索を打ち切るというアプローチを提案した．実験の結果，解の精度を高水準に保ったまま，約3倍～6倍の高速化を実現することができた．学習データを基に，検索の後始末を排除する適用型距離計算回数限定質問は，同程度のコスト改善率 (IE) で比較した場合に他の手法よりもエラー (EP) を同程度か，若しくはより低く抑えることができた．

今後の課題として，今回は近傍質問に対する性能の評価を行ったので，範囲質問に対する評価も行う必要があると考える．

参考文献

[1] Antonin Guttman. R-trees: A dynamic index structure for spatial searching. In Beatrice Yormark, editor, *SIGMOD'84, Proceedings of Annual Meeting, Boston, Massachusetts, June 18-21, 1984*, pp. 47–57. ACM Press, 1984.

[2] Giuseppe Amato, Fausto Rabitti, Pasquale Savino, and Pavel Zezula. *Region proximity in metric spaces and its use for approximate similarity search.* AACM Transactions on Information Systems. ACM Press, 2003.

[3] Pavel Zezula, Giuseppe Amato, Vlastislav Dohnal, and Michal Batko. *Similarity Search: The Metric Space Approach.* Advances in Database Systems. Springer Verlag, 2006.

[4] 二宮大輔. 実時間動画検索システムの実現-空間索引を用いた最近傍質問の応答時間の均一化-. 九州工業大学 卒業論文, 2005.

[5] 浦郷祐希, 田島圭, 青木隆明, 岩崎瑶平, 篠原武. 空間索引による近似画像の高速検索を用いた動画同定システムの実現. 火の国情報シンポジウム 2009, 2009.