4B-1

メトリック空間における最近傍ペア探索アルゴリズムの高速化

倉沢 央 † 高須 淳宏 †† 安達 淳 ††
† 東京大学大学院 †† 国立情報学研究所

1 はじめに

k 最近傍ペア問い合わせ (k-closest pair query) は，データセットの中から距離の近い k 個のオブジェクトペアを探す検索タスクである．k 最近傍ペア問い合わせは，レコードリンケージやマルチメディアデータベースなどで使われる．例えば,「大学とその最寄駅の対を距離の近いものから 5 件」を探したい時に役に立つ．本研究は，メトリック空間における k 最近傍ペア問い合わせ処理コストの削減を目指している．本手法は距離の公理を満たす類似度すべてを対象とし，ベクトル間のユークリッド距離や文字列間の編集距離などを扱える．

k 最近傍ペア問い合わせの最も単純なアルゴリズムは Nested loops である．データセット中のすべてのオブジェクト間の距離を計算する手法で，N 個のオブジェクトに対して $\frac{N\cdot(N-1)}{2}$ 回の距離計算を必要とする．この距離計算コストを削減すべく，類似検索索引を使った手法 [1] が提案されている．これらの手法では共通して，k 番目の類似ペア間の距離の上限値を，初期値 ∞ から類似ペアを見つけるたびに減少させる．さらに，*Pivot* と呼ぶ参照オブジェクトから各オブジェクトまでの距離と三角不等式を使って，上限値よりも距離が遠いと判断できるオブジェクトのペアを枝刈りする．この枝刈りは，上限値が小さく，Pivot からオブジェクトまでの距離が分散しているときほど効果が大きい．しかしながら，従来手法は，k 番目の類似ペア間の距離の上限値が収束していく特徴を枝刈り手法に利用していなかった．

そこで我々は，適応型空間多分割による分割統治法の k 最近傍ペア探索手法，Adaptive Multi Partitioning (AMP) を提案する．AMP は，Pivot からオブジェクトまでの距離が分散している空間から順に分割・統治のステップで k 最近傍ペアを探索する．距離に対するオブジェクトの分散は，距離の分布の歪度をもとに判断する．本手法は，距離に対するオブジェクトの分布が密な空間のほうが，収束した上限値による枝刈りの効果が大きいことを利用している．閾値よりも距離の近いペアを探索する分割統治法のアルゴリズムは従来から研究されていたが [2]，これを k 最近傍ペア問い合わせに適応させて枝刈りの効果を向上させたのは，我々の知る限り本研究が初めてである．

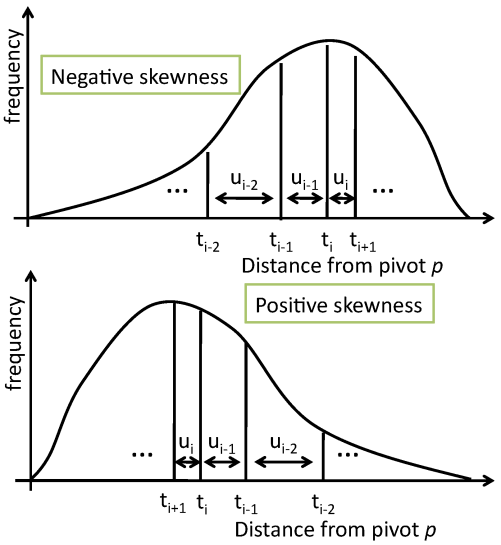


図 1: Adaptive Multi Partitioning

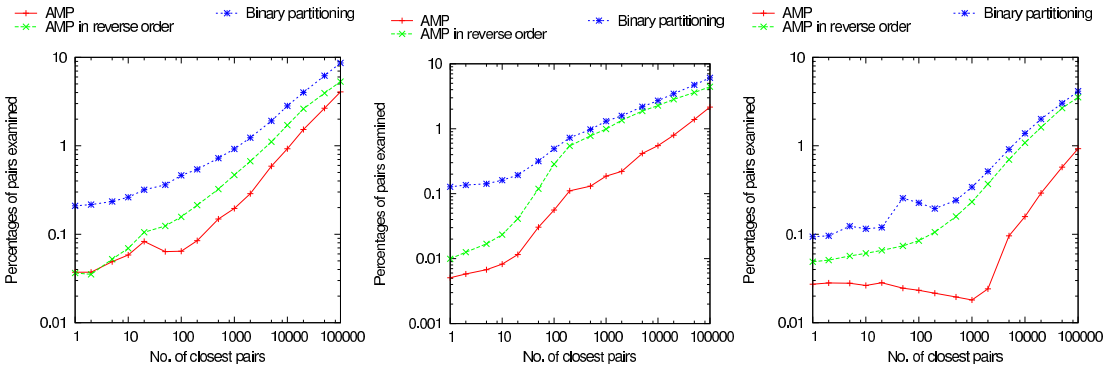
本稿では，AMP の分割方法について説明した後，評価実験の結果を紹介する．3 つの実データを使って探索時に発生する距離計算コストを比較し，提案手法の有効性を示した．

2 Adaptive Multi Partitioning

Adaptive Multi Partitioning (AMP) は，適応型空間多分割による分割統治法の k 最近傍ペア探索手法である．k 番目の類似ペア間の距離の上限値を更新しながら再帰的に空間の多分割を繰り返し，上限値よりも距離が遠いと判断できるオブジェクトのペアを枝刈りする．分割の順序は Pivot からオブジェクトまでの歪度をもとに決める．図 1 に AMP の概要図を表す．

2 つのオブジェクト集合 X と $Y(|X| \leq |Y|)$ から k 個の最近傍ペアを探索する問い合わせ，AMP(X,Y) ($X = Y$ のときは AMP(X)) を想定する．暫定的な最近傍ペア集合を A とし，A は上限 k 個のペアをヒープ構造で管理する．また，u を，$|A| < k$ のときは ∞，$|A| = k$ のときは A における k 番目の類似ペア間の距離とする．オブジェクト集合 S は，管理する各オブジェクト o について初期値を nil とした参照距離 $\mathrm{Ref}_{o,S}$ を持つ．以下ではオブジェクト集合 S の部分集

Fast k-Closest Pair Algorithm in Metric Spaces
† Hisashi Kurasawa(kurasawa@nii.ac.jp)
†† Atsuhiro Takasu(takasu@nii.ac.jp)
†† Jun Adachi(adachi@nii.ac.jp)
The University of Tokyo (†)
National Institute of Informatics (††)
2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan



(a) Nasa (b) Corel Image Datasets (c) Color Histogram

図 2: Computational Cost

合を $S_i = \{o | o \in S \wedge t_i \leq d(o,p) < t_{i+1}\}$ と定義する．

まず，オブジェクト集合 X と Y それぞれから，$\mathrm{Ref}_{o,S} \neq nil \wedge \mathrm{Ref}_{o,S} > u$ $(S = X, Y)$ を満たすオブジェクトをすべて除く．

次に，分割統治の事前計算を行う．まず，X からランダムにオブジェクトを 1 つ選び，Pivot p とし，p を X から除く．そして，p から X と Y に含まれる各オブジェクトまでの距離を計算する．p から各オブジェクトまでの距離の最小値，最大値，平均値，歪度をそれぞれ $d_{\min}$，$d_{\max}$，d_{mean}，s とする．歪度は 3 次のモーメントである．$d(o_i, p) < u$ $(o_i \in Y)$（$X = Y$ のときは $d_{o_i,p} < u$ $(o_i \in X)$）を満たすオブジェクトのペア (o_i, p) を見つけたら，A と u を更新する．

そして，分割･統治ステップをはじめる．i 番目の分割距離を t_i とする．$s < 0$ の Negative skewness のとき，初期値 t_0 は $d_{\min}$ とし，$t_{i+1} = t_i + u$ とする．ただし，t_1 に限り $t_1 > d_{\mathrm{mean}}$ のときは $t_1 = d_{\mathrm{mean}}$ とする．分割ステップとして，X_i と Y_i に対して，$\mathrm{AMP}(X_i, Y_i)$（$X = Y$ のときは $\mathrm{AMP}(X_i)$）を実行する．統治ステップでは，$X'_{i-1} = \{o | o \in X_{i-1} \wedge t_i - d(p,o) < u\}$，$X'_i = \{o | o \in X_i \wedge d(p,o) - t_i < u\}$ の新たな部分集合をつくる．オブジェクトの参照距離は，$\forall o \in X'_{i-1}, \mathrm{Ref}_{o,X'_{i-1}} = \min\{\mathrm{Ref}_{o,X}, t_i - d(p,o)\}$，$\forall o \in X'_i, \mathrm{Ref}_{o,X'_i} = \min\{\mathrm{Ref}_{o,X}, d(p,o) - t_i\}$ とする．同様に，Y'_{i-1} と Y'_i を作る．そして，$X \neq Y$ のときは $\mathrm{AMP}(X'_{i-1}, Y'_i)$ と $\mathrm{AMP}(X'_i, Y'_{i-1})$ を，$X = Y$ のときは $\mathrm{AMP}(X'_{i-1}, X'_i)$ を実行する．

$s \geq 0$ の Positive skewness のときは，$t_0 = d_{\max}$，$t_{i+1} = t_i - u$，t_1 に限り $t_1 < d_{\mathrm{mean}}$ のときは $t_1 = d_{\mathrm{mean}}$ として，p から距離の離れたオブジェクトから順に分割・統治のステップを実行する．

3 評価実験

実験には，Nasa（20 次元のヒストグラムデータ，40,150 件），Corel Image Datasets（32 次元の画像ヒストグラムデータ，68,040 件），そして Color Histogram（112 次元のヒストグラムデータ，112,544 件）の 3 つのデータセットを使用した．Corel Image Datasets は UCI KDD Archive で，それ以外は Metric Space Library で提供されている．いずれもユークリッド距離を類似度とした．実験では，AMP（提案手法），AMP in reverse order（歪度の値をもとにした多分割を AMP と逆順にした手法），Binary partitioning（2 分割の手法），そして Nested Loops（Naive な手法）の 4 つの手法を実装した．1 から 100,000 の最近傍ペアを探索する問い合わせを 500 回実行し，距離計算回数の平均値を求めた．

Nested Loops の実験結果を 100%としたときの結果を出力したものが図 2 である．提案手法は最も距離計算を削減できている．また，2 分割手法と比べると，多分割のほうが性能が良い．これは，分割のステップのたびにオブジェクト集合のサイズだけ Pivot との距離の計算が発生するが，多分割にすることで分割ステップの回数を低減できたことによるものだと思われる．

4 まとめ

我々は，適応型空間多分割による分割統治法の k 最近傍ペア探索手法，AMP を提案した．3 つの実データを使った実験から，距離計算コストを効果的に削減できることを確認した．今後は人工データを使って詳細に評価したい．

参考文献

[1] R.Paredes, et al., Solving similarity joins and range queries in metric spaces with the list of twin clusters. J. of Discrete Algorithms, 2009.

[2] E.H.Jacox, et al., Metric space similarity joins, ACM Trans. Database Syst., 2008.