D-022

クラスタリングを利用した距離尺度の組み合わせによる Top-k 検索

Top-k Query Processing with Clustering for Combinatorial Objects Using Euclidean Distance

鈴木 貴敦 † 高須 淳宏 ‡ 安達 淳 ‡
Takanobu Suzuki Atsuhiro Takasu Jun Adachi

1 はじめに

距離尺度を用いた Top-k 検索問題の中で，オブジェクトの組み合わせを対象とした問題を取り上げる．具体的には，データセット U とクエリ Q，距離尺度 d が与えられたときに，U 中のオブジェクト v_i を重複なしで組み合わせたもの g_l に対して，クエリとの距離が最も小さくなる上位 k 件の組み合わせを求める検索タスクである．

通例の検索では，データベース中の各オブジェクトに対して，クエリとの類似度で順位付けを行い，結果として出力する．それに対して，我々が提案する検索は，オブジェクトの組み合わせに対して，クエリとの類似度で順位付けを行い，結果として出力する．これは，オブジェクトの組み合わせのほうが，ユーザの要求に適確に答えられる場合があるためである．この検索の応用は，例えば，日々の生活の履歴をクエリとして，1 日にどのようなものをどれくらい食べて，どのような運動をどれほど行えばよいかを提示する，健康的な生活を送るための支援システムなど，複数の要素を組み合わせることでよりよい結果が求められるものを想定している．

組み合わせによる検索を実装する最も単純な方法は，ループを入れ子にして全ての組み合わせを調べることである [1]．全ての組み合わせに対してクエリとの類似度を計算する場合，オブジェクト数を N，1 つの組み合わせにおけるオブジェクト数の上限を n とすると，$\sum_{i=1}^{n} {}_N C_i \sim O(N^n)$ のコストがかかってしまい，現実的とは言いがたい．そこで，我々はクラスタリングを利用した組み合わせ Top-k 検索の高速化手法を提案した [5]．ここでは，あらかじめ k-means でクラスタリングを行い，できたクラスタの組み合わせに対して類似度の計算を行うことで，解候補の絞り込みを行う．しかし，クラスタの組み合わせからオブジェクトの組み合わせを求める段階において，解になるかどうかの判定をクラスタ重心からの距離のみで行っていたため，効率的な枝刈りができていない問題があった．本稿では，事前処理でクラスタリングに加えて主成分分析を行い，第 1 主成分上へオブジェクトの射影をとることで，解候補の判定の効率化を行った．

これまでの Top-k 検索の研究では，評価関数の単調性を利用し，評価値が降順になるようにあらかじめソートしておくことで高速化するものが多い [2]．しかし，今回のように，類似度評価を距離で行う場合では，ソートによる高速化が難しいため，上位 k 件に入る可能性のあるオブジェクトを効率よく見つけ出すことに主眼が置かれている [3]．組み合わせを対象とした Top-k 研究では Skyline 検索をベースとしたものが提案されている [4]．この研究では，Skyline となるオブジェクトに対して各次元の値の大小で順位付けを行う．本稿の扱うタスクは，評価関数が距離であり，オブジェクト全体を対象としている点で異なっている．

† 東京大学大学院 情報理工学系研究科

‡ 国立情報学研究所

2 問題定義

データセット U と，クエリ Q，クエリとオブジェクトの類似度を定量化するための距離尺度 d が与えられたとする．n を非負整数としたときに，以下の 2 つの条件をみたすようなオブジェクト $v \in U$ の組み合わせ $g \subset U$(ただし，g の要素数は n 以下) の列 $G = (g_{i_1}, g_{i_2}, \ldots, g_{i_k})$ を求めることが，本稿で扱う検索タスクである．

1. 列 G は k 件のオブジェクトの組み合わせからなり，全ての $g_j (g_j \notin G)$ について $d(Q, g_{i_k}) \leq d(Q, g_j)$ をみたす
2. 列 G はクエリからの距離で整列されている．すなわち，$\forall 1 \leq j < k : d(Q, g_{i_j}) \leq d(Q, g_{i_{j+1}})$

オブジェクトとクエリを多次元ベクトルで表現し，オブジェクトの組み合わせをベクトル同士の加法，距離尺度としてユークリッド距離を利用する．オブジェクト，クエリともにベクトルの要素は非負とした．

3 提案手法

提案手法の概要を以下にまとめる．

1. 事前処理
   (a) 主成分分析を行い，最も寄与の大きい第 1 主成分を求める．
   (b) k-means によるクラスタリングを行い，クラスタ重心とクラスタ半径を求める．クラスタ重心と，クラスタ重心から各オブジェクトへの差分ベクトルに関して，第 1 主成分上への射影を求める (ただし，k-means における k は，解の数 k と無関係である)．
2. 探索処理
   (a) (1-b) で求めたクラスタに関して，クラスタ重心の組み合わせを調べる．上位 k 組のオブジェクトの組み合わせを含む，クラスタの組み合わせを保持する．
   (b) (2-a) で求めたクラスタの組み合わせを調べ，上位 k 組のオブジェクトの組み合わせを決定する．この際，まずは第 1 主成分上において解になり得るか否かを判定した後，距離計算を行う．

事前処理 本手法では，クエリが与えられる前に，データセットに対して主成分分析とクラスタリングを行う．始めに主成分分析を行い，最も寄与の大きい第 1 主成分を求める．次に，k-means を用いてクラスタリングを行い，クラスタの重心およびクラスタ半径を求める．そして，第 1 主成分上にオブジェクト及びクラスタ重心の射影を求める．

探索処理 まず始めに，上位 k 組のオブジェクトの組み合わせを含むクラスタの組み合わせを求める．クラスタ同士の組み合わせには，重心ベクトルの和を利用する．クラスタの組み合わせの半径は，選んだクラスタの半径の総和とする．上位 k 組のオブジェクトの組み合わせを含むクラスタは，以下の 2 つの条件を満たす．ただし，k' とは，クラスタの組み合わせからできるオブジェクトの組み合わせが，k 個以上となるクラスタの組

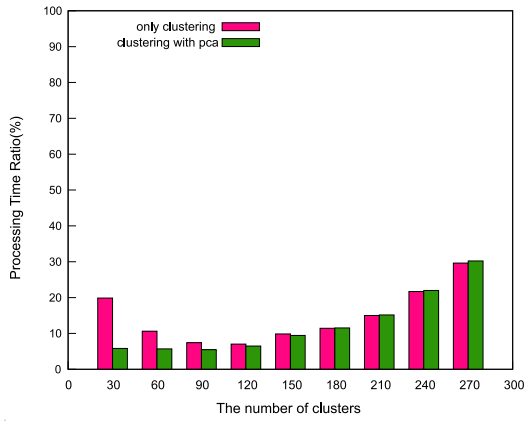

図 1　クラスタ数を変化させた場合

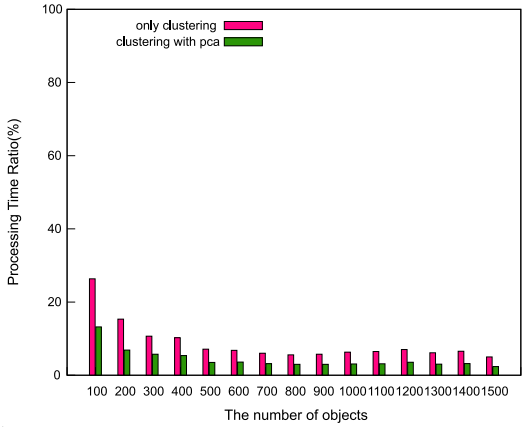

図 2　オブジェクト数を変化させた場合

み合わせの数である．

1. クラスタ重心の和とクエリとの距離が最も近い上位 k' 組のクラスタの組み合わせ (グループ 1 とする)
2. クラスタの組み合わせ C の半径の総和を r_C，クエリとの距離を d_C，とすると，上位 k' 番目のクラスタの組み合わせを $C_{k'}$ とした場合，以下の式 1 を満たすクラスタの組み合わせ $\hat{C}$ (グループ 2 とする)

$$d_{C_{k'}} + r_{k'} > d(\hat{C}, Q) - r_{\hat{C}} \tag{1}$$

得られたクラスタの組み合わせから，オブジェクトの組み合わせを求める．まず，グループ 1 内のクラスタの組み合わせとクエリとの距離が小さいものから順に，オブジェクトの組み合わせを k 組求める．k 組求めたら，残ったクラスタの組み合わせ $\hat{C}$ について，$max\{d(\hat{C}, Q) - r_{\hat{C}}, 0\}$ が小さいものから順に調べる．このとき，全ての可能な組み合わせに対して距離を計算するのではなく，第 1 主成分上への射影を利用し，以下の式 2 を満たすオブジェクトの組み合わせのみ，解になる可能性のあるものとして実際に距離計算を行う．ただし，現在注目しているクラスタを $\hat{C}$，差分ベクトル $x \in \hat{C}$ の第 1 主成分への射影を $proj(x)$，d'_k を k 番目の距離とする．

$$proj(Q) - d'_k > proj(\hat{C}) + \sum proj(x) \tag{2}$$

$max\{d(\hat{C}, Q) - r_{\hat{C}}, 0\}$ が，k 番目の距離よりも大きくなったら探索を終了する．

4　評価実験

実験には，UCI Machine Learning Repository[6] より提供されている，あわびの個体の数値データ (Sex，Rings を除いた 7 次元，4177 個) を利用した．クラスタリングのみを用いた手法，クラスタリングに主成分分析を加えた手法について，3 個以下のオブジェクトの組み合わせのなかから，上位 1 件を求める場合について，全解探索のコストを 100% としたときの計算コストで評価を行った．なお，クエリはランダムに選んだ 100 個オブジェクトの平均をとり，そこで得られたベクトルを 3 倍したものを利用した．

図 1 に，オブジェクト数を 300 個に固定し，クラスタ数を 0 から 270 まで 30 刻みで変化させた場合の結果を示す．横軸がクラスタ数を表し，縦軸が全解探索との比率を表す．各クラスタ数につき 50 回探索を行い，計算コスト比率の平均をとって結果とした．クラスタリングと主成分分析両方を用いた場合，最大で 95% 計算コストが削減できることがわかった．クラスタリングのみを用いた場合，適切なクラスタ数を選ばなければ，コストが最大で 20% 程度変動してしまうため，チューニングの面で問題がある．しかし，主成分分析を利用すれば，クラスタ数が適正値よりも少ない場合であれば，計算コストの変動を抑えることが可能であり，その上計算コストを減らせる利点がある．

図 2 に，クラスタ数をオブジェクトの数に対して 20% に固定した場合に，オブジェクト数を変化させた場合の計算コストの変化を示す．オブジェクト数が少ない場合では，コストが比較的大きくなっているが，オブジェクト数が増加しても安定して 90% 以上の高速化を実現した．

この手法の問題点は，データ数や，1 つの組み合わせにおけるオブジェクト数の増加に対応できない点である．現在の決定的手法では計算量のオーダを変えることは難しいと考えられるため，今後はよりスケーラブルな近似解法を検討する．

5　おわりに

本稿では距離尺度を用いた組み合わせに関する Top-k 検索を，クラスタリングと主成分分析を用いて高速化した．提案手法では，k-means によるクラスタリングを行い，クラスタの組み合わせを調べることで上位 k 件の組み合わせ候補を絞り込み，そこからオブジェクトの組み合わせを求める際に，オブジェクトとクラスタの第 1 主成分上への射影を利用した枝刈りを行い，高速化を達成した．現在の決定的手法では，スケーラビリティの面で問題があるため，今後はより大規模なデータを扱うことができる近似解法を模索していく．

参考文献

[1] P.Zezula, et al.,Simirality Search The Metric Space Approach, Springer,2006
[2] R.Fagin, et al., Optimal Aggregation Algorithms for Middleware, PODS, 2001.
[3] N.Augsten, et al., TASM: Top-k Approximate Subtree Matching, ICDE, 2010.
[4] I-fang Su, et al.,Top-k Combinatorial Skyline Queries, DASFAA, 2010.
[5] 鈴木他, 距離尺度の組み合わせによる Top-k 検索の提案, 第 73 回情報処理学会全国大会, 2011.
[6] UCI Machine Learning Repository, http://archive.ics.uci.edu/ml/