4N-7

距離尺度の組み合わせによるTop-k検索の提案

鈴木 貴敦 † 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

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．例えば，栄養バランスを考慮した食材の検索のような，1つのオブジェクトだけではユーザの要求を満たすことは難しいが，複数のオブジェクトの組み合わせを考えるとユーザの意図をよりよく反映した結果が得られる検索へ応用できる．本研究の目的は，このような組み合わせを求める検索での問い合わせ処理の高速化である．

組み合わせによる検索を実現する最も単純な方法は，ループを入れ子にすることである[1]．しかし，全ての組み合わせに対してクエリとの距離を計算する場合，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うため現実的ではない．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する．本手法では，あらかじめk-means法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距離計算をすることで上位k件の組み合わせになり得るオブジェクトを限定する．この手法は，オブジェクト数が$1/x$になったときに，計算コストが$1/x^n$倍になることを利用した．

組み合わせによる検索の先行研究としては，Skyline検索で求めたオブジェクトに対して，与えられた評価関数の値が最も大きくなるような組み合わせの上位k件を求める検索タスクが挙げられる[2]．本研究は，距離尺度を利用する点，オブジェクト全体に対して組み合わせを考える点の2つで異なっている．

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近さを定量化する距離尺度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$d(Q, g_{i_k}) \leq d(Q, g_j)$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，$\forall 1 \leq j < k : d(Q, g_{i_j}) \leq d(Q, g_{i_{j+1}})$

さらに，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．また，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. k-means法によるクラスタリングを行う
2. 1.で求めたクラスタに関して，組み合わせに対して探索を行い，上位k件の組み合わせに含まれうるオブジェクトを要素に持つクラスタの候補を求める
3. 2.で求めた候補について，上位から順に再度探索を行い，上位k件のオブジェクトの組み合わせを決定する

本手法では，まず始めにk-means法を用いてクラスタリングを行う．そこで得られたクラスタの組み合わせに対して距離計算を行い，上位k件の組み合わせの要素になり得るオブジェクトを絞り込む．距離計算には各クラスタの重心を用いる．クラスタ半径は，重心から最も離れたベクトルと重心との距離とした．

クラスタの組み合わせg'から生成されるオブジェクトの組み合わせgとクエリとの距離は，三角不等式の関係から，

$$max\{d(c, Q) - R, 0\} \leq d(g, Q) \leq d(c_k, Q) + R_k \quad (1)$$

A Top-k Query Processing for Combinatorial Objects on Metric Spaces

† Takanobu SUZUKI(suzuki@nii.ac.jp)
†† Atsuhiro TAKASU(takasu@nii.ac.jp)
†† Jun Adachi(adachi@nii.ac.jp)
Graduate School of Information Scienece and Technology, The University of Tokyo (†)
National Institute of Informatics (††)
2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan

を満たす．ここで，c は g' のクラスタの重心の総和，R はクラスタの半径の総和を表す．また，c_k，R_k はそれぞれ上位 k 番目のクラスタの重心の総和と，クラスタの半径の総和を表す．我々はこの不等式を利用して枝刈りを行った．

オブジェクトやクラスタの組み合わせの探索は，深さ優先探索をもとにして行う．オブジェクト，クエリともにベクトルの要素が非負であるため，距離の変化は下に凸の曲線となる．そのため，一度距離が増加傾向に転じたら，それ以降の組み合わせは探索しない．

クラスタの組み合わせに対して上位候補が求められたら，上位から順にクラスタを構成している要素について再度探索を行い，上位 k 件の組み合わせを決定する．

4 評価実験

実験には，UCI Machine Learning Repository[3] 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，データ数は 500 件) を利用した．提案手法の比較対象として，クラスタリングせずに深さ優先探索のみを行うものを実装し，それぞれの手法について上位 1 件を求める場合について，距離計算の回数を求めて比較した．なお，クエリはランダムに選んだオブジェクトの平均をとり，そこで得られたベクトルを m 倍したものを利用した．m が大きい場合，オブジェクトを組み合わせてもクエリ近辺に到達しないことになる．

図 1 に組み合わせの最大要素数を 3 個として，クエリの m を変化させたときの結果を示す．各 1 から 20 まで各乗数につき 20 回探索を行い，そのときの距離計算の回数の平均を結果とした．横軸はクエリの m，縦軸はループを入れ子にしたものを 100%としたときの距離計算の回数の比率を表す．なお，本実験では，クラスタ数を 172 個に固定した．図 2 にオブジェクト数を 500 個に固定し，組み合わせの最大要素数を 3 個として，クラスタ数を変化させた場合の結果を示す．クエリは $m = 100$ のときのものを使用した．

図 1 より，提案手法は，クエリとオブジェクトの組み合わせとの距離が大きくなった場合に有効であることがわかった．また，図 2 より，クラスタ数の変化に伴って処理コストが大きく変化していることがわかる．実験ではクラスタ数を人手で設定した．適切なクラスタ数を選ばなければ処理コストが大きくなってしまうが，状況に合わせた調整は難しいと考えられる．そのため，今後はクラスタ数を調整しなくとも処理コストを小さく抑えることができるクラスタ手法を検討する必要がある．

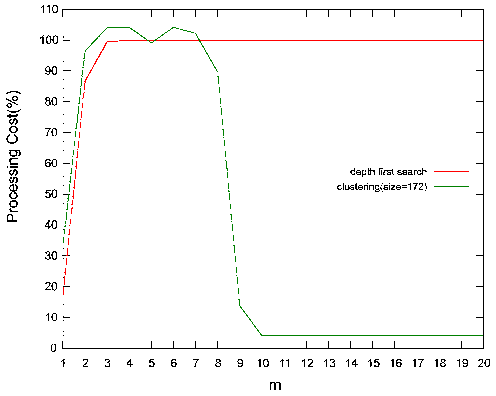

図 1: クエリを変化させた場合

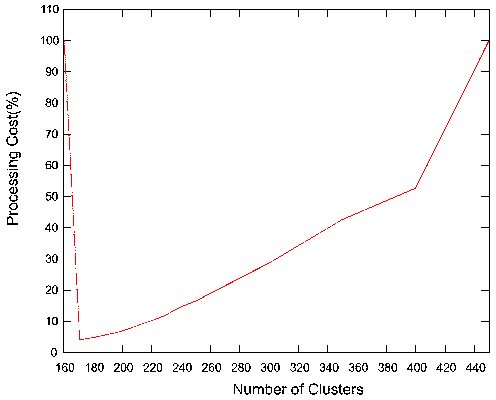

図 2: クラスタ数を変化させた場合

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 検索を提案した．提案手法では，k-means 法でクラスタリングし，クラスタの組み合わせを調べて上位 k 件の組み合わせ候補をあらかじめ絞り込むことで高速化を行った．オブジェクトの組み合わせとクエリとの距離が大きい場合に有効であるがクラスタ数を適切に設定しなければ効果は小さくなってしまうことがわかった．今後は，クエリとオブジェクトの組み合わせとの距離に非依存で，かつ状況に合わせて自動的にパラメータを調整可能なクラスタ手法を検討する．

参考文献

[1] P.Zezula, et al.,Simirality Search The Metric Space Approach, Springer,2006

[2] I-fang Su, et al.,Top-k Combinatorial Skyline Queries, Database Systems for Advanced Applications, 2010.

[3] UCI Machine Learning Repository, http://archive.ics.uci.edu/ml/