ISS-A-033

代表点再選択を用いた mCK 検索の多様性と正確さの向上

大石貴也 †　　津野貴大 ‡　　大森匡 ‡

† 電気通信大学情報理工学域 I 類　　‡ 電気通信大学大学院情報理工学研究科

1. 背景と目的

最近の Web データは位置情報とキーワードつき写真のように，個々のデータ (オブジェクト) が発生場所の緯度経度情報と内容を表すタグ複数を持つ．こうしたデータ集合からの領域発見問題に，m-最近接キーワード検索 (mCK 検索) がある．mCK 検索とは，問い合わせ Q としてキーワード m 個を与えたとき，Q を満たす高々m 個のオブジェクト集合 O のうちその要素間の相互近接度を表す「直径」$diam(O)$ が最小となる O_{opt} を求める問題である [1](図 1). ($diam(O) = \max_{\forall o_i, o_j \in O} dist(o_i, o_j)$)

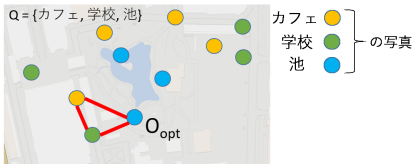

図 1: mCK 検索 (Q={ カフェ，学校，池 }) の検索結果例

mCK 問題で直径の小さい順に k 個の答えを列挙すると，地図上のごく少数の領域に答えが集まりやすい．著者らは，入力データ側で Drosou の代表点選択方法 [2] を工夫してこの解決を試みてきた [3]．本稿では再サンプルを使って列挙する解の多様性と正確性を向上させる手法を報告する．

2. 再サンプルを利用した mCK 検索

西野ら [3] は DiSC-diversity の手法 [2] を mCK 検索の各キーワード Ki を満たす点集合 Di に適用して代表点集合を作り，上位 k 個の mCK 解検索の解集中を改善した．つまり，Di を被覆する半径 r の円の集合を作り，各円の中心点で Di を代表して，各キーワードの代表点集合を併合したものを入力データとして mCK 検索を行った．この方法は，半径 r が大きいと多様性（異なる 2 解の間の平均距離）が上がるが，代表点から作った答え（代表解）は直径に最大+-2r の誤差を持つ．

そこで，本稿では，$r = 1$km による初回列挙の上位 100 解に対して，それらの解を構成している各キーワードの点が集約した元データ点を対象に，$r' = 50$m による代表点を再取得することにした．そして，この再取得した代表点を用いて，初回列挙解についての詳細な結果となる解を求める．これにより，解の散らばりを $r = 1$km 対応のまま損なわずに r'=50m に応じた正確な解の列挙が可能になるはずである．ただし，2 回目列挙のとき解 a を見つけると，「初回列挙の上位 $k1$ 位までの解のうちある A_i が存在し，その A_i について，a のキーワード別頂点が A_i の対応するキーワードの頂点の集約範囲に入る場合にのみ，a を再列挙解として認める」とし，この制約下で再列挙で上位 100 解までを数えた．また，初回列挙解 1 つあたり最大 μ 個までの 2 回目列挙解を許すとした．

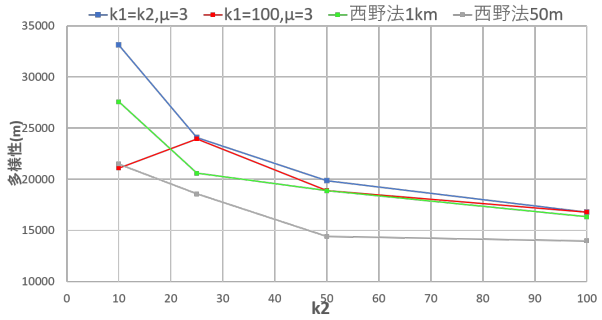

図 2: 上位 K 解の持つ多様性尺度 (2 解間の平均距離 (m))

3. 実験結果と議論

Flickr 写真データ約 20 万件を使った上位 100 解の検索結果における西野の手法 (r=1km と 50m の 2 種類) の場合と，今回の提案手法 (初回列挙を r = 1km, 再列挙時が r'=50m) の場合について，それぞれの上位 k 解までの集合が持つ多様性尺度（相異なる 2 解の距離の平均値）を，図 2 に示す．ここで，提案手法は，2. の記述に沿って初回列挙の上位 $k1$ 解を使って再列挙で上位 $k2$ 解までを求めるとして，$k1 = 100$ で $\mu = 3$ を標準的な設定とし，$k1 = k2$(= 再列挙で上位 k 個まで) も試した．同図から，提案手法が，r'=50m の場合の正確さのまま多様性では r = 1km のときの値とほぼ同等にできたとわかる．一方，μ のために上位 20 位以内の正解を拾わない現象も見られた．以上から，r=1km に応じた高い多様性を保ったまま初回列挙の上位解のうち何個までを使って再サンプルするのか，そのときパラメタ μ の設定による解の正確さへの影響をどう制御するか，が課題である．

参考文献

1. T.Guo,et.al., ”Efficient Algorithms for Answering the m-closest Keywords Query，” ACM SIGMOD, pp.405-418, 2015.

2. M.Drosou, et.al., ”DisC Diversity: Result Diversification based on Dissimilarity and Coverage，” VLDB 2013，pp.13-24，2013.

3. 西野，津野，大森，他, ” 空間 Web データ上の m-最近接キーワード検索における代表解の探索，” 情処全大, 4L-06, 2021.