2M-04

mCK 検索における独立性と被覆性のある代表解の列挙 *

塙 雪耶 †　大森 匡 †　藤田 秀之 †　新谷 隆彦 †

電気通信大学大学院　情報理工学研究科　情報・ネットワーク工学専攻 ‡

1　背景と目的

Flickr や Twitter といった緯度経度情報つきの Web データを用いた情報抽出は最近のデータベース研究の話題の一つであり，その一つに m-最近接キーワード検索 (mCK 検索) がある．これは，Flickr のように，個々の写真データ (オブジェクト) が場所と写真内容に関するタグ複数を持っている状況で，問い合わせとしてキーワード m 個の入力 Q 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，その要素の相互近接度 (直径の小ささ) が最も高い集合 O_{opt} を求める問題である [1].

mCK 検索では，オブジェクト組を答えとするため，直径が小さい順に k 個の答えを列挙すると，地図上の数か所に答えが集まり解の多様性が低下する．その解決方法として，r-DisC Diversity[3] の考えに基づいた独立性・被覆性のある代表解を列挙する方法を提案する．

2　mCK 検索問題と独立性・被覆性

mCK 検索問題の定義: 本稿では地図上のテキスト情報と位置情報を持つ空間 Web データをオブジェクトと呼び，各オブジェクトは地図上の 1 点を表すと仮定する．このとき mCK 検索とは，次のように定義される．ユーザーが m 個のキーワード Q を入力したとき「オブジェクト o が持つテキスト情報は Q の少なくとも 1 つ以上のキーワードに該当する」という制約下で高々 m 個のオブジェクトの集合 $O = \{o_{i1}, o_{i2}, \ldots, o_{il}\}$ を対象としたとき，Q の各キーワードが少なくとも 1 つの O のオブジェクトによって満たされており，かつ，そのような O のうち，直径 $diam(O)$ が最小となるものを最適解 O_{opt} として返す検索問題である [1].（ここで，O の直径 $diam(O)$ とは O の 2 オブジェクトの距離 $dist$ の最大値．即ち $diam(O) = \max_{\forall o_i, o_j \in O} dist(o_i, o_j)$.）

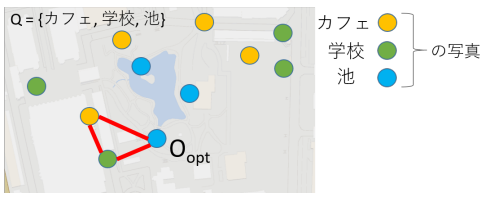

図 1: mCK 検索の例

mCK 検索の例を示す．図 1 では，地図上にカフェ，学校，池というテキスト情報を持った 3 種類の写真が存在する．図 1 の例で Q =｛カフェ，学校，池｝で mCK 検索を行うと，3 種類の写真からなるオブジェクト集合で直径が最も小さい集合を探す．すると，赤い線で結ばれたオブジェクト集合が最も直径が小さいのでこれを O_{opt} として返す．O_{opt} が Q に応じた最適な領域となる．

従来手法: mCK 検索は m について NP 困難であり [1]，本稿では PE 法 [2] を用いる．PE 法では，データベース上で Q のキーワードを 1 つでも満たすようなオブジェクト全てを一つの四分木に格納する．四分木の各ノードには，そのノードに属する点集合の MBR(最小包囲矩形) を持たせている．その後最近接二点探索の戦略を使って mCK 解の直径の候補となる 2 つのオブジェクトペア (o_1, o_2) を葉ノード組列挙を介して直径の小さい順に列挙し，各 (o_1, o_2) が直径となりうる mCK 解を構成するかを順に検査する．

Top-k mCK 検索における問題の指摘: mCK 検索はオブジェクトの組み合わせを解とするため，直径が小さい順に上位 k 解を求める Top-k mCK 検索を行ったとき，上位の解付近にオブジェクトが複数あると，それらのオブジェクトを使って別の解を作るので，解が重なって出現する．そのため，上位 k 個の解が地図上の数か所に局所的に集中して出現するという問題が発生する．

r-DisC Diversity の独立性と被覆性: 多様性の高い情報検索技法の研究として Drosou の r-DisC Diversity[3] がある．これは，二次元上の点集合について，円の中心同士の距離が r 以上の条件で半径 r の円の集合によって，点集合全体をできるだけ少ない数の円で被覆して，各円内の点をその円の中心点によって支配 (代表) させるというものである．(mCK 検索では円の中心点が代表解となる.) このことは独立性と被覆性の 2 つの性質によって定義され, 代表点集合に独立性があるとは, 代表点集合の任意の要素 o_i, o_j に対して, $dist(o_i, o_j) > r$ が成立していることであり , 被覆性があるとは, 元の点集合の要素は必ず代表点集合の少なくとも 1 つの点によって代表されているということである．

3　独立性・被覆性のある列挙の方法

本稿では，mCK 解 X の中心点を，X の直径となる 2 端点の中点と定義して，mCK 解列挙に DisC Diversity を適用する．つまり，ある解 Z の中心点から半径 r で描いた円を C_Z とおき，円 C_Z の中に別の解 W の中心点が入るときに Z dominates W と決めて，Z を代表解に選んだら W を skip すれば良い．この考えに沿い，独立性・被覆性のある代表解の列挙方法として以下の方法を提案する．

3c 方式: Y dominate X を「解 Y の方が別の解 X より順位の強い解 (Y の直径が X のそれ以下) でありかつ，Y の円 C_Y に X の中心点 pX が含まれる」と定義したとき，

*Finding independent and coverage answers for m-Closest Keywords Query

†Y.Hanawa, T.Ohmori, H.Fujita, T.Shintani

‡The University of Electro-Communications

残す解 X を「 X の追加時点で，X を dominate する Y が現在の Top-k 解候補リスト L1 に存在しない」とした方法が 3c である．3c によって得られた解の集合は dominate の定義に基づいて DisC Diversity を考えたときに，上位 k 解の直径未満の全解を被覆する性質を維持した集合となるが，独立性は保証できない．

3d 方式: Y dominates X を「 (直径の大小に関係なく) 解 Y の円 C_Y に別の解 X の中心点 pX が含まれる」と定義したとき，この基準で PE 法を使って他の代表解に dominate されない X を上位 k 個列挙探索する方法が 3d である．具体的には，葉ノード組ごとに直径の小さい順に解リスト L1 を作り，以降葉ノード組から作られる解候補が L1 に入るか検査する (一度選んだ代表解はそのまま残す.)．3d によって得られた解の集合はこの dominate の定義に基づいて Disc Diversity を考えたときの独立した支配頂点の集合になっているが，探索中に上位 k 解から追い出された解 Y が上位 k 解に入る他の解 Z を dominate している可能性があるため，上位 k 解の直径未満の全解の被覆を保証できない．

3c，3d は独立性と被覆性のどちらか一方の性質しか保証しない．そこで本稿ではそれら両方の性質を保証するような次の方法を提案する．

3d_all 方式: Y dominates X の定義は先の 3d 方式と同じまま，検査の対象を直径 εkm 以下の解集合とし Top-k による排除を行わない．直径が εkm より大きいものは代表解候補としないことでそれらに支配されてしまう直径の小さい解を漏れなく検査することが可能になる．すなわち，3d_all で得られる代表解は dominate の定義から独立性を保証し，直径 εkm 以下の全解に対して被覆性を保証するものとなる．

4　評価とまとめ

実験は 3d_all 方式の他，比較のため多様性の制御を一切かけない場合 (mCK3) と 3c 方式で行い，地図と数値を使って評価する．Q={sakura，river，temple} とし，mCK3，3c は k=100，また 3c，3d_all は円半径 r=0.5km で固定し，3d_all は ε=1km 以下の全解列挙とした．

地図での比較は浅草周辺で行った．赤点が sakura，緑点が river，青点が temple で，青い丸に数字 (解の順位) が書いてある三角形が代表解を表す．最初に，図 2 と図 3，4 を比較する．mCK3 の場合は 3c における解 4 番周辺に図 2 のように解が集中している一方，3c，3d_all では，mCK3 の解が集中している所を数個の解で支配するなど，広い範囲で代表解を形成していることが分かる．次に，図 3 と図 4 を比較する．3c の解 45 番付近は極小解を取るために代表解が直近に現れ独立性が崩れている一方，3d_all の同じ箇所は解 42 番 1 つに要約されているため独立性を維持している．このような箇所は他にも見られ，3c の直径 1km 以下の解を 3d_all の代表解のいずれかが必ず支配していて被覆性を維持している．数値評価で用いるのは異なる 2 解の平均距離 diverstiy(m) と解の直径 diam(km) である．mCK3 と 3c の上位 100 解の diversity はそれぞれ 1501 と 7091 であった．3d_all は代表解が 66 個しか見つからなかったので上位 66 解での diversity を計算するとその値は 7272 であった．3c, 3d_all ともに制御しない場合と比べて多様性が大きく向上している．3d_all は独立性を保証し解の個数が 3c に比べて少ないため僅かに値が大きい．また，3d_all の解 66 番の diam は 0.99969 で 1km 以下に抑えられている．

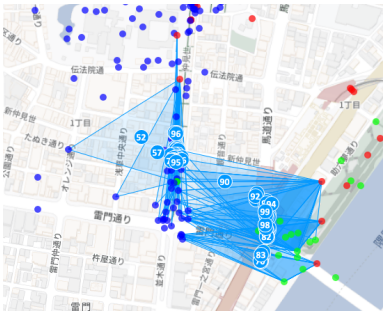

図 2: 多様性制御をかけない場合 (mCK3)

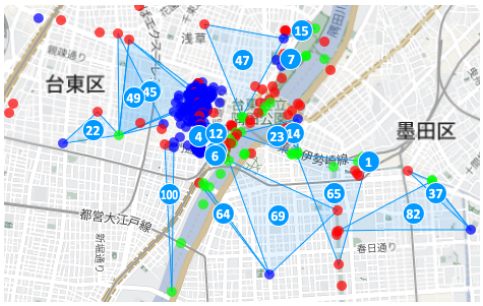

図 3: 3c(r=0.5km)

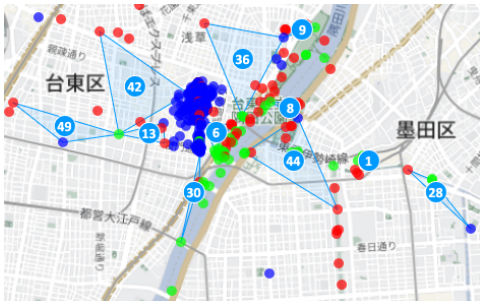

図 4: 3d_all(r=0.5km，ε=1km)

まとめとして，3c では独立性は保証しないが極小解を落とさず被覆性を保証する一方，3d_all では検査する解集合を限定することで 3c の代表解が支配している解を支配し，独立性と被覆性のある代表解を列挙することができた．

謝辞：本研究は科研費 23K11115 の助成による．

参考文献

[1] T.Guo,X.Cuo,G.Cong"Efficient Algorithms for Answering the m-closest Keywords Query，" ACM SIGMOD, pp.405-418, 2015.

[2] Qiu, Hei, Ohmori, Fujita, "An Object-Pair Driven Approach for Top-k mCK Query Problem by Using Hilbert R-tree，" 2019 IEEE BigDataSE pp.655-661, 2019

[3] M.Drosou，E.Pitoura "DisC Diversity: Result Diversification based on Dissimilarity and Coverage，" VLDB 2013，pp.13-24，2013