ISS-P-044

空間データにおける多様性を考慮した m-最近接キーワード検索方式

道根星眞，藤田秀之，新谷隆彦，　大森匡
電気通信大学大学院情報理工学研究科

1. 背景と目的

近年の Web 上のデータは，写真データのように，内容を表すテキスト情報だけでなく，地図上の（撮影）位置も持ち，空間上の点データ（以下，オブジェクト）でもある．これら空間的な Web データを使って「入力した m 個のキーワード群にとって最適な場所」を抽出する研究の 1 つに，m 最近接キーワード検索 (mCK 検索) 問題がある [1]．本稿では，mCK 検索における検索結果の空間的な多様性を向上させる解法を提案する．

2. mCK 検索の従来の問題

mCK 検索は，データベース D（Flickr なら，写真の集合）が与えられたとき，入力キーワード m 個に最も適切な地図上の位置を探す問題である．形式的には，高々 m 個の写真の集合 O のうち，O の全写真で m 個全てのキーワードを満たせるような集合 O を考え，最適な O として，O の全要素が地図上に最も近接して配置されており, かつ, 冗長な要素がないものという $O(\subset D)$ を決定する問題である．

ここで, O の「直径」を, $diam(O) = \max_{o_i,o_j \in O} dist(o_i, o_j)$ で与える．直径が小さければ，O の全オブジェクトは相互により近接していると言える．$diam(O)$ が最小となる O が，Q に最適な解である．

従来手法 [1] では，直径の小さい順に上位 K 解を地図上に表示させることができるが，それらは，ほぼ同一の箇所に集まる．(例：sakura の写真は，その近くに大量に存在するため，ほぼ同じ場所の解が増える)．即ち，mCK 検索の解による地図上の分布全体を知ることはできず，情報検索で言う多様性が極端に低い．

3. 提案手法

そこで本稿では，原論文 [1] で使われた Apriori に基づいた空間分割木のノード組み合わせにおいて, 新たに全解列挙方式に着目した設計を採用して，上記の問題を解決する．

今, キーワード m 個からなる問い合わせ Q が出ると解法の手順は，以下の通りである：

1. Q の少なくとも 1 キーワードにあたるオブジェクトを D からロードし，同一地点のデータを 1 つにした後, オブジェクト群から 1 つの KD 木を作成する．葉ノードには MBR を与えておく．

2．KD 木の個々の葉ノードを 1 アイテムとして，Apriori 法を使って葉ノードのアイテムセットを長さ 1, 2, 3, ..., m の順に作成する．このとき, 長さ k の部分アイテムセット $L[k]$ は，必ず，次の条件を満たす必要がある：

（ルール 1）r 個 ($r \leq m$) のキーワードを含むとされた l 個のアイテムからなるアイテムセット L は, 各アイテムにユニークなキーワードの割り当てができなければならない. ゆえに $l \leq r$ の必要がある．

（ルール 2）アイテムセット $L = \{C_1, C_2, ..., C_l\}$ において，L の中の 2 つのアイテム (C_x, C_y) 間の最小距離 $mindist(C_x, C_y)$ と最小直径 θ の関係が $mindist(C_x, C_y) < \theta$ でなければならない. $(C_x, C_y \in L)$

列挙では, 長さ $k+1$ のアイテムセット $L[k+1]$ は, 上の 2 つを満たす $L[k]$ を用いて生成し, その後, ルール 1,2 を満たす $L[k+1]$ だけを残していく．

3．全アイテムセットを計算したら，Q を満たすものを解とする．

手順 3 では，解となるアイテムセット L に，特定のキーワード w にヒットする全オブジェクト数のうち L が持つ数の比率をカバー率として与え，カバー率の高い順に上位 K 個をとることにした．

実験では，関東上空の緯度経度に入る Flickr 写真共有データ 514918 件の写真で評価を行った．$Q = \{sakura, river, temple\}$ のとき, 原論文のオブジェクトを使った上位 100 解は全て，地図上の 1 箇所（スカイツリー付近）に集中していたのに比べ，提案手法は葉ノードアイテムセットで解が存在する 99 箇所をもれなく列挙でき，多様性尺度（相異なる 2 つの解の間の距離の平均）も, 従来手法では 21m で, 提案手法では 17579m となり, 大幅な向上を得た．

4. まとめ

本稿では, 空間 Web 上の mCK 検索問題を扱い，KD 木で対象データ空間を分割して，その葉ノードの組み合わせ (アイテムセット) から，m キーワード全てを満たし, 直径が閾値以下になるものを全てを求める設計と計算方法を提案し, 検索結果の多様性を向上できることを示した．

参考文献

[1] Zhang, et al.”Keyword search in spatial databases: Towards searching by document,” IEEE ICDE, pp.688-699, 2009.

Copyright © 2018 IEICE