審 査 の 結 果 の 要 旨

氏　名　　倉　沢　　央

本論文は、「A Study of Fast Similarity Search Techniques in Metric Spaces （メトリック空間における類似検索の高速化に関する研究）」と題し、英文 5 章から構成されている。情報検索の基礎的な研究分野として、距離空間での類似検索において検索コストを削減するための手法について論じたものである。検索対象となるオブジェクトの空間における分布の特徴を活用して効率的な索引を構成する手法として二つの新しい方式を提案し、従来手法よりもすぐれた特性を持つことを検証している。

第 1 章は「Introduction（序章）」であり、距離空間にあるオブジェクト群に対する類似検索の問い合わせコスト削減の問題を提示し概観している。従来から多数の類似検索のための索引付け手法が提案されてきたが、いずれの手法も検索対象のオブジェクトの特徴を索引構築に活かせていなかった。ここで新たに、オブジェクトの分布の特徴にもとづいた data distribution-based approach と呼ぶ索引付けの方策を提案している。そしてこの方策に基づく 2 つの新しい Pivot 分割手法として Maximal Metric Margin Partitioning (MMMP)と Pivot Capacity Tree (PCTree)の開発の経緯を述べている。

第 2 章は「Related Work（関連研究）」であり、前半では、距離空間の定義や距離関数の具体例、さらに検索タスクについて説明している。後半では、従来提案されている類似検索索引における枝刈り手法と索引付け手法を紹介してそれらに見られる課題について論じている。ほぼすべての枝刈り手法には Pivot と呼ばれる参照オブジェクトが使われているが、この Pivot の選び方によって性能が大きく左右される。本論文では、従来研究においては Pivot によって形成される分割面について考慮が十分でないために枝刈りに効果的な Pivot を選べていないという、既存手法の問題に注目している。

第 3 章は「MMMP: Margin-based Pivot Selection Scheme」と題し、第一の提案手法 MMMP を説明するものである。MMMP は、クラスタ間のマージンを利用した類似検索索引である。MMMP ではまず、データの分布傾向のうち特にクラスタの境界を抽出する。そして、クラスタ形状にもとづいて Pivot とその分割距離を決める。MMMP の分割面は、隣り合うクラスタの端からの距離を最大にするように置かれる。MMMP は偏った分布のデータに対して効果的な手法である。MMMP における索引作成手法を提示した後、従来手法との間で性能評価を行っている。人工の 2 から 30 次元のベクトルデータと三つの実データに対して、iDistance、D-Index、List of Clusters という三つの先行手法との間で、検索応答時間、距離計算回数、ページアクセス回数の比較を行い、いずれにおいても提案する MMMP が良好な特性を持つことを示した。しかし課題として Pivot 選択コストの大きさが上げられ、クラスタ化していないオブジェクト空間において弱いことも判明した。

第 4 章は「PCTree: Pivot Selection Scheme for Optimizing both Pruning and Balancing」と題して、第二の提案手法 PCTree についてのべている。提案手法 PCTree は、データの分布だけでなく、索引木のバランスも考慮した手法である。PCTree では、Pivot によって分割される部分空間のバランスと、Pivot による検索時の枝刈りの効果の、2 つを考慮して Pivot を選択する。その結果、PCTree はデータの分布に合わせて索引構造を効果的に変化させている。PCTree は、MMMP の索引木が不均衡になりうる欠点を改善した手法と言える。先行する研究 GHT、MVP、List of Clusters および SAT の四つの手法との間で人工データ（2 から 64 次元のベクトルデ

ータ、Euclid 距離）、実データ（画像データで Euclid 距離、英単語で編集距離等、5 種）に対して評価実験を行った。近傍検索に必要な距離計算回数、そして索引木の高さなどを比較し、提案する PCTree は、類似検索索引における枝刈り効果とバランスを考慮した分割手法であることが判明し、先行する手法に対して、様々な分布のデータに対して全般的に有効な索引であることが明らかになった。

第 5 章は「Conclusion （結論）」であり、本研究の全体をまとめている。本研究の貢献は、Pivot 選択の新しい方策 data distribution-based approach を確立したことである。従来研究を概観すると、クラスタ形状や索引木のバランスを Pivot 選択に考慮しているような研究は他に見られず、本研究が初めてのものである。これら 2 つの提案手法の有効性を、既存研究と比較した評価実験を通して明らかにした。

以上これを要するに、本論文は、極めて基礎的な距離空間にあるオブジェクト群に対する類似検索において、問い合わせコストを削減するための一般的な手法を論じたもので、オブジェクトの分布の特徴にもとづいた data distribution-based approach と呼ぶ索引付けの一般的概念を提示し、その具体的な実現方法として 2 つの新しい Pivot 分割手法、すなわち Maximal Metric Margin Partitioning (MMMP)および Pivot Capacity Tree (PCTree)を提案し、複数のデータセットに対して性能評価を行い、提案手法が先行する手法に比して全般的に優れた性能を持つことを実証することにより、情報検索の基礎分野で有用な知見を明らかにしたものとして、電子情報学上貢献するところが少なくない。

よって本論文は博士（情報理工学）の学位請求論文として合格と認められる。