論 文 内 容 の 要 旨

博士論文題目 Studies on Spatial Index Schemes for High-Dimensional Data Sets
(高次元データ集合のための空間索引機構に関する研究)

氏　名　　櫻井 保志

(論文内容の要旨)

大規模なマルチメディアデータベースにおいて、高速な内容検索は優れたヒューマンインタフェースを実現する上で重要である。マルチメディアデータベースシステムにおいては、テキスト情報を用いた検索のみならず、マルチメディアデータから抽出した特徴ベクトルに基づく内容検索に関する支援を必要としている。例えば、画像データベースシステムの内容検索は、画像処理によって抽出した特徴ベクトルのマッチングを行うことにより、問い合わせ画像と類似した特徴を有する一つもしくは複数個の画像を画像データベースから探索する。特にデータ集合が大規模になれば、特徴ベクトルを用いた内容検索のための処理が高負荷となるため、マルチメディアデータ検索を用いる多くのアプリケーションは、様々な次元数による空間探索技術を必要としている。

本研究では、新たな索引機構である部分空間符号化法 (SCM ; Subspace Coding Method) を提案する。SCM の動機付けは、従来手法の中で、最も優れた二つのアクセス手法である SR-tree と VA-File の比較分析に基づいている。SR-tree と VA-File の比較はこれまで報告されていないため、本論文では、人工データと実データを用いた、これら二つの手法に関する比較実験の結果を提示している。結果では、SR-tree が 40 次元までは優れた性能を有することを示している。しかしながら、次元数が上昇するにつれて、ノンリーフノードに関するエントリの容量の大きい点が、枝の数の縮小を招き、SR-tree の性能低下につながることが、分析により明らかとなった。

実験結果の分析に基づき、最小包囲矩形 (MBR ; Minimum Bounding Rectangle) や最小包囲球 (MBS ; Minimum Bounding Sphere) を用いる木索引に適用可能な新たな索引機構である SCM を導入する。SCM の基本的なアイデアは、各々 MBR と MBS を包含し、近似する仮想包囲矩形 (VBR ; Virtual Bounding Rectangle) と仮想包囲擬球 (VBS ; Virtual Bounding quasiSphere) の導入にある。また、部分空間符号は VBR と VBS をコンパクトに表現し、このことによって索引におけるノンリーフノードに関する枝の数の増大を可能にする。VA-File において用いられている絶対的なベクトル位置の近似と異なり、部分空間符号は親の VBR と VBS に基づいて、VBR と VBS の相対的位置を表現するものである。この特徴は、実際のアプリケーションにおいて一般に見られるような、一様に分布していないベクトルに特に有効である。

実験では、ノンリーフノードが VBR と VBS を含む木索引を用いて最近傍探索を実施し、実データを用いた結果では SCM の有効性が明らかとなった。SR-tree に適用した場合、実験において実施した 4 次元から 56 次元までの全次元において、SCM はオリジナルの SR-tree と VA-File より性能が優れている。56 次元に関して、SCM は VA-File と比べ 71.9 %、オリジナルの SR-tree と比べ 74.7 % のページアクセスの低減化を達成した。

氏　名	櫻井　保志

(論文審査結果の要旨)

本論文は、空間データとして表現された大量のマルチメディアデータから、所望のデータを効率良く検索するための索引構造とアルゴリズムに関する研究成果をまとめたものである。

特徴ベクトルを用いたマルチメディアデータの内容検索を効率的に実施する空間探索手法に関しては、次元数、もしくはデータ量が多くなると探索コストが極めて大きくなることが知られている。申請者の提案したアルゴリズムでは、R-tree とその派生手法において用いられている最小包囲領域を符号化することによって、木構造におけるファンアウトを増大させ、探索性能の向上を図っている。

本論文における研究成果は次の 2 点に要約される。

第 1 の成果は、従来手法の比較、分析を通じ、従来手法に関する新たな知見を導出したことである。高次元の空間探索に関して最も優れた手法として VA-File と SR-tree が挙げられるが、これまで両手法の比較は報告されていなかった。本論文において両手法の優劣が明らかにされており、また VA-File では探索空間全体からオブジェクトを近似するために分布が偏ったデータ集合の探索においては性能劣化を招くこと、SR-tree においてはファンアウトの小さいことが性能向上の妨げになっていることを指摘している。これらの知見および指摘は、高次元空間探索において本質的、かつ重要であり、評価できる。

第 2 の成果は、空間探索の分野における新手法の提案と手法の有効性の実証である。申請者が提案した部分空間符号化法は、(1) 最小包囲領域を包含し、近似する仮想包囲領域の概念、(2) 仮想包囲領域を表現する部分空間符号を導入したデータ構造、(3) 仮想包囲領域を用いた探索アルゴリズム、(4) 効率的に、部分空間符号による木構造を構築し、構築した木構造を維持するための更新アルゴリズム、によって構成される。(1) によって、仮想包囲矩形は最小包囲矩形を、仮想包囲擬球は最小包囲球を近似しており、領域の位置表現コストを低減化させている。(2) に関して、領域の近似によって索引構造を構築することにより、構造中のノンリーフノードにおけるファンアウトを増大させ、また、親ノードの領域に基づいて子ノードの領域を相対的に表現することにより、近似誤差の低減化に成功している。(3) では、探索処理において、領域の近似誤差が小さく、またファンアウトの大きい構造を用いることによって探索性能の向上を達成している。実際に、提案手法が従来手法を上回る性能を有することを評価実験によって明らかにしており、評価できる。(4) では、部分空間符号に基づくデータ構造を効率的に構築するためのアルゴリズムを提供している。これら本論文において提示されている新たな知見と手法は、空間探索に関する重要な成果である。

以上のように、本論文はデータ工学の分野において高い貢献があり、博士 (工学) の学位論文に値すると認める。