デジタル機器のデータ検索能力を向上させる多次元インデックス技術

より多様なデータを高速，かつ簡単に検索する

デジタル家電や高度なオフィス機器の普及により，これらに蓄えられるデータの種類や量，活用方法が多様化しています。組込み機器でも，このような状況に対応できるよう，多様なデータを簡単に扱うための仕組み“データベース（DB）”が利用されるようになってきています。

今回東芝は，組込みDBに多次元インデックス機能を追加し，次元数の大きなデータを従来方式より高速に検索する方法を開発しました。これにより，画像や音声などのマルチメディアデータやセンサからの時系列データなどを，簡単に，かつ高速に検索することが可能になり，機器のデータ検索能力を向上させることができます。

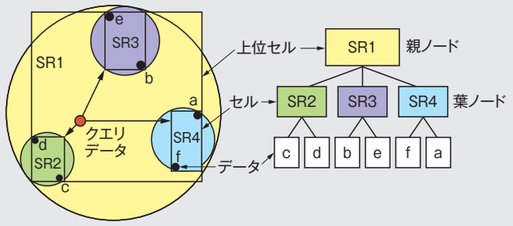

図1. SR-Treeインデックス — 木のノードを部分空間に，ノードの親子関係を部分空間の包含関係にそれぞれ対応付けます。木の探索を行うことで多次元空間を効率的に探索できます。

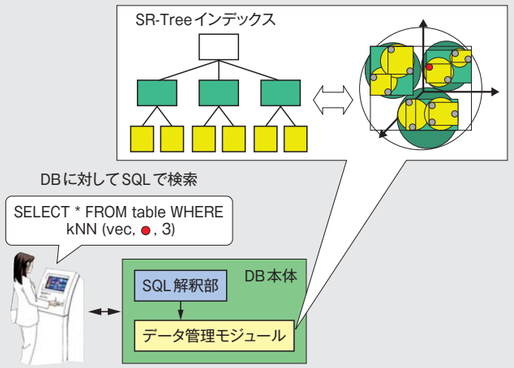

図2. DB内部に搭載した多次元インデックス — ユーザーは格納されているデータを意識することなく，DBと同様に，操作言語SQLを使って検索することができます。

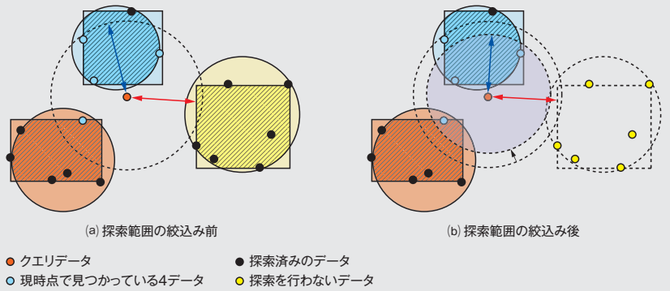

図3. 近似を用いた4近傍検索 — クエリデータから近傍4データまでの距離に比べて黄色のセルまでの距離は短いが，ANNでは，近傍データまでの距離に係数を掛けることで探索するセルの数を減少させ，検索を高速化します。

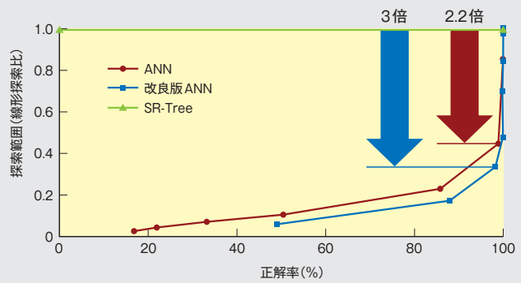

図4. 探索範囲と正解率 — 128次元の画像特徴データ20万件から20件の近傍検索を行い，探索効率を測定しました。SR-Treeでは線形探索と同等ですが，ANNの導入で2.2倍，更に，改良版ANNで3倍程度まで改善しています。

デジタル機器における高機能検索技術

近年，大量のデータを電子化して機器に蓄える事例が増えており，この中から様々な視点でデータを検索して活用したいという需要が高まっています。検索エンジンの普及に見られるように，データの活用は様々な発見や気づきを与えてくれます。このような検索をデジタル機器の中でコンパクトに実現できれば，機器の魅力を高めることが可能です。

東芝は，DBをデジタル機器に組み込むための研究開発を行い，データの格納や，検索，分類などのソフトウェアを共通的に提供することで，高信頼で多様なデータ処理機能を持つ機器を，効率的に開発できるように研究を進めています。

近傍検索の高速化

最近よく使われる画像や音声などのマルチメディアや時系列のデータは，複数の値を組み合わせて表現されます。蓄積したこのようなデータからクエリデータに類似するものを検索するには，複数の値で定義される空間上でクエリデータから距離が近いデータを指定個数だけ取り出す，近傍検索が必要となります。しかし，従来のDBは，複数の値を同時に比較して近傍を取り出す機能がないため，高速に近傍検索することが困難でした。

そこで当社は，高速に近傍検索できるSR-Tree（Sphere/Rectangle-Tree）インデックス機能を組込みDBに追加するとともに，近似による高速化手法を用いた近似近傍検索（ANN）を取り入れ，更にその近似方法も改良しました。デジタル機器にこの組込みDBとSR-Tree，改良版ANNを搭載することで，データ操作を共通で行えるようになり，また多様なデータを高速に検索できるようになります。

SR-Treeインデックス機能

SR-Treeは，木構造による探索絞込み機能を多次元空間に適用した技術で，木を構成するノードをセルという部分空間に対応付け，更にノードの親子関係をセルの包含関係に対応させたものです。セルは球と直方体を組み合わせた形状となり，蓄積される多次元データは葉ノードのセルに格納されます（**図1**）。

このSR-Treeをデータ管理モジュールに組み入れることで，多次元データ検索を高速化するためのインデックスとして機能させます。更に操作言語SQL（Structured Query Language）の解釈部を変更することで，DBから直接多次元データを扱えるようになります（**図2**）。

検索時にはクエリデータからセルまでの距離と見つかったデータまでの距離とを比較しつつ，深さを優先探索することで探索範囲を絞り込み，余分なノードやデータにアクセスすることなく効率的に近傍検索できます。

しかし，次元数が高い場合，近傍検索の鍵となる距離の区別が付きにくくなり，インデックスを用いてもほとんど探索範囲を絞り込むことができません。

高次元空間における探索範囲の絞込み

そこで，検索結果の完全性を保証しない代わりに，探索範囲を積極的に絞るANNの概念をSR-Tree上に実装しました。これは，探索候補となるセルまでの距離と現時点で見つかっている近傍データまでの距離を比較する際，近傍データまでの距離を実際より短く評価することで積極的な絞込みを実現する方法です。この方法では，セルまでの距離とデータまでの距離の差があまりない場合，探索候補のセルは，中のデータまでの距離が探索済みのデータより遠いことが多いため，探索対象から外し，近傍検索を高速化します（**図3**）。

更に，セルに格納されるデータの密度が低い場合は近傍解がないと判断し，より積極的に探索対象から外します。SR-Treeのセルは体積を計算できませんが，一般的に高次元空間ではデータの個数が増えると包含領域が爆発的に広がって密度が低下すると考えられるので，多次元データの格納数から絞込みの度合いを算出しています。

性能評価

SR-Treeのインデックスと改良版ANNの機能を組込みDBに追加し，128次元の画像特徴データ20万件からANNを行って効率と精度を評価しました。その結果，線形及び通常のSR-Treeによる近傍検索で得られる解の99％を，従来のANNで2.2倍，改良したANNでは3倍程度の効率で導出でき，検索の能力を改善できることを確認しました（**図4**）。

今後の展望

多次元インデックス技術をデジタル機器のDBに組み込むことによって，類似画像検索を用いたシーン検索やセンサデータからの診断を容易，かつ高速に行うことが可能となり，デジタル機器の更なるデータ検索能力向上が期待できます。

鹿山　俊洋
ソフトウェア技術センター
先端ソフトウェア開発担当主務