

ORDBMS(LiteObject)における複合メディア検索機構の実現

2Q-6

西岡 秀一 鬼塚 真 黒岩 淳一 芳西 崇

NTT サイバースペース研究所

1 はじめに

近年、計算機環境の著しい発展により、テキスト、画像、音楽等の情報のデジタル化が加速し、情報が増加している。これに伴い、デジタル化された情報の管理が重要になってきた。従来の市販DBMSでは、数値・文字情報の管理は容易であるが、画像・音楽等の蓄積・操作は困難であるため、我々はマルチメディア情報の操作に適したDBエンジンLiteObjectの開発を進めている。

現状、LiteObjectでは単一メディアの検索処理は可能となっているが、複数のメディアにまたがる検索を行えない。本稿では、複数のメディアにまたがる検索(複合メディア検索)を検討し、LiteObjectにおける実現法について述べる。

2 LiteObject

LiteObjectのアーキテクチャ[1]の特徴を以下に示す。

- (1)多様なデータ定義機能により、複合メディア情報の複雑なデータ構造を表現することが可能。
- (2)ユーザ定義に対応した拡張SQLをサポートすることにより、類似検索などの特殊な操作を記述したSQLの処理が可能。
- (3)組込み機能を用意することにより、メディア特化の機能追加が可能。組込み例として類似検索に適したデータ操作機能(多次元インデックス[2])がある。

3 音楽検索システムの概要

LiteObjectを適用した複合メディアシステムの

一つとして、音楽検索システム[3]がある。本システムの概略を、図1に示す。

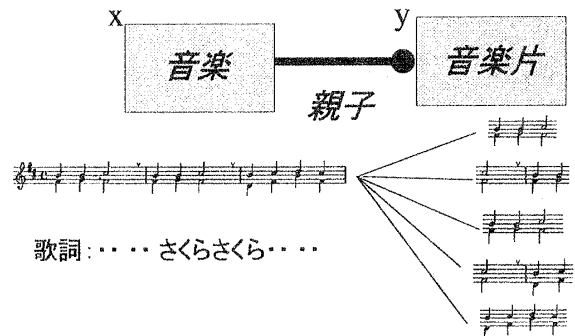


図1：音楽検索システムの概略

曲全体の楽音情報および歌詞情報を音楽クラス(親クラス)に格納する。曲全体から抽出した音楽片の特徴量を音楽片クラス(子クラス)に格納する。

本システムに対し、ユーザからメロディの一部をハミングで入力し、歌詞の一部をテキストで指定し、「それらに似た曲は？」という問合せが行われる。この問合せをSQL文に変換すると、下記のようなになる。

```
SELECT x.曲名, x.similarity
FROM 音楽 x, x.親子 y
WHERE y.特徴量 similar 'humming.mid'
AND x.歌詞 includes 'さくらさくら';
ORDER BY x.similarity
```

similarityはメソッドで、WHERE句で指定した条件の総合的な類似度を計算する。similarは、右辺と左辺に記述されている楽音情報の特徴量が類似しているかを判定するオペレータである。上記の場合、左辺は音楽片クラスに格納されている属性を示し、右辺はユーザがハミング入力したデータ(MIDI)である。

includesは、右辺と左辺に記述されている文章が類似しているかを判定するオペレータである。上記の場合、左辺は音楽クラスに格納されている

属性を示し、右辺はユーザが入力した文章である。

このSQL文のような複数のメディアにまたがった検索(複合メディア検索)は、単一メディアを指定した検索より、絞り込み効率が上がるため、ユーザにより正確に所望の曲を検索できると考えられる。

4 実現方法

複合メディア検索機構を実現するために、LiteObject に類似検索用オプション(インデックスおよびオペレータ)と文章検索用オプション(インデックスおよびオペレータ)を組込む。類似検索用オプションは2章で示したものを利用する。以下、文章検索用オプションについて検討する。

4.1 文章検索用インデックス

文章検索つまり全文検索については、従来より様々な研究がなされている。その代表的な手法として、 $tf \cdot idf$ 法がある。 $tf \cdot idf$ 法とは、あるキーワードが対象となる文書中に含まれる出現回数(tf : term frequency)と全文書数のうちのキーワードを含むファイル数の比(idf : Inverse Document Frequency)を利用し、単語の頻度をあらかじめ計算する方法である。この手法を用いた文章検索用インデックス[4]をLiteObjectに組込むことにする。

4.2 文章検索用オペレータ

前述した文章検索用インデックスに対して、LiteObject からデータアクセスを行うため、オペレータ(`includes`)をLiteObjectに定義する。

4.3 検索機構

類似検索用オプションおよび文章検索用オプションを組込んだLiteObjectの構造は図2のようになる。

LiteObjectでは、送信されたSQL文を以下のように処理する。

- `similar`に関して、入力されたハミングと類似度が高い音楽片クラスのインスタンス集合を演算する。
- a. で得られた結果から、親子関係を用いて、音楽クラスのインスタンス集合を求める。

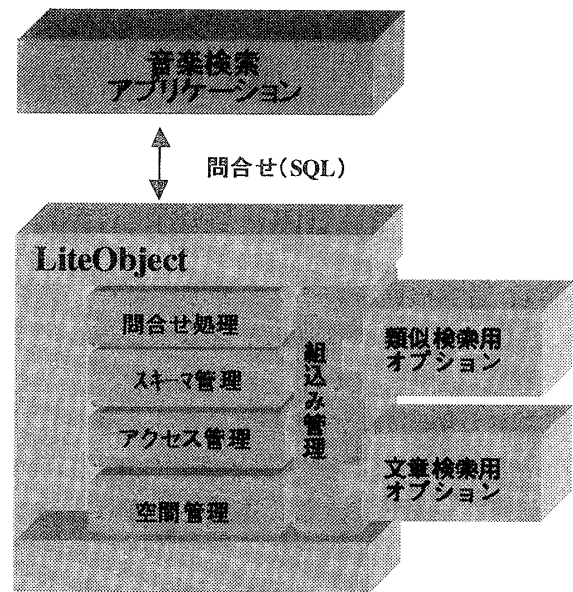


図2: LiteObjectの構造

- `includes`に関して、入力された文章と類似度が高い音楽クラスのインスタンス集合を演算する。
- bおよびcより得られた集合同士をマージし、総合的な`similarity`を求める。
- dの演算結果をソートし、アプリケーションに返却する。

5 おわりに

複合メディアシステムの実現に必要な複合メディア検索機構をLiteObject上に実現した。

今後、ユーザからの各メディアに対する重要度の付与に対応できるように拡張する予定である。

参考文献

- 岡田他, "高速ORDBMS LiteObjectの設計と実装," 信学会, 第9回データ工学ワークショップ, 1998.
- K.Curtis et al., "A Comprehensive Image Similarity Retrieval System that Utilizes Multiple Feature Vectors in High Dimensional Space," Proc. of ICICS '97, pp.180-184, 1997.
- 西原他, "時間正規化を用いたハミング検索システム," 音楽情報科学, 30-6, pp.27-32, 1999.
- 馬場 肇, "日本語全文検索システムの構築と活用," SOFTBANK, 1998.