

マルチメディアデータのための索引技術

吉川 正俊 奈良先端科学技術大学院大学情報科学研究科 yosikawa@is.aist-nara.ac.jp

植村 俊亮 奈良先端科学技術大学院大学情報科学研究科 uemura@is.aist-nara.ac.jp

大量に蓄積されたマルチメディアデータから、所望のデータを高速に検索するための索引技術を概観する。まず、マルチメディアデータ検索をメタデータ検索と内容検索に大別して整理する。高精度の内容検索のために特徴量が多次元ベクトルとして抽出され、その高速検索のために多次元ベクトル索引が開発されている。多次元ベクトル索引の設計因子をまとめ、最近の研究のうち特色のあるものを紹介する。

マルチメディアデータの索引付け

メタデータ検索と内容検索

マルチメディアデータの検索は、メタデータ検索と内容検索に大別することができる。メタデータ検索とは、マルチメディアデータの書誌情報や内容情報を表現したメタデータをもとにした検索である。たとえば、写真を例にとると、撮影日時、撮影場所、撮影条件などの書誌情報や内容の説明文などがメタデータになる。もとのマルチメディアデータからメタデータを得る作業を注釈付けと呼ぶ。メタデータは多くの場合、文字列や数値の形態をとるが、メタデータ自身がマルチメディアデータであってもよい(たとえば、写真のメタデータが音声として存在する場合も考えられる)。一方、

内容検索とはマルチメディアデータそのものの比較に基づく検索である。たとえば、ある写真を入力キーとし、それと類似した写真を求めることがそれに相当する。

メタデータは必ずしも存在するとは限らない。また、たとえ存在しても、メタデータのうち書誌情報は客観情報であり自動的に取得可能なものもあるが、内容情報は、通常、データの作成に人手を介する必要があるため、作成者の主観が入ることや、作成コストがかかるという問題点がある。また、メタデータの内容情報は、対象となるデータのある観点から表現したデータに過ぎないことから、マルチメディアデータの検索としては限界がある。さらに、前述のようにメタデータ自体がマルチメディアデータの場合もある。そのため、マルチメディアデータ自身の解析に基づく内容検索技術が必要となる。

マルチメディアデータの内容検索の枠組み

マルチメディアデータの索引付けのための枠組みを図-1に与える。注釈付けの結果得られるメタデータのうち、文字列や数値データの索引付けのためには、古典的な索引であるB木やハッシュを使える。一方、内容検索のためには、通常もとのマルチメディアデータから抽出した特徴量を用いる。多くの場合、特徴量は、多次元ベクトル空間内のデータで表現される。たとえば、静止画などの特徴量は、多次元ベクトル空間内の1つの

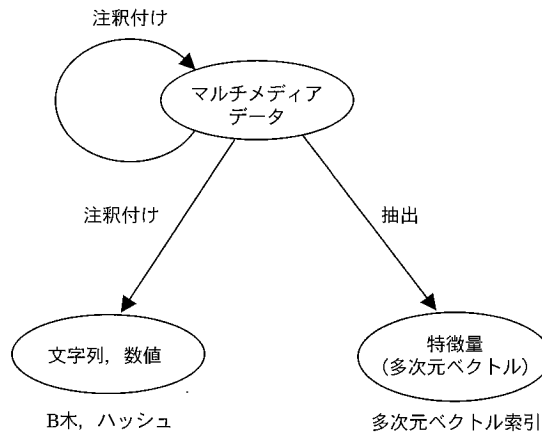


図-1 マルチメディアデータ索引の枠組み

点で表され、動画画像や時系列データなど時間とともに変化するデータの特徴量は、多次元ベクトル空間内の点の軌跡で表されることが多い。

対象とするデータベースは巨大であるため、逐次検索は实际的ではない。そこで、多次元ベクトルのための索引のデータ構造を工夫することにより、検索を高速化することが重要になる。

したがって、大量のマルチメディアデータ検索技術の開発においては、

- 特徴量の開発
 - 検索を高速化する索引データ構造の開発
- の2つの点が重要な課題となる。

特 徴 量

マルチメディアデータの内容検索を必要とする実際の応用では一致検索に加え類似検索が重要となる。類似検索の具体的な問合せ例としては、「ある写真と類似している写真を検索したい」、「ハミングで入力した音と類似している曲を知りたい」、「ある銘柄の株価と似た値動きをしている銘柄を知りたい」などがある。一般に類似度判定で着目する点は応用ごとに異なる。たとえば、画像の類似検索では、撮影対象は問わず画面全体の色合いを重視する場合や、色ではなく撮影対象の形状を重視する場合などがある。そこで、原データの中から類似度判定のために有効な情報だけを抽出し簡潔に表現した特徴量を用いる。特徴量は、各メディアごとに種々のものが開発されてきている。たとえば、

画像の特徴量としては、画像全体や部分画像の色情報や模様情報、画像中のオブジェクトの位置、形状情報などがある¹³⁾。また、データの種類によっては、データ量削減を主目的として特徴量を抽出する場合もある。たとえば、株価のような数値時系列データでは、データ間の類似度はユークリッド距離などで定義できるが、原データのままだでは容量が大きく類似度判定に時間がかかるため、よりデータ量の少ない特徴量を用いることがある。

特徴量データは次の性質を持っていることが望ましい。

- (1) もとのマルチメディアデータの持つ情報のうち類似度判定に必要な部分を可能な限り保存している。
- (2) 特徴量間の類似度を簡単に定義できる。
- (3) (特にデータ削減を目的とした特徴量については、) 特徴量データに基づく類似度検索を行っても検索漏れ (false dismissal) を生じない。

特徴量同士の非類似度の尺度として距離関数が導入される。距離関数としては、2つの n 次元ベクトルデータ $x = (x_1, x_2, \dots, x_n)$, $y = (y_1, y_2, \dots, y_n)$ の距離を $L_p(x, y) = (\sum_{i=1}^n |x_i - y_i|^p)^{1/p}$ で定義する L_p 距離が一般的であるが、ユークリッド距離 L_2 が用いられることが多い。

特徴量の例

時系列数値データ³⁾を例にとって、特徴量の説明をする。データベース中には、長さ n の時系列数値データ (以降、時系列データ) が大量に格納されており、2つの

データベース索引技術

時系列データ $x = [x_i], y = [y_i], i=1, \dots, n$ の距離はユークリッド距離 L_2 に基づくものとする。問合せとしても長さ n の時系列データを与え、問合せとの距離が ϵ 以内の時系列データを求めるものとする。 $X = [X_0, \dots, X_{n-1}]$ を x の n 点離散フーリエ変換 (DFT) (Y についても同様) とするならば、Parseval の公式から、

$$L_2(x, y) = L_2(X, Y)$$

が成立する。ここで、時系列データ x の特徴量 $F(x)$ として、DFT の最初の f ($\leq n$) 個の係数を取り、特徴量間の距離をユークリッド距離とすれば、前述の、特徴量データが持つべき性質のうち、(2) は満足される。また、

$$L_2(F(x), F(y)) \leq L_2(X, Y) \\ = L_2(x, y)$$

が成立することから、特徴量同士の距離はもとのデータ同士の距離の下限を与えるため、特徴量間の距離が ϵ 内のデータだけを残しても正しい答えが検索漏れになることはない。すなわち、この特徴量は前述の性質 (3) を満足する。また、性質 (1) については、 f の値が大きいほどよいことは明らかであるが、実験では $n=512$ のときに $f=6$ で十分な性能を得られることが報告されている³⁾。

多次元データ索引

これまでに、多くの多次元データのアクセス方法が提案されてきている⁴⁾。多次元データ索引の主要な設計因子には次のようなものがある。

- **前提とする距離関数:** 多くのものはユークリッド距離に基づくが、後述のように一般な楕円体距離関数を前提とするものもある。さらに、M木²⁾ などのように、距離の公理を満足している任意の関数を対象とするものもある。
- **対象とする問合せの種類:** 通常は次のいずれかまたは両方を対象とする。
 - k -近傍問合せ (k -nearest neighbor query)

与えられた問合せオブジェクトに近いものから順に k 個のオブジェクトを求める問合せ。
 - 範囲問合せ (range query)

与えられた範囲 (矩形または球など) に存在するすべてのオブジェクトを求める問合せ。
- **動的なデータ更新:** データを動的に挿入したり、削除することを許すか否か。

索引のデータ構造は、従来の文字列、数値のための索引と同様に、データを階層的にクラスタ化した木構造やハッシュに基づくものがある。さらに、データや包囲矩形を近似表現してデータ量を削減する方法や、多次元データのある関数で1次元化し、B木など従来の索引を利用する方法やこれらの方法を組み合わせる方法も提案されている。

木構造の代表的な索引はR木⁵⁾ である。R木は、地図などの2次元データを対象として考案されたファイル構造であるが、原理は、3次元以上の多次元データにも適用できる。データの更新に伴うデータ構造の動的な変更を許すことや分枝限定法に基づく k -近傍探索法⁶⁾ が開発されたことなどにより、R木をもとにそれを改良した索引構造が数多く提案されてきている。R木のファイル構造は本特集の第3編「空間データの効率的な管理と高速空間検索のためのデータ構造」で解説されるため、ここでは省略する。1次元データを対象とするB木の場合は、データをクラスタ化し木構造にすることにより、データ数に対して対数オーダの時間でアクセスを可能とするが、R木の場合は、多次元データを対象にすることと包囲矩形同士が重なることのために、検索時にバックトラックがある。そのために、木構造ではあっても、B木のようにデータ数に対して対数オーダの時間でデータアクセスが可能であることは保証されておらず、次元が高くなると線形オーダの時間に近づくことが指摘されている¹²⁾。そこで、R木の改良版ではバックトラックの回数を減らすためのさまざまな工夫がされている。

多次元データ索引については、研究面ではし烈な性能競争が展開されているが、文字列、数値データのためのB木に相当するような決定版が現れるには至っていない。ここでは、最近の研究のうち、B木を利用するデータ構造と距離関数として楕円体距離関数を用いるものを紹介する。

B木の利用

一般に、新しい索引構造をデータベース管理システム (DBMS) に組み込むためには、問合せ言語、問合せ最適化器、並行処理機構、障害回復機構などを拡張する必要があるため、多大な開発コストを要する。したがって、このような開発コストをかけるに値する多次元データ索引でない限りDBMSへの組み込みは行われなない。索引データ構造によって性能に極端に大きな差がない場合は、実装が容易かどうかが重要な点となる。このことを考慮し、多次元データ索引をB木の上に構築する手法も提案されている。B木は標準的な索引であり、これを用いた問合せ最適化アルゴリズムや並行処理、

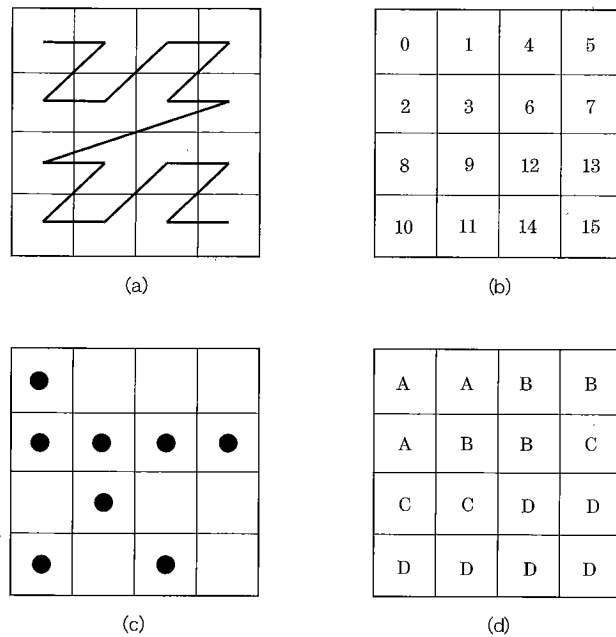


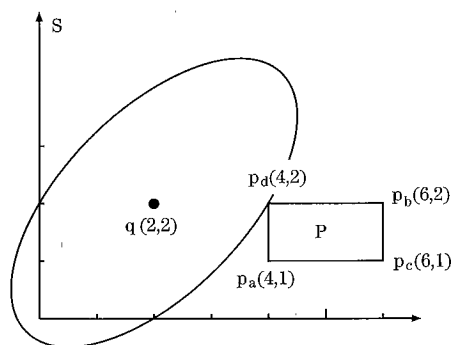
図-2 UB木における空間のアドレス付け

障害回復などについても長年の技術の蓄積があり、ほぼすべてのDBMSで安定稼働している。B木に格納されるデータは全順序が付けられているデータであるため、多次元データをB木上に構築するためには、それらに何らかの方法で全順序を付ける必要がある。このためには、通常、空間充填曲線 (space filling curve) を用いて、多次元空間を1次元空間に写像する手法を用いる。たとえば、B木の考案者であるBayerのグループで開発されたUB木⁷⁾ではZ曲線を用いている。たとえば、4×4の2次元平面の場合は、図-2 (a) のようなZ曲線でセルの順序が付けられる。図-2 (b) は各セルのZアドレスと呼ばれる。各ページの容量をオブジェクト2個とし、図-2 (c) のようにオブジェクトが分布していたと仮定すると、全平面は図-2 (d) に示すように4ページで被覆される。多次元空間上の範囲問合せは、B木上の複数の範囲問合せに変換され処理される。たとえば、Zアドレスの2, 3, 8, 9の範囲問合せは、B木上での2から3の範囲問合せと8から9の範囲問合せに変換されることになる。

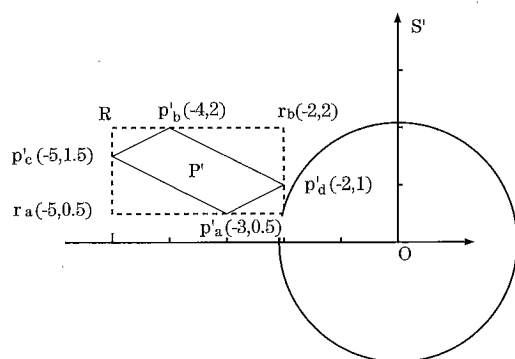
楕円体距離関数

多くの多次元データ索引が前提としているユークリッド距離は各次元の独立性を仮定しているが、実際の特徴量の中には必ずしも独立性を仮定できないものがある。たとえば、画像の色ヒストグラムを多次元ベクトルで表した場合、単純にユークリッド距離を用いたのでは、赤色と橙色の距離は赤色と青色の距離と等しくなり、人間の直観と合わなくなる。そこで、より現実に即した距離関数として楕円体距離関数を用いた索引も提案されている。問合せ行列を M とすると、2つのベクトル $\mathbf{x}, \mathbf{y}$ の楕円体距離は、 $d_M^2(\mathbf{x}, \mathbf{y}) = (\mathbf{x} - \mathbf{y})^t M (\mathbf{x} - \mathbf{y})$ で表される。次元間の依存関係を表現する問合せ行列 M は、利用者の嗜好を反映し適応的であることが望ましい。このような条件のもとで、近傍検索を高速化するための方法が提案されている^{10), 1), 9)}。

櫻井らの空間変換法⁹⁾では、問合せ行列を利用して空間変換を行う。例として2次元空間の場合を考えると、図-3に示すように、もとの空間における問合せ点と矩形との楕円体距離は、変換後の空間における問合せ点と平行四辺形とのユークリッド距離に変換される。さらに平行四辺形をその最小包囲矩形で近似することにより問合せ点との距離の下限を高速に計算する。問合せ



(a) もとの空間における矩形



(b) 変換後の空間における平行四辺形とその最小包囲矩形

図-3 空間変換の例⁹⁾

結果に残らないデータをこのように高速に枝刈りすることにより、検索性能を向上させている。

今後の展望

今後、応用の拡大や計算機環境の進化により、マルチメディアデータの索引技術は、これまでの蓄積の上にさらなる展開が予想される。最近の研究の中から将来重要性を増すと予想されるものを例として挙げる。

新たな応用への対応: たとえば、巨大な仮想3次元空間をウォークスルーする場合には、ウォークスルーするに従って変化する利用者の視野範囲の景観情報を順次取得する必要がある、問合せ点が連続的に変化する¹¹⁾。このように、応用範囲の拡大とともに、従来想定していなかった種類の問合せが必要となり、その

高速処理に適した索引構造を開発することが必要となる。

メモリ上の索引: これまで提案されたほとんどすべてのマルチメディアデータ索引は、データが磁気ディスクに格納されていることを前提としていた。したがって、ディスクアクセスコストを削減することが索引の重要な設計目標であった。ところが、最近のメモリの大容量化と低価格化により、索引全体をメモリに格納することも可能となってきた。メモリ常駐型でキャッシュを意識したR木の提案⁶⁾などもあり、今後の重要な方向であろう。

参考文献

- 1) Ankerst, M., Braunmüller, B., Kriegel, H-P. and Seidl, T.: Improving Adaptable Similarity Query Processing by Using Approximations, In Proc. of the 24th International Conference on Very Large Data Bases (VLDB), pp.206-217, New York City, NY (Aug. 1998).
- 2) Ciaccia, P., Patella, M. and Zezula, P.: M-tree: An Efficient Access Method for Similarity Search in Metric Spaces, In Proc. of the 23rd International Conference on Very Large Data Bases (VLDB), pp.426-435, Athens (Aug. 1997).
- 3) Faloutsos, C., Ranganathan, M. and Manolopoulos, Y.: Fast Subsequence Matching in Time-Series Databases, In Proc. ACM SIGMOD International Conference on Management of Data, pp.419-429 (May 1994).
- 4) Gaede, V. and Günther, O.: Multidimensional Access Methods, ACM Computing Surveys, Vol.30, No.2, pp.170-231 (June 1998).
- 5) Guttman, A.: R-Trees: A Dynamic Index Structure for Spatial Searching, In Beatrice Yormark, editor, SIGMOD'84, Proceedings of Annual Meeting, Boston, Massachusetts, June 18-21, 1984, pp.47-57, ACM Press (1984).
- 6) Kim, K., Cha, S. K. and Kwon, K.: Optimizing Multidimensional Index Trees for Main Memory Access, In Proc. ACM SIGMOD International Conference on Management of Data, pp.139-150 (May 2001).
- 7) Ramsak, F., Markl, V., Fenk, R., Zirkel, M., Elhardt, K. and Bayer, R.: Integrating the UB-Tree into a Database System Kernel, In Proc. of the 26th International Conference on Very Large Data Bases (VLDB), pp.263-272 (Sep. 2000).
- 8) Roussopoulos, N., Kelley, S. and Vincent, F.: Nearest Neighbor Queries, In Proc. ACM SIGMOD International Conference on Management of Data, pp.71-79 (May 1995).
- 9) Sakurai, Y., Yoshikawa, M., Kataoka, R. and Uemura, S.: Similarity Search for Adaptive Ellipsoid Queries Using Spatial Transformation, In Proc. of the 27th International Conference on Very Large Data Bases (VLDB), Roma, Italy (Sep. 2001).
- 10) Seidl, T. and Kriegel, H-P.: Efficient User-Adaptable Similarity Search in Large Multimedia Databases, In Proc. of the 23rd International Conference on Very Large Data Bases (VLDB), pp.506-515, Athens (Aug. 1997).
- 11) Tan, K-L., Shou, L., Huang, Z., Chionh, J. and Ruan, Y.: Walking Through a Very Large Virtual Environment in Real-time, In Proc. of the 27th International Conference on Very Large Data Bases (VLDB), Roma, Italy (Sep. 2001).
- 12) Weber, R., Schek, H-J. and Blott, S.: A Quantitative Analysis and Performance Study for Similarity-Search Methods in High-Dimensional Spaces, In Proc. of the 24th International Conference on Very Large Data Bases (VLDB), pp.194-205, New York City, NY (Aug. 1998).
- 13) 串間和彦, 赤間浩樹, 紺谷精一, 山室雅司: 色や形状等の表層的特徴量にもとづく画像内容検索技術, 情報処理学会論文誌データベース, Vol.40, No.SIG3 (TOD1), pp.171-184 (Feb. 1999).

(平成13年8月10日受付)

