ムービングオブジェクトデータベースにおける類似検索のための 1 次元索引付け手法とその評価

One-Dimensional Indexing of Moving Object Database for Similarity Search and its Evaluation

北原 由美子♥　　増永 良文♠

Yumiko KITAHARA　Yoshifumi MASUNAGA

我々は，3 次元空間において物体の動きに関するデータを市販のモーションキャプチャシステムQuickMAGIVを用いて取得し，様々な問合せを実現するムービングオブジェクトデータベースシステムの構築を進めてきた．本論文では，このシステムで類似検索処理を効率化するために，データに索引付けするにあたり，多次元の空間索引を付与するのではなく，現在ビジネスデータ処理の分野で普及している 1 次元の索引であるB+木を用い索引付けすることを考えた．この索引付けは，データベース内のある一つのデータと他のデータ間の相違度を探索キーとしてデータを 1 次元で構造化する．この索引法が類似検索でどれほど有効かを，データを構造化しない場合と比較して示す．

A moving object database system has been under development at Ochanomizu University. It captures motion object data in the three dimensional space, where the data are captured using a commercial motion capturing system named QuickMAGIV. In this paper, a novel indexing method of moving object data using B+-tree, a widely used index in business data processing, is investigated. The essential difference between the traditional spatial indexing and our approach is that the former is multi-dimensional while the latter is one dimensional. The one dimensional index is given according to the dissimilarity measure between the data and one fixed data a moving object database. Effectiveness of this approach is examined by comparing it with non-indexed case under similarity search.

1. はじめに

近年，モーションキャプチャリングシステムやGPSなどを用いて，動く物体の位置や姿勢などのデータを計測するセンシング技術が発達している．それに伴い，こうした動きのデータに対して様々な問合せや分析を行いたいという要求が高まっている．そこで我々は，3次元空間において，物体の動きに関するデータを取得し，様々な問合せを実現するムービングオブジェクトデータベースシステムの構築を進めてきた[1,2,3]．本システムの主要な機能である類似検索機能は，Query-By-Exampleの発想に基づいて行い，データ間の相違度をユークリッド距離関数によって計算している．これまでは類似検索を行うのに，問合せデータを与えたとき，データベースに格納されている全データとの相違度を総当りで計算していた．しかし，これでは格納されているデータ量が多くなるにつれ，検索時間が増加してしまう．そのために，データを索引付けすることが考えられるが，ムービングオブジェクトのデータは，空間的な位置や時間を属性として有するために，一般にはR木といった多次元の索引付け手法を必要とする．しかし，現在，多用されている索引付け手法であるB+木は“1次元”の索引であるので，Oracleなど市販のデータベースシステムを使ってムービングオブジェクトデータを格納して検索するにはB+木を使ったムービングオブジェクトデータの索引付けを考えることが必要である．この観点から，我々は，先行研究 [4] で索引を1次元のスカラー値で与える方法を提案した．本論文では，その手法のもとで類似検索処理時間が索引を付与された場合と付与されない場合とでどれほどの改善があるのか，その有効性を論じる．

2. ムービングオブジェクトデータベースと類似検索

2.1 システム概要

本研究でのシステム概要を図1に表す．ムービングオブジェクトデータの計測には，センシング装置には光学式のモーションキャプチャリングシステムであるQuickMAG Ⅳ（応用計測研究所製）を使用する．

オブジェクトの 3 点にカラーマーカを付け，ステレオカメラを用いオブジェクトの動きを 3 次元で計測する．計測後，ユーザは格納インタフェースで計測データのシーンやオブジェクトに関する情報を登録し，挿入エンジンによりムービングオブジェクトデータモデルに従ったデータに変換した上でムービングオブジェクトデータベースに格納される．また，ユーザは問合せインタフェースによって検索エンジンを制御し，Query-By-Example の発想に基づく問合せを行い，検索エンジンを通して検索結果を返す．

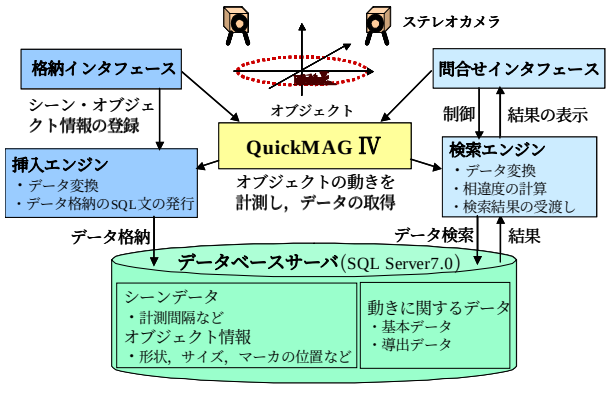

図 1 システム概要
Fig.1 A System Overview

2.2 動きとは

計測されたオブジェクトの動きは，オブジェクトの中心座標，向き，傾き，及びそれに付与された時刻印の4要素で構成される．正式には，計測周波数fで時刻t_sからt_eまで計測されたオブジェクトの動きをベクトル$\vec{M}=\langle\vec{m}_1,\vec{m}_2,\cdots,\vec{m}_n\rangle$で表す．このとき$\vec{m}_i=\langle\vec{p}_i,\vec{or}_i,\vec{gr}_i,t_i\rangle$であり，各要素はオブジェク

♥ 学生会員　お茶の水女子大学大学院人間文化研究科博士前期課程　yumiko@db.is.ocha.ac.jp

♠ 正会員　お茶の水女子大学理学部情報科学科 masunaga@is.ocha.ac.jp

トの位置ベクトル $\vec{p}_i = \langle x_i, y_i, z_i \rangle$，向きベクトル $\vec{or}_i = \langle ox_i, oy_i, oz_i \rangle$，傾きベクトル $\vec{gr}_i = \langle gx_i, gy_i, gz_i \rangle$ を表している．また，$t_s = t_0$，$t_e = t_n$，$n = (t_e - t_s) \times f$ である．

2.3 動きの同一性と基本類似性

動きの同一性の定義について述べた後，動きの各要素の基本類似性の定義について述べる．

【定義 2.1】（動きの同一性）

動き $\vec{M}$ と $\vec{M}'$ が次の条件を満たすとき，2 つの動きは同一であるという．

1. $f = f'$
2. $t_s = t'_s \wedge t_e = t'_e$
3. $(\forall i)\vec{p}_i = \vec{p}'_i$

【定義 2.2】（位置の基本類似性）

動き $\vec{M}$ が $\vec{M}'$ に対して次の条件を満たすとき，位置について ε-類似しているという．

1. $f = f'$
2. $t_s = t'_s \wedge t_e = t'_e$
3. $D_p(\vec{M}, \vec{M}') = \dfrac{\sqrt{\sum_{i=1}^{n} |\vec{p}_i - \vec{p}'_i|^2}}{n} \leq \varepsilon$

なお，定義から分かるように，これらは開始時刻と終了時刻が同じでかつ計測周波数が等しい2つのムービングオブジェクトデータに対する定義である．そこで，格納データの長さが異なる場合に対応する検索法として，タイムワーピング距離関数[5]を用いてデータ間の相違度を計算する方法や，データ自身を伸縮し長さを等しくした後ユークリッド距離関数で相違度を計算する方法を取入れた．それぞれの方法で類似検索を行ったときの検索結果の比較を次節で示す．

2.4 ユークリッド距離関数の採用

ムービングオブジェクトデータには，模型列車を円軌道上を等速(train_c)，速-遅(train_fs)，速-遅-速-遅(train_fsfs)の3速度変化パターンで走らせた300個データを用いる．表1に，問合せデータに train1_c と train1_fs を与えたときの検索結果をランキングで示したものと検索に要した時間を示す．

表 1 類似検索結果と検索時間

Table 1 Similarity Search Results and Execution Time

	train1_c			train1_fs	
	ユークリッド	Time Warping		ユークリッド	Time Warping
1	train1_c	train1_c	1	train1_fs	train1_fs
2	train2_c	train2_c	2	train2_fs	train2_c
3	train2_fsfs	train2_fsfs	3	train1_fsfs	train1_c
4	train1_fsfs	train1_fs	4	train2_fsfs	train2_fsfs
5	train2_fs	train2_fs	5	train1_c	train2_fs
6	train1_fs	train1_fsfs	6	train2_c	train1_fsfs
	2.82 (s)	72.95 (s)		2.84 (s)	74.76 (s)

この結果から，タイムワーピング距離を用いた場合，走行パターンがうまく識別されていないことが分かる．またタイムワーピング距離を用いると，要素同士の比較回数が膨大になるため，検索に要する時間は平均約 26 倍もかかってしまう．そのため，ムービングオブジェクト同士の類似性を取り扱うためには，ユークリッド距離関数を用いることにする．

3. ムービングオブジェクトデータの 1 次元索引付け

ムービングオブジェクトデータは時間属性を有しているという意味で時系列データである．これまで，時系列データを構造化するための技術として様々な索引付けの研究が行われてきた．よく知られた方法として，離散フーリエ変換などの特徴抽出関数を用いて時系列データを2次元の周波数空間にマッピングして，R木やK-D木に代表される空間アクセス法を適用する方法がある[6,7]．注意すべきこととして，このような変換により，一般には多次元の索引構造が誘引される．一方，売上や社員管理といったビジネスデータ管理を主目的とするリレーショナルデータベースではB木やB+木といった索引付けの手法が使われているが，それはデータのある一つの属性値に基づいた索引であるという意味で，1 次元の索引構造である．これらの索引付け手法は現在稼動しているほとんどのデータベース管理システム上で使用可能となっている．さらに，リレーショナルデータベースでは，BLOB データ型の導入で，時系列データを属性値として直接格納できる状況となっている．したがって，Oracleなど既存のリレーショナルデータベースシステムを用いてムービングオブジェクトデータ管理を行う場合，ムービングオブジェクトデータにB+木などの 1 次元の索引付けができるか否かを研究しておくことが重要となる．

3.1 相違度に基づく索引付け

まず，データベースに格納されているムービングオブジェクトデータの中からある一つのデータを選定し，それを“基準データ”とする．次に，基準データと他の格納データとの相違度をユークリッド距離関数によって計算する．相違度は1次元の値なので，この値を索引としてB+木を用いてデータを構造化する．B+木はデータの挿入順で木構造が変わってくるが，ここではデータベースに格納されている順にデータを索引付けすることとする．

こうして構造化したデータに対して類似検索を行う，しかし，この基準データとの相違度，つまり各データに付与された索引は相対的なものであるので，実データ同士の距離関係が正確に保存されていない．そのため，本来類似していないデータを検索結果として返してしまう過多誤認が生じる恐れがある．一方，過小誤認は発生しないことは明らかにされてる[4]．そこで次に，この索引を利用しながらも過多誤認を排除する類似検索法について説明する．

3.2 索引を利用した類似検索法

ある問合せデータ $\bar{Q}$ が与えられたとする．

【手順Ⅰ】まず，基準データ $\bar{R}$ と $\bar{Q}$ との相違度 $D_p(\bar{R}, \bar{Q})$ を求める．格納されている全データには，$\bar{R}$ との相違度が索引として付与されているのでその値に着目し，

$$D_p(\bar{R}, \bar{Q}) - \varepsilon \leq D_p(\bar{R}, \bar{O}') \leq D_p(\bar{R}, \bar{Q}) + \varepsilon$$

となる索引をもつ全てのデータ $\bar{O}'$ を検索する．この $\bar{O}'$ が全検索結果候補データとなる．

【手順Ⅱ】手順Ⅰで検索された候補データ $\bar{O}'$ と問合せデータ $\bar{Q}$ との実際のユークリッド距離を計算し，

$$D_p(\bar{Q}, \bar{O}) \leq \varepsilon$$

となる索引をもつ $\bar{O}$ が最終的に $\bar{Q}$ と ε-類似している検索結果となる．

例えば，与えられた問合せデータの索引が 4.8 とすると，4.8 に一番近い値の葉ノードにたどり着く．その前後で 4.8

±εに入る探索キーをもつ葉ノードが候補データとなる．またε＝0のときは特殊な場合であるが，問合せデータに格納データを用いれば検索可能である．

3.3 基準データの選定法

以上に述べた手法により類似検索を行う場合，基準データの選定法が問題となってくる．シミュレーションによりその最適解を探ったところ，データベースに格納されたデータの分布が正規分布に近い偏りである場合，その中心からある程度離れた位置のデータを基準データに取ることで，類似検索時に検証すべきデータ数をある程度抑えられることが明らかになったので，格納されている全データの時刻ごとの座標値を平均して一つのデータを作成し，この平均データ $\bar{A}$ との相違度が最も大きいデータを基準データ $\bar{R}$ と定めている[4]．こうして選定した基準データとの相違度を用いた類似検索の検索範囲を簡単のため，2次元の点データを用いたものを図2に示す．問合せデータ $\bar{Q}$ と ε-類似しているものを検索するとき，図2の環状の網掛け部分に含まれるデータが検索結果の候補データ，円状の斜線部分に含まれるデータが最終的な検索結果である．これまでは問合せデータを与える度にデータ間の相違度を総当りで計算していたが，データをこの手法を用いて索引付けすることにより，データ間の相違度を計算する回数が減り，検索に要する時間の削減が期待できる．

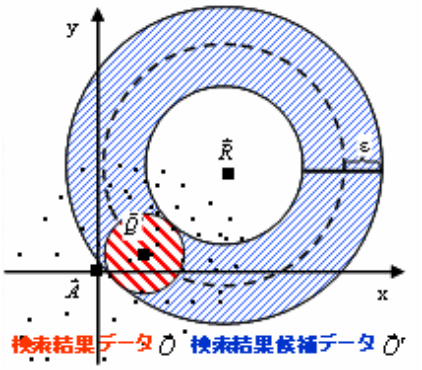

図2 検索範囲
Fig.2 Search Ranges

4. 検索時間のモデル化

総当りで検索を行った場合と，1次元索引付けしたデータに対して検索を行った場合の検索時間を第1次近似でモデル化したものについてそれぞれ論じる．

4.1 総当りによる検索時間

ムービングオブジェクトデータが計測順にデータベースに格納されているとする．ある一つの問合せデータを与えたとき，格納データ一つ一つとの相違度を計算するのにかかる平均コストをそれぞれC_{path0}とすると，全データ数がnのときの検索時間C_{Ex}は(1)式の通りである．

$$C_{Ex} = C_{path0} \times n + C_0 \quad \text{(s)} \qquad (1)$$

ここでC_0はデータ数とは無関係な定数項である．

4.2 1次元索引付けによる検索時間

ムービングオブジェクトデータが前章で述べた方法でB⁺木を用いて1次元で索引付けされているとする．ある一つの問合せデータを与えたとき，そのデータの索引と一番近い値

表2 検索時間（ε＝500）
Table 2 Execution Time（ε＝500）

問合せデータ	data1	data2	data3	data4	data5	data6
総当り	13.473	13.088	13.204	13.492	13.268	13.183
一次元索引構造	14.721	14.65	14.416	15.189	15.03	14.862
検索結果数	600	600	600	600	600	600
性能比(倍)	0.915	0.893	0.916	0.888	0.883	0.887

表3 検索時間（ε＝100）
Table 3 Execution Time（ε＝100）

問合せデータ	data1	data2	data3	data4	data5	data6
総当り	13.272	13.39	13.663	13.28	13.387	13.274
一次元索引構造	2.592	3.86	7.602	7.157	4.912	7.99
検索結果数	8	62	256	178	122	210
性能比(倍)	5.12	3.47	1.8	1.86	2.73	1.66

表4 検索時間（ε＝50）
Table 4 Execution Time（ε＝50）

問合せデータ	data1	data2	data3	data4	data5	data6
総当り	13.264	13.53	13.274	13.249	13.11	13.205
一次元索引構造	0.739	2.037	4.793	2.74	1.936	2.861
検索結果数	4	20	114	34	32	44
性能比(倍)	17.9	6.64	2.77	4.84	6.77	4.62

表5 検索時間（ε＝0）
Table 5 Execution Time（ε＝0）

問合せデータ	data1	data2	data3	data4	data5	data6
総当り	13.21	13.319	13.503	13.09	13.247	13.469
一次元索引構造	0.328	0.329	0.343	0.338	0.344	0.341
検索結果数	1	1	1	1	1	1
性能比(倍)	40.3	40.5	39.4	38.7	38.5	39.5

の葉ノードまでたどるのにかかるコストをC_{path1}，類似検索法の手順ⅠでかかるコストをC_{path2}，手順Ⅱでかかる平均コストをそれぞれC_{path3}とする．C_{path3}=C_{path0}に注意する．過多誤認検証の対象となる候補データ数をm_εとすると検索時間C_{B+tree}は(2)式の通りである．

$$C_{B+tree} = C_{path0} \times m_\varepsilon + C_{path1} + C_{path2} + C'_0 \quad \text{(s)} \qquad (2)$$

ここでC'_0はデータ数とは無関係な定数項である．

5. 検索時間の性能評価

卓球のラケットの3点にカラーマーカを付け，そのフルスイングの状態をQuickMAG Ⅳで計測しムービングオブジェクトデータを取得する．この計測データを上記の1次元の索引構造を用いて構造化したものに対して類似検索を行い，検索に要した時間を計測し，従来の総当り法の検索時間と比較する．また今回は木が最も深くなる最悪の場合を想定し，木のファンアウト数は2としている．さらにデータベースから問合せデータをランダムに抽出し，εの値を変えて類似検索を行う．

データ数を600とし，ε＝500，100，50，0としたときの検索結果を表2，3，4，5に示す．これらの表から従来の総当りの場合，どんな問合せデータを与えても，またεがどんな値であっても検索時間はほぼ一定であることが分かる．一方，1次元索引付けした場合，問合せデータによって検索時間は異なる．この差は(2)式より候補データ数m_εの影響といえる．表2よりε＝500では，索引付けしたほうが平均約

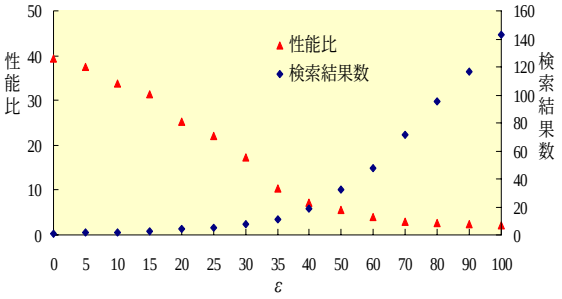

図 3 ε と性能比および検索結果数の関係（データ数 600）
Fig.3 Relation between ε and its performance (600data)

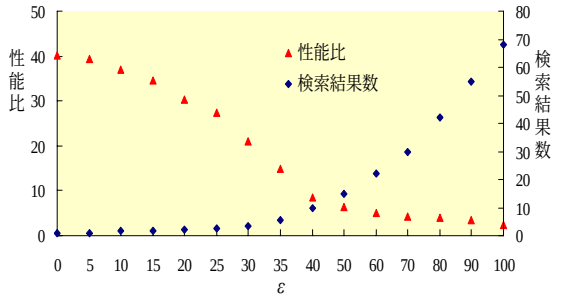

図 4 ε と性能比および検索結果数の関係（データ数 300）
Fig.4 Relation between ε and its performance (300data)

0. 897 倍性能が低下してしまっている．この原因としては，検索結果数が全データ数となっていることから，候補データ数が全データ数（m_{500}= 600）となってしまったためと考えられる．また，検索結果数はεとともに増減するが，問合せデータによって検索結果数の絞込みに差が生じる．そこで両極端のデータ，ここではdata1 とdata3 を除いたデータの平均値に基づいてεと性能比，および検索結果数の関係を調べたところ図 3 のようになった. 同様にデータ数を 300 と固定した検索結果より図 4 の関係が導かれた．図 3，図 4 から，全データ数の約 5%分のデータを検索結果として取得したいときはε = 50 とすればよい．このように，εの値が増加するにつれ性能比は減少していくが，データ数に関わらずこの 1 次元索引付け手法がムービングオブジェクトデータベースにおける類似検索において有効であることが実証された．

また，問合せデータに yumiko1 を与えたときの検索結果は，総当り法と索引付けした場合の両方において表 6 のようになった．検索結果は類似順に並んでいる．これより過多誤認や過少誤認が発生していないことが分かる．

6. まとめと今後の課題

本稿では，現行のムービングオブジェクトデータベースにおいて，より効率的な類似検索を実現するため，ムービングオブジェクトデータをB^+木を用いて 1 次元で構造化する方法について述べ，実際に索引付けしたデータに対し類似検索を行い，その有効性を示した．今回の実装では，動きの要素である位置，向き，傾きデータを各々索引付けしたため，各要素に重み付けした検索には対応していない．今後は，オブジェクトが回転するなど全ての要素データが重要となる複雑な動きの場合の類似検索が可能となるよう，動きの要素を統合させた索引付けを検討する必要がある．

表 6 総当り法および 1 次元索引構造の検索結果
Table 6 Search Result Ranking

Rank	SceneID	SceneName	ObjectID	ObjectName	Dissimilarity
1	251	yumiko001	251	pingpong	0
2	262	yumiko012	262	pingpong	19.2246
3	222	yukari022	222	pingpong	20.1776
4	95	yukari015	95	pingpong	21.2704
5	260	yumiko010	260	pingpong	24.9248
6	94	yukari014	94	pingpong	30.2494
7	53	nakamura003	53	pingpong	31.6042
8	259	yumiko009	259	pingpong	31.6314
9	175	nahoko015	175	pingpong	35.5882
10	264	yumiko014	264	pingpong	37.5081
11	176	nahoko016	176	pingpong	38.0964
12	263	yumiko013	263	pingpong	39.9760
13	173	nahoko013	173	pingpong	43.7946
14	192	ray022	192	pingpong	46.9150
15	96	yukari016	96	pingpong	46.9440
16	180	nahoko020	180	pingpong	47.6504
17	129	yokokawa009	129	pingpong	50.2071
18	196	ray026	196	pingpong	50.6375
19	131	yokokawa011	131	pingpong	53.4804
20	62	ozaki002	62	pingpong	54.1152
21	182	ray012	182	pingpong	55.9660
22	188	ray018	188	pingpong	56.7749
23	54	nakamura004	54	pingpong	57.2358
24	97	yukari017	97	pingpong	57.2407
25	65	ozaki005	65	pingpong	57.6135

[文献]

[1] Y. Masunaga and N. Ukai, “ Toward a 3D Moving Object Data Model -A Preliminary Consideration-, ” Proceedings of the 1999 International Symposium on Database Applications in Non-Traditional Environments, pp.306-316, November 1999.

[2] 水崎聡子，増永良文，“ムービングオブジェクトデータベースシステムのための類似検索機能の実現に向けて，”情報処理学会データベースシステム研究会報告，vol.2001, no.70, pp.217-223, 2001.

[3] 河内聡恵，増永良文，“ ムービングオブジェクトの速度変化パターンを識別できる類似検索機能の導入，” 日本データベース学会 Letters，Vol.2，No.1, pp.15-18, May 2003.

[4] 澤井美弥，“ 基準データとの相違度に着目したムービングオブジェクトデータの一類似検索法，” 電子情報通信学会技術研究報告（DBWS2003），Vol.103，No.191，pp.247-252, July 2003.

[5] B.-K. Yi, H. Jagadish, and C. Faloutsos, “ Efficient Retrieval of Similar Time Sequences Under Time Warping, ” Proceedings of the International Conference on Data Engineering, pp.201-208, IEEE CS, Orlando, FL, February 1998.

[6] R. Agrawal, C. Faloutsos, and A. Swami, “ Efficient similarity search in sequence databases , ” Proceedings of the 4th International Conference on Foundations of Data Organizations and Algorithms (FODO ’93), pp.69-84, Chicago, October 1993.

[7] D. Rafiei and A. O. Mendelzon, “Querying Time Series Data Based on Similarity,” Proceedings of the 8th DASFAA, pp.267-274, Kyoto, March 2003.

北原 由美子 Yumiko KITAHARA

2005 お茶の水女子大学大学院人間文化研究科博士前期課程数理・情報科学専攻修了．ムービングオブジェクトデータベースシステムの研究に従事．日本データベース学会学生会員．

増永 良文 Yoshifumi MASUNAGA

お茶の水女子大学理学部情報科学科教授．1970 東北大学大学院工学研究科博士課程修了，工学博士．データベースシステムの研究・開発に従事．情報処理学会および電子情報通信学会フェロー．日本データベース学会会長．著書に「リレーショナルデータベース入門［新訂版］」(サイエンス社)など．