2003－DBS－131　(66)
2003／7／17

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ベースの一類似検索法

澤井 美弥[†]　北原 由美子[†]　増永 良文[‡]

†お茶の水女子大学人間文化研究科数理情報科学専攻 〒112-8610 東京都文京区大塚 2-1-1
‡お茶の水女子大学理学部情報科学科 〒112-8610 東京都文京区大塚 2-1-1

E-mail: †{miya, yumiko}@dblab.is.ocha.ac.jp, ‡masunaga@is.ocha.ac.jp

あらまし　従来, ムービングオブジェクトデータのような時系列データの類似検索に関しては, データを例えばフーリエ変換し周波数ドメインに写像して, そこで R-木や K-D木で索引付けする方法が用いられている．しかしながら，変換後の空間は多次元空間となり，データベース管理システムのアクセス法として広く用いられている B-木や B^+-木を使えないという問題点があった．そこで本稿では，ムービングオブジェクトデータの索引を１次元のスカラー値で与える方法を考案した．この方法では，ムービングオブジェクトデータベースに格納されている適当なムービングオブジェクトの動きを基準データとして選定し，その他のムービングオブジェクトデータにはそれと基準データとの相違度(＝類似度というも同じ)を計算して, 得られた相違度をそのデータの索引として付与することによりデータを構造化する．このとき少なくとも２つのことが問題となる．一つはこの索引付けは距離保存ではないので，この索引付けで近隣とされる２つのムービングオブジェクトデータ同士が類似しているとは必ずしもいえないこと，もう一つは基準データをどのように選定すると類似検索に要する時間を短縮できるかという問題である．本稿ではこれらを論じる．

キーワード　ムービングオブジェクト，時系列データ，類似検索，相違度，基準データ

A Similarity Search Method of Moving Object Databases based on Dissimilarity with respect to a Reference Data

Miya SAWAI[†]　Yumiko KITAHARA[†]　and　Yoshifumi MASUNAGA[‡]

† Graduate School of Humanities and Sciences, Ochanomizu University
2-1-1 Otsuka, Bunkyou-ku, Tokyo, 112-8610 Japan

‡ Department of Information Science, Faculty of Science, Ochanomizu University
2-1-1 Otsuka, Bunkyou-ku, Tokyo, 112-8610 Japan

E-mail: †{miya, yumiko}@dblab.is.ocha.ac.jp, ‡masunaga@is.ocha.ac.jp

Abstract　R-trees and K-D-tress are often used to index time series data such as moving objects data for efficient similarity search which are constructed by translating the data into a frequency domain using a feature extraction function such as Fourier transformation. However, there is a problem in that one-dimensional index structures such as B-tree and B^+-tree, which are widely used in traditional database management system, cannot be used because the frequency domain is multi-dimensional. In order to resolve this problem, we propose an indexing method of moving object data whose index values are scalar values. This method need to select one moving object stored in the moving object database to be a reference data, and the dissimilarities to all other moving object data are calculated so that the dissimilarity

measures are attached to them as indices. Two problems happen when we employee this method: First, since this indexing does not preserve distance, it is not always true that two moving object data are similar even though index values are in neighbor. Second, it is necessary to make clear what data should be selected as the reference data in order to shorten the average time for similar search. These are addressed in this paper.

Keyword Moving object, time series data, similarity search, dissimilarity, reference data

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(Global Positioning System) 等を用いて，ムービングオブジェクトの位置や姿勢など動きに関するさまざまなデータを計測するためのセンシング技術が発達してきている．これらは，２次元または３次元の座標空間におけるオブジェクトの位置を表す座標値や姿勢を表すパラメタ値を一定時間間隔毎に取得するためのものである．これに伴い，このような計測データをデータベースに蓄え，動きに関する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次元空間を動き回るオブジェクトの動きに関するデータを時系列データとして蓄え，様々な問合せを可能に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いる．既に，動きに関する基本データとしてオブジェクトの位置・向き・傾きデータを考慮に入れ，これらのデータを時刻データと共に統合的に扱うことのできるムービングオブジェクトデータモデルと，このモデルに基づいた問合せ言語である MOQL を提案している[1]. また，動きに関する類似検索機能の実現に向けて，ムービングオブジェクトデータに対する類似性を体系的に定義している[2]. このとき，動きを言葉で表すことが困難な場合もあるので Query-By-Example の発想に基づき，実際に所望の動きをモーションキャプチャシステムを利用して計測して，取得したデータと類似したデータをデータベースから検索できる機能を実現している[3]. さらに, 加速度を考慮することにより, 同じ軌跡を描く動きであっても移動速度の変化パターンが異なる動きの違いも認識することのできる類似性を定義することができるようになった[4].

しかしながら，類似検索は問合せの動きと格納されているデータとを直接総当りでチェックしてきたので，格納されているデータ量が多くなると検索に時間がかかることが懸念され，効果的なムービングオブジェクトデータの構成法を導入することが必要となっていた．また，オブジェクトの効率の良い検索法を研究することが必要である．言うまでもなく，類似検索の問題に対してはこれまでに多くの研究がなされてきているが, 大別すると (1) 測度法と (2) 変換法に分かれる．測度法は問合せデータと格納データの相違度を直接あるいは間接に計算する．（ここに, 間接とは例えばムービングオブジェクトデータをフーリエ変換して変換先の周波数ドメインで相違度を計算することを指す.）変換法では，問合せデータと格納データを一致させるに要する変換の工数で決める方法である．一般に，測度法は様々な概念を数学的に展開できる特性があり数多く研究されてきた．最も典型的なアプローチは，例えば時系列データをフーリエ変換し，変換後の主要項の係数が成す一般にｎ次元の周波数ドメインに写像して，R-木や K-D 木で索引付けする方法である[5,6]．しかしながら，変換後の空間は多次元空間となり，現在データベース管理システムのアクセス法として多用されている B-木や B^+-木を使えないという問題点があった．

そこで本稿では，ムービングオブジェクトデータの索引を１次元のスカラー値で与える方法を考案した．この方法ではムービングオブジェクトデータベースに格納されている適当なムービングオブジェクトの動きを基準データとし，その他のムービングオブジェクトデータにはそれと基準データとの相違度(＝類似度というも同じ)を計算して，得られた相違度をそのデータの索引として付与することによりデータを構造化する．このとき少なくとも２つのことが問

題となる．一つはこの索引付けは距離保存ではないので，この索引付けで近隣にある２つのムービングオブジェクトデータ同士が類似しているとは必ずしもいえないこと，もう一つは基準データをどのように選定すると類似検索に要する時間を短縮できるかという問題である．本稿ではこれらを論じる．

2. ムービングオブジェクトデータベースシステムの構築

2.1. システムの概要

本節では，我々がこれまで構築してきたムービングオブジェクトデータベースシステムの概略を示す．ムービングオブジェクトデータの計測には，センシング装置として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)を使用している．具体的には，円軌道を走行する模型電車の動きや卓球のラケットのスイングをカラーボールマーカ３点をつけて３次元的に計測している．

図１にシステムの概要を示す．格納インタフェースでは，モーションキャプチャリングシステムで計測して取得したオブジェクトに関するデータとシーンに関するデータ，計測データのファイルの所在等の入力をユーザから受ける機能を提供する．問合せインタフェースでは，ユーザからの Query-By-Example に基づく問合せを受付ける機能や，検索結果の表示機能を提供する．本システムの類似検索機能は２つのデータ間の相違度を 2.3 節で述べる距離関数によって測るという方法をとっている．これまでに，位置データに加えて，それから計算される速度や加速度に基づく相違度の計算や，ユークリッド距離を Time Warping 距離に置き換えた計算などの試みも行っている．

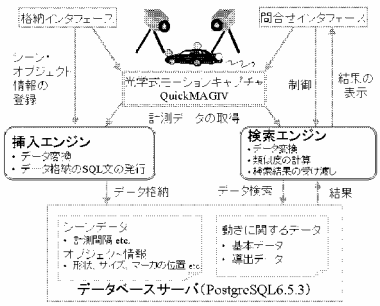

図 1: システム概要
Fig.1: A System Overview

なお，本システムで使用しているモーションキャプチャシステム QuickMAG IV では，動きを計測する際に予め計測時間を設定しておくため，必要なデータは取得したデータの一部であることが多い．それゆえデータの一部を切り出す必要が発生する．しかしながら計測データは図２のように計測点の一定時間間隔毎の座標列であるためそこから所望の一部を特定することは困難である．そこで我々は，計測シーンの画像を取得し，その画像を見ながら所望の動きの区間を指定すると，それに従って計測データを編集するような機能及びインタフェースを開発している[3].

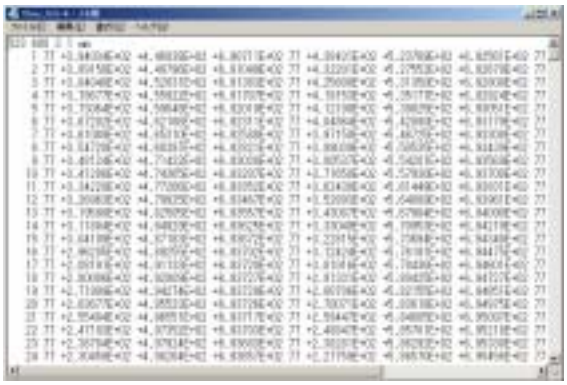

図 2: 計測データ
Fig.2: Measured Moving Object Data

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・速度・加速度，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 $\vec{M}=(\vec{m}_0,\vec{m}_1,\cdots,\vec{m}_m)$ で表す．このとき $\vec{m}_i=(\vec{p}_i,\vec{v}_i,\vec{a}_i,t_i)$ である．各要素はオブジェクトの位置ベクトル $\vec{p}_i=(x_i,y_i,z_i)$，速度ベクトル $\vec{v}_i=(vx_i,vy_i,vz_i)$，加速度ベクトル $\vec{a}_i=(ax_i,ay_i,az_i)$ を表す．また，定義から，$(t_e-t_s)\times f=m$ である．

2.3. 動きの同一性と基本類似性

動きの類似性を定義する前に，まずはその基本として動きの同一性の定義を行う．

定義 2.1：(動きの同一性)

動き $\vec{M}$，$\vec{M}'$ 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. $(\forall i)\ \vec{p}_i=\vec{p}_i'$

続いて，動きの基本類似性の定義を行う．

定義 2.2：(動きの基本類似性)

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. $D_p(\vec{M},\vec{M}')=\dfrac{\sqrt{\sum_{i=0}^{n} d_p(\vec{m}_i,\vec{m}_i')}}{n}\le\varepsilon$

但し $d_p(\vec{m}_i,\vec{m}_i')=|\vec{p}_i-\vec{p}_i'|^2$

なお，定義から分かるように，本稿では同一性や基本類似性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に対して定義されていることに注意する．一般に長さが異なる2つのムービングオブジェクトの動きに関する同一性や類似性を定義できるが，本稿はそれを論じるのが主題ではないので2つのムービングオブジェクトの開始時刻と終了時刻は同じであると仮定して議論を進めている．

3. ムービングオブジェクトデータの組織化

さて，問合せを効率よく処理するためには，ムービングオブジェクトデータに索引をつけるなど，それをうまく組織化をしておくことが必要である．特に，類似検索は点問合せ(point query)や範囲問合せ(range query)と並んで重要な問合せであるので，本稿ではムービングオブジェクトデータの類似検索を念頭においてデータ組織化を議論する．まず，従来よく研究されてきた特徴抽出関数による索引付けに言及して，続いて我々が新しく提案する基準データとの相違度に基づく索引付け法を述べる．

3.1. 特徴抽出関数による索引付け

時系列データのデータ構造と関わる類似検索の方法として，特徴抽出関数を用いて次元を圧縮し，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知られている[5]．このとき索引には，時系列データを特徴抽出関数により別空間にマッピングした値が使われる．特徴抽出関数として，主に離散フーリエ変換 (DFT: Discrete Fourier Transform) や離散ウェーブレット変換 (DWT: Discrete Wavelet Transform) などが用いられる．これらの特徴抽出関数は，時系列データを時空間ドメイン上での距離関係を保存しながら周波数ドメインにマッピングするために使われている．この変換により生成される周波数係数の先頭または末尾のいくつかがR-木やK-D木のような多次元索引構造を通して索引として使われる．このような索引構造を使用した上で，様々な類似性に対応した問合せ処理が数多く研究されている（例：[6]）．

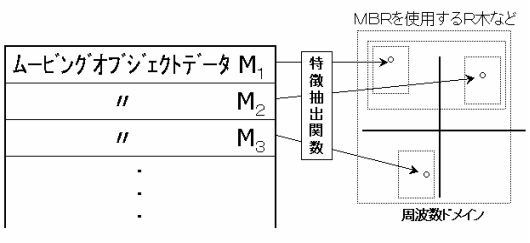

図 3: 特徴抽出関数による索引付け
Fig.3: Indexing using Feature Extraction Function

3.2. 基準データとの相違度に基づく索引付け

さて，上記の特徴抽出関数による変換は一般に多次元の索引構造を必要とし，これまでデータベースシステムで用いられてきた１次元の索引付けの構造である B-木や B^+-木を利用することができないという問題点があった．そこで我々は，格納されているムービングオブジェクトデータの一つを適当に選び，それを基準データとして，“基準データとの相違度”を計算することにより，スカラー値による新たな索引付け方法を提案する．その相違度は，定義 2.2 で示したユークリッド距離関数によって測ることとする．

定義 3.1：(基準データとの相違度)

基準データを $\vec{R}$ として，$\vec{R}$ と $\vec{M}$ の相違度を以下のように定義する．

$$D_p\left(\vec{R},\vec{M}\right)=\sqrt{\sum_{i=0}^{n} d_p\left(\vec{r}_i,\vec{m}_i\right)}$$

以下に索引付けの手順を述べる．

1. 動きのデータを１つ選出し，それを基準データとして，現在格納されている全てのデータとの相違度を計算する．（この基準データの選出方法については 3.6 節で検討する．）
2. データベースに格納されるデータに索引としてこの相違度を付加する．
3. 付加された相違度をもとに，B^+-木などのアクセス法を用いて索引付けを行う．

しかしこの基準データとの相違度は相対的なものであるため，これを絶対的な尺度として使用し，索引付け及びデータの組織化を行うと，実データ同士の距離関係が正確には保存されないため，過少誤認 (false dismissal) や，過多誤認 (false alarm) が発生する恐れがある．以下で，過少誤認が存在し得ないことを示し，過多誤認の排除のための類似検索方法を提案する．

3.3. 過少誤認の無発生

過少誤認とは，本来類似検索により類似しているとみなされるべきデータをとりこぼしてしまうことである．例えば図４のような２次元のデータがあり，M と ε -類似しているものを検索することを考え，過小誤認が発生しないことを示す．これには，実際に $\vec{M}$ と ε -類似している全ての $\vec{M}'$ に関して，以下の式が成り立てばよい．

$$D_p\left(\vec{M},\vec{M}'\right)\le\varepsilon$$

$$\Rightarrow\quad \left|D_p\left(\vec{R},\vec{M}\right)-D_p\left(\vec{R},\vec{M}'\right)\right|\le\varepsilon$$

しかし，定義 3. 1 で相違度計算に使用したユークリッド距離は三角不等式を満たすため，上の式は成立する．よって，過少誤認は発生しない．

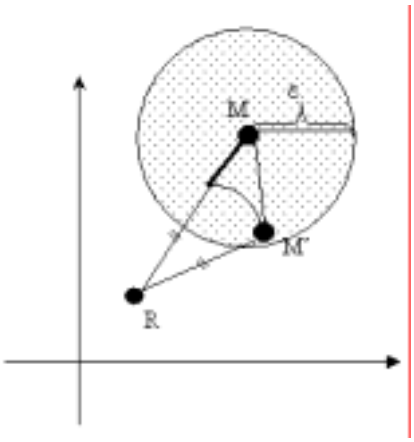

図 4: 過少誤認の発生しないこと
Fig.4: No Occurrence of False Dismissal

3.4. 過多誤認の発生

過多誤認とは，本来類似していないデータを類似しているとみなしてしまうことである．例えば図５のような２次元データがあり，M と ε -類似しているものを検索する

とき，$D_p\left(\vec{M},\vec{M}'\right)\geq\varepsilon$であるにも関わらず

$$\left|D_p\left(\vec{R},\vec{M}\right)-D_p\left(\vec{R},\vec{M}'\right)\right|=0\leq\varepsilon$$

であることから$\vec{M}'$が検索されてしまう．

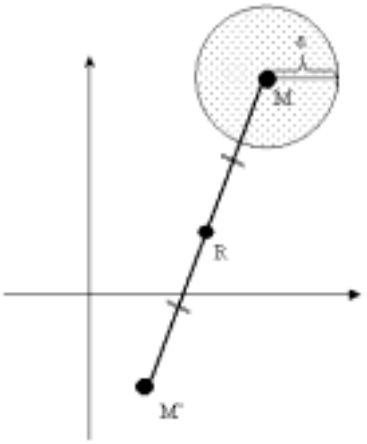

図 5: 過多誤認の発生

Fig.5: Occurrence of False Alarm

3.5. 過多誤認排除のための類似検索方法

過多誤認を排除するため，以下のような類似検索法を提案する．これまでと同様に基準データを$\vec{R}$，また問合せデータを$\vec{Q}$とする．

【第１段階】索引として付加した基準データとの相違度のみに着目し，

$$D_p\left(\vec{R},\vec{M}\right)\leq D_p\left(\vec{R},\vec{Q}\right)\pm\varepsilon$$

となる$\vec{M}$を全て検索結果の候補とする．

【第２段階】第１段階で候補として挙げられた全てのデータと，問合せデータ$\vec{Q}$との実際のユークリッド距離を計算し，

$$D_p\left(\vec{Q},\vec{M}'\right)\leq\varepsilon$$

となる$\vec{M}'$を，最終的に問合せデータ$\vec{Q}$とε-類似している検索結果とする．

3.6. 基準データの選定方法

以上に述べた手法により類似検索を行う場合，基準データの選定方法が問題となってくる．つまり,基準データとして何を選定するかによって，前節で述べた過多誤認排除のために検証すべきデータの数が多くなったり，あるいは少なくてすむことが考えられる．

そこで，シミュレーションによりその最適解を探った．そのために，第一次近似としてムービングオブジェクトデータは図６に示すように１次元の標準正規分布(分散σ^2は１，平均μは０)に従って分布するものであるとし，図の網掛け部分に相当する第１段階で求められる検索候補数を計算した．ここに許容誤差εは微少量の定数である．

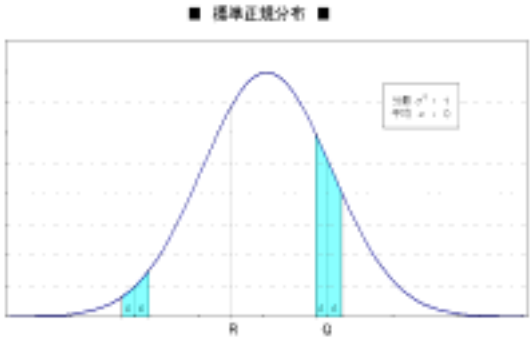

図 6: 標準正規分布

Fig.6: Standard Normal Distribution

さて，一般に正規分布の式は以下の様に表される．

$$f\left(x\middle|\mu,\sigma^2\right)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

基準データＲの座標値をｒ，問合せデータＱのＲとの相違度をｑとする．第１段階で決まる検索候補数（図６の網掛け部分）Ｋは全データ数をｎとすると，以下の式で表すことができる．

$$K=n\int_{r+q-\varepsilon}^{r+q+\varepsilon}f\left(x\middle|\mu,\sigma^2\right)dx+n\int_{r-q-\varepsilon}^{r-q+\varepsilon}f\left(x\middle|\mu,\sigma^2\right)dx$$

このとき，許容誤差εは充分に小さいため，検索候補数Ｋは次の式で近似することができる．

$$K\cong 2\varepsilon n\,f\left(r+q\middle|\mu,\sigma^2\right)+2\varepsilon n\,f\left(r-q\middle|\mu,\sigma^2\right)$$

$$=\frac{2\varepsilon n}{\sqrt{2\pi\sigma^2}}\left\{\exp\left(-\frac{(r+q-\mu)^2}{2\sigma^2}\right)+\exp\left(-\frac{(r-q-\mu)^2}{2\sigma^2}\right)\right\}$$

以上の式を用いて，ｒ及びｑを変数とした

シミュレーションを行った．シミュレーション実行時の各定数の値は以下のとおりに設定した．

平均　：$\mu = 0$　　分散　　：$\sigma^2 = 1$

データ数：$n = 1000$　許容誤差　：$\varepsilon = 0.1$

図７に，シミュレーション結果を表示する．これは横軸にｒの座標値を，縦軸に検索結果候補数Ｋをとったもので，問合せデータＱの基準データＲとの相違度ｑを変化させて曲線を描画している．

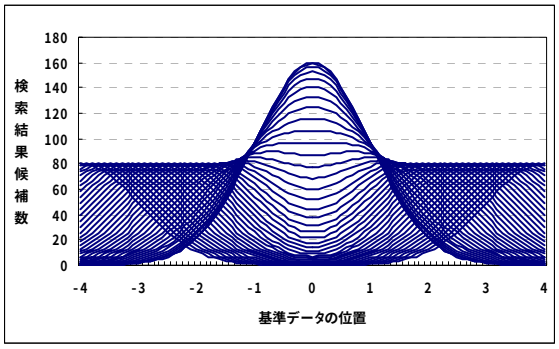

図 7: シミュレーション結果
Fig. 7 A Simulation Result

図７から，問合せデータＱと基準データＲとの相違度ｑがどのような値をとっても，基準データＲが平均（μ=0）からある程度離れると，検索結果候補数Ｋは一定の値以下に抑えられることがわかる．この場合では，基準データＲを分布の平均値から 1.8σ 以上離れた位置にとると，検索結果候補数は８０以下に抑えられていることが分かる．

なお，最近 NB-木が報告されている[7]．そこでは相違度の替わりに Euclidean Norm が用いられているが，それを我々のケースに当てはめると基準データを原点にとった相違度と等しいものである．しかし，[7]ではデータの分布が正規分布とは仮定されていなく，今後より的確な比較を行う予定である．

4. まとめと今後の課題

現行のムービングオブジェクトデータベース内のデータに，基準データとの相違度に着目した索引を付けデータを構造化することを考え，基準データの選出方法ならびに索引を利用した一類似検索法を提案した．今後の課題として，以上の論理に基づいた実装を行い，類似検索時に必要とされる計算量などの検証や，以上の類似検索方法に最適なデータ構造の探求が挙げられる．

末筆ながら，ムービングオブジェクトデータベースに関する増永研究室諸氏の討論に感謝する．

文　献

[1] Ｙ．Masunaga, Ｎ．Ukai: “Toward a 3D Moving Object Data Model - A Preliminary Consideration -,” In Proceedings of the 1999 International Symposium on Database Applications in Non-Traditional Environments, Kyoto, pp. 306-316, 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] 澤井美弥, “ムービングオブジェクトデータベースにおける画像編集インタフェースの作成,” お茶の水女子大学卒業論文, 2002.

[4] 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.

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

[6] Davood Rafiei and Alberto O.Mendelzon: “Querying Time Series Data Based on Similarity,” IEEE Trans.on Knowledge and Data Engineering, 12(5), pp. 675-693, September/October 2000.

[7] M.J. Fonseca and J.A. Jorge: “Indexing High -Dimensional Data for Content-Based Retrieval in Large Databases,” In Proceedings of the 8th DASFAA, pp. 267-274, Kyoto, March 2003.