様式 C-19

科学研究費補助金研究成果報告書

平成　23 年　6 月　24 日現在

機関番号：34315
研究種目：基盤研究(C)
研究期間：2008～2010
課題番号：20500104
研究課題名（和文）　内部特徴による時間データ類似検索方式の研究

研究課題名（英文）　Research on Temporal Data Similarity Search with Internal Features

研究代表者
川越　恭二（KAWAGOE KYOJI）
立命館大学・情報理工学部・教授
研究者番号：４０２９８７２４

研究成果の概要（和文）：
本研究は，気象や医学等の分野で用いられる時系列データの分析の際に必要な類似検索を効率化する新たな方法の開発を目指す．これまでは時系列データにおける時刻ごとの値の差の大小により類似した時系列データを求める方法が代表的であった．しかし，時系列データが発生している対象物に内在する特徴の類似性考慮していないために，利用者から見れば得られた類似データが必ずしも類似しているとは言いがたい場合が存在していた．本研究では，対象物の内部的な特徴に着目する新たな方法を開発することで性能の向上および方法を支援する環境基盤を実現した．
研究成果の概要（英文）：
In this research, we aim to develop a new method of temporal data similarity search in order to improve performance of similar time series data search used in time series data analysis. The method we developped was introduced by internal features with which a time series data is produced. Through our research work, we realized performance improvement and some environment infrastructure to support the method.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2008 年度	1,100,000	330,000	1,430,000
2009 年度	1,200,000	360,000	1,560,000
2010 年度	1,000,000	300,000	1,300,000
総　計	3,300,000	990,000	4,290,000

研究分野：　情報技術
科研費の分科・細目：　情報学，メディア情報学・データベース
キーワード：　時間データ，時系列データ，類似検索，内部特徴，情報検索，マルチメディア

１．研究開始当初の背景

工業統計，商業統計，経済統計，国勢調査統計，気象情報，株式情報，交通情報など社会で利用されている様々な時間データに関する類似検索方法がこれまでに多数開発されている．これまでの類似検索方法には，DFT(離散フーリエ変換)，DWT(離散ウェーブレット変換)，APCA(適応区分定数近似)，PLA(区分線形近似), SAX(記号集約近似)などが提案され，さらに類似度（距離）定義として, ユークリッド距離, DTW(動的時間伸縮), 編集距離などが用いられている．

また，今後多くのセンサから発生し大量に格納される時間データを管理・処理・利用するであろうユビキタス社会において，予測，分類，分析等を効果的に行えるデータ処理の実現への期待は非常に大きいものがあると考える．
たとえば，医療分野では診断治療のために患者に接続された各種センサーから時間とともに変化する大量のデータが送られ，時々刻々変化する患者の状態をモニタリングすることができる．しかしこれまでの技術や方法を大量の時系列データの分析管理するに

は十分ではない状況が今後発生するものと予想される．さらに多種の混在するデータの分析を同時に行うことも重要であるため，新たな方法の開発が強く求められている．

２．研究の目的

本研究では，株価等の金融データ，シミュレーションによる科学技術データ，温度や気圧等の気象データなどの時間的に変化するデータである時間データの類似検索に関し，その適用能力の向上を目指す．すなわち，従来の類似性定義ではなく時間データが生成される内部構造表現に着目して記述した特徴の類似性を定義することで，特定の応用に依存せず広範囲の応用に適用可能なこれまでにない高い適用能力を持つ新たな類似検索手法の開発を目指す．

先にのべたように，時間データ類似検索方法でこれまで利用されていた類似度定義，逆にみれば距離定義はユークリッド距離や動的時間伸縮距離などが提案されている．また，時間データを直接検索するのではなく，時間データの特徴を抽出し特徴空間で類似性の判定を行う方法が存在する．特徴空間としては，離散フーリエ変換(DFT),離散ウェーブレット変換(DWT),区分集約近似(PAA),適応型区分定数近似(APCA),区分線形近似(PLA)などが使用されており，各々の特徴空間では特徴ベクトル間でのユークリッド距離が使用されている．しかし，このような従来手法には，応用適用能力,類似検索精度,類似検索処理効が重要であるが多くの問題を抱えている．本研究では，これらの問題のうち，最初の問題への解決を図ることを目的とする．応用領域ごとに適切な特徴空間を選択・記述することが，時間データの類似検索手法の利用の立場からは重要となる．本研究は，応用領域ごとに特徴空間を選択するのではなく，応用領域の時間データの性質を記述できるような仕組みを開発するものである．応用領域に依存した特徴的な性質を記述することが可能となれば，その記述内容を内部表現として使用し内部的な特徴空間を自動的に構築することが可能となり，対象とする応用領域に最も適した内部特徴空間で時間データの類似検索方式を使用することが可能となると考える．

３．研究の方法

本研究を実施するため，まず，内部特徴モデル表現の開発と内部特徴モデルを用いた類似検索方法の開発を行い，さらに本提案方法の特定応用分野への適用の研究項目を行うと同時に研究を支える環境基盤の構築を進める．まず，内部特徴モデル表現の開発には内部特徴モデルとして制約付線形の状態方程式の活用に取り組み，線形モデルで高範囲の応用分野に適用可能であることを確認する．同時に，性能向上のための索引構造を導出するとともにマルチメディア類似検索方法の基本性質と枠組みの構築を行う．提案した類似検索方法が広範囲の応用分野に適用できることを示すために複数の応用分野を設定する．具体的には，音楽情報，医療情報,Web の分野を設定し従来方法との比較実験を実施する．環境基盤については，ネットワーク環境，マルチメディア環境，検索環境の視点からの各々の基盤整備を進める．

４．研究成果

本研究により開発した研究成果について記述する．

(1) 内部特徴による時系列データ類似検索

これまでの時間データの類似検索は基本的に時間データの時系列曲線の幾何的類似性を基本として行われている．しかし，生命科学分野のような動的モデルを用いて分析をおこなう分野がある．この動的モデルを内部的なモデルとすることで，外部に現れる時間データはこの動的モデルの振る舞いを示すにすぎず，外部的振る舞いの類似性を判断するのは意味がない．したがって，時系列データの類似性よりも内部モデルの類似性を考慮すべきである．しかし，外部的振る舞いである時系列データから内部表現である動的モデルを同定するには多大な計算量と同定誤差が発生するために非常に作業的に困難なことである．また，動的モデルの構築自体が試行錯誤的な研究に繋がる作業であるため，あらかじめすべての時系列データから動的モデルを同定しておくことは現実的ではない．そこで，本研究では，ある時系列データと類似した動的モデルを持つ他の時系列データを求める問題を扱い，そのための効率的検索手法を提案した．Naive な方法は，与えられた質問時間データから時間空間内で離れた距離にある時間データをフィルタリングしておき，フィルタリング後のすべての時間データに対して内部モデルの同定を行う方法である．特徴空間内で質問に対応した内部モデルとの類似度を計算し，類似した内部モデルの時間データを出力する．これに対して，提案方法は，内部モデルを記述する特徴空間内で，質問時間データに対する内部モデルを中心点として，適当な方向にある複数の変動点を計算しその変動点の外部的振る舞いを計算する．その振る舞いを示す時間空間内の点からある距離範囲内にある時間データについてのみ，その内部モデルを同定し類似度を計算し類似した内部モデルの時間データを出力する．本提案方法を薬学や医療分野で用いられているコンパートメントシステムに適用し方法の有効性を調べた．コンパートメントシステムは制約つきの多次元

線形差分方程式で記述されるが，この方程式を用いて擬似時間データをランダムに発生したテストデータを用いた．評価実験の結果，同一再現率（66%）について，Naive 方法では 10%の適合率であったが，提案手法では 66%の結果を得た．さらに，ひとつの質問に必要な同定回数が全データ量の 7%であり，これだけの同定で類似した時間データを得ることができることを実験で示した．コンパートメントシステムへの応用以外にも，遺伝子やたんぱく質の相互作用の分析に用いら

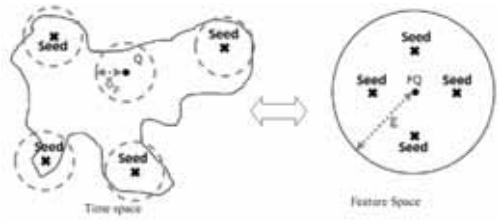

図 1 概要

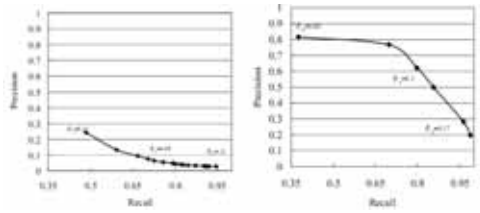

図 2 評価結果
(左:比較手法,右:提案手法)

れる制約つき多次元非線形微分方程式で記述される S-System への応用も可能であると考える．図 1 と図 2 に本研究成果の概要と評価実験の結果を示す．

(2) 時間単語による時間データ検索
時間的に変化する時間データの類似検索やクラスタリングなどでは，大量の高次元の時間データをより効率的に処理するために，低次元の特徴によって時間データを近似することが用いられている．代表的な近似手法には，時間データを周波数領域の特徴として表現するフーリエ変換法，時間データを Haar と呼ぶ関数で近似するウェーブレット法，時間データを時間区分的定数で近似する APCA（適応区分定数近似法），時間データを区分的線分で近似する PLA(区分的線形近似法)，時間データをその値によって文字で表現するシンボリック近似法などが提案されている．これらの近似方法は時間に沿った形で特徴化を行う点で共通であるが，雑音が多数存在する時間データや微妙に変化する類似した時間データの場合には，類似性が十分に機能せずに，よい検索精度が得られないという問題が存在していた．そこで，本研究では，時間データを時間単語と呼ぶ単語の集合で近似する新たな方法を提案する．時間データの時間単語のシーケンスでの近似ではなく，時間単語の集合での近似であるため，これまでの方法のような時間的変化のみを特徴に反映させる方法ではない．時間単語は，時間データから特徴点と呼ぶ時間データを特徴付ける複数点を検出し，その特徴点の特徴ベクトルを求める．すべての特徴点のベクトルの集合を求め，その集合のクラスタリングによって，複数のグループに分割する．分割されたグループの代表点（クラスター中心）をひとつの時間単語に対応づける．特徴点の特徴ベクトルから最も近い時間単語のベクトルですべての特徴点を表現する．時間データごとに時間データベクトルを構築し，この時間データベクトルで個々の時間データの近似を行う．ここで，時間データベクトルの個々の要素は，その時間データ内での要素に対応した時間単語の出現回数を値と設定する．本提案手法は，文書検索で用いられる文書を索引語ベクトルで近似するベクトル空間モデルを基本的手法とする時間データ検索方法としては非常に新しい方法である．本提案の応用として，音が時間的変化するデータと考えられる音楽情報を取り上げた．まず，ハミングによる楽曲検索において利用者が入力したハミング内に含まれる誤りを検出し，キーワード検索で用いられる「もしかして」機能と同等の入力ハミングの時間単語スペル修正機能を時間単語抽出および既存文書処理手法の適用により実現した．入力時間単語スペル修正機能の評価実験の結果，20 回のハミングによる楽曲検索作業において，うち 3 回の作業で本スペル修正機能を利用して利用者の望む楽曲を取り出すことが可能となった．さらに，楽曲類似検索に対しても時間単語の適用を図った．75 曲の楽曲を対象として評価実験を行った結果，ポップス風の音楽については 11 点平均適合率が 0.59 というよい値を得た．

(3) Web 編集履歴からの信頼度算出
Web 上には様々な情報が時系列で編集されているが，その情報は信頼できない情報も含まれている．これらの情報を自動的に特定するために，これまではシステムの利用者によって情報に対して信頼度を付与する方法が利用されてきた．しかし，信頼度の精度が非常に低いと同時に高度な知識が必要であった．そこで，時系列データへの信頼度付与のために，利用者による明示的判定ではなく，暗黙的判定方法を開発した．すなわち時系列データから利用者からの入力を必要とせずに，利用者の行動から自動的に特徴抽出し，

信頼度算出する方法を提案した．Wikipediaを信頼度算出の対象とし，多くの編集を経て残存している記述部分を信頼度が高い部分であると特定した．特定方法として，残存率が高い記述を特定する方法を利用するとともにした．さらに信頼度の高い記述を行うことが多い著者を選び出し，その著者による記述は信頼度を高く設定した．この結果，記述残存率が算出できない記事に対して，高い精度で信頼度の判定ができた．日本語Wikipediaを利用した評価実験により，抽出記事100件のうち，61件が信頼度が高いと判定できる記事として得た．

(4)マルチメディアの類似検索
時系列データの内部特徴を用いることで類似検索の効率化を目指す方法の開発の一環でマルチメディアデータとテキスト文書の類似性に着目し，テキスト文書処理技術を用いたマルチメディアデータ類似検索の研究を行った．マルチメディアデータの中でも画像データを中心に画像データ内に含まれる内部オブジェクトの特徴を用いた類似検索方法を開発した．画像内に占める内部オブジェクト特徴量を用いて画像(文書)ベクトルを求め，通常のテキスト文書検索で利用される特徴ベクトル間類似度を画像検索に適用する方法を実現した．本提案方法はキーワード検索で得た画像集合の中から選択した画像を問合せ画像として入力する際にその画像と類似した画像を検索するのに用いることができる．310 枚の画像に対して評価実験を行った結果，画像内オブジェクトを考慮しない場合に比べて 11 点平均適合率が約 10%向上した．

(5)その他研究成果および研究環境基盤整備
発育時系列データやイベントデータ，ストリームデータへの適用を行うことで，応用可能性の拡大を行った．加えて，前述のような環境基盤構築として，アドホックネットワークにおける各種情報探索手法の提案と評価やWeb サーバの新たなキャッシュ方法の開発，複数検索エンジンの適切な選択方法の開発と評価等を実施した．これらの研究成果は国内外の学会で発表を行った．

５．主な発表論文等
〔雑誌論文〕（計３件）
① 著者名：大野和久,鈴木優,川越恭二,論文標題：楽曲全体における特徴量の傾向に基づいた類似検索手法,査読：有,雑誌名：日本データベース学会論文誌,巻：7,発行年：2008,ページ：233-238
② 著者名：辻健太，川越恭二,論文標題：画像内オブジェクトの相対的特徴に基づく類似検索手法,査読：有,雑誌名：日本データベース学会論文誌,巻：9,発行年：2010,ページ：53-58
③ 著者名：鈴木優，吉川正俊,論文標題：Wikipedia におけるキーパーソン抽出による信頼度算出精度および速度の改善,査読：有,雑誌名：情報処理学会論文誌：データベース,巻：3,発行年：2010,ページ：20-32

〔学会発表〕（計 24 件）
①発表者名：Kazuhisa Ono，Yu Suzuki and Kyoji Kawagoe，発表標題：A Music Retrieval Method based on Distribution of Feature Segments,学会名等：MIPR2008,発表年月日：2008年12月16日,発表場所：バークレー(アメリカ)
②発表者名：清田寛信,鈴木優，川越恭二,発表標題：P2Pでの多次元検索のための熱度を考慮した索引分散配置手法,学会名等：電子情報通信学会情報ネットワーク研究,発表年月日：2009 年 1 月 23 日,発表場所：名古屋工業大学（愛知県）
③発表者名：鈴木優，大野和久，光川正弘，川越恭二,発表標題：テキスト情報検索手法を利用したマルチメディアデータ検索手法,学会名等：電子情報通信学会情報ネットワーク研究,発表年月日：2009年3月8日,発表場所：掛川市（愛知県）
④発表者名：水野美沙，鈴木　優，川越恭二，発表標題：楽曲演奏学習のための時間的揺らぎの特徴抽出,学会名等：情報処理学会第 71 回全国大会,発表年月日：2009年 3 月 11 日,発表場所：立命館大学（滋賀県）
⑤発表者名：Salma Nasrin　Kyoji Kawagoe，発表標題：SG2 Index Structure for Super Peer Architecture based Databases,学会名等：マルチメディア，分散，協調とモバイル(DICOMO2009)シンポジウム,発表年月日：2009 年 7 月 9 日,発表場所：杉乃井ホテル（大分県）
⑥発表者名：辻健太，川越恭二,発表標題：画像内オブジェクトの相対的特徴を用いた類似画像検索手法,学会名等：電子情報通信学会 Web インテリジェンスとインタラクション研究会,発表年月日：2009年10 月 22 日,発表場所：学術総合センター(東京都)
⑦発表者名：鈴木督史，川越恭二,発表標題：スペル修正技術を用いた楽曲検索システム,学会名等：情報処理学会音楽情報科学研究会,発表年月日：2009 年 11 月 3 日,発表場所：東京大学(東京都)
⑧発表者名：鈴木優，吉川正俊,発表標題：Wikipedia におけるキーパーソン抽出による信頼度算出精度および速度の改善,

学会名等: 人工知能学会第21回セマンティックウェブとオントロジー研究会,発表年月日: 2009年11月22日,発表場所: 東京大学(東京都)

⑨発表者名: Kodai Mizuno and Kyoji Kawagoe, 発表標題: Dynamic Selection Method of The Best Search Engine for A User's Query,学会名等: IUCS 2009,発表年月日: 2009年12月4日,発表場所: 日本科学未来館(東京都)

⑩発表者名: 吉本和紀, 鈴木優, 吉川正俊, 発表標題: マイクロブログにおける他者への影響を考慮した投稿者の重要度推定手法,学会名等: 第2回データ工学と情報マネジメントに関するフォーラム(DEIM 2010),発表年月日: 2010年2月28日,発表場所: 淡路夢舞台国際会議場(兵庫県)

⑪発表者名: 辻健太, 川越恭二,発表標題: 画像内オブジェクトの相対的特徴を用いた類似画像検索,学会名等: 第2回データ工学と情報マネジメントに関するフォーラム(DEIM 2010),発表年月日: 2010年3月1日,発表場所: 淡路夢舞台国際会議場(兵庫県)

⑫発表者名: 山本めぐみ, 川越恭二,発表標題: 発育時系列データの類似検索による育児ブログ推薦システム,学会名等: 情報処理学会第72回全国大会,発表年月日: 2010年3月10日,発表場所: 東京大学(東京都)

⑬発表者名: Salma Nasrin, Kyoji Kawagoe, 発表標題: SG2: A Novel Index Structure for Efficient Data Management in Super-Peer Architecture, 学会名等: IEEE International Conference on Ubiquitous and Future Networks,発表年月日: 2010年6月18日,発表場所: 済州島（韓国）

⑭発表者名: Kyoji Kawagoe, Thomas Bernecker, Hans-Peter Kriegel et al, 発表標題: Similarity Search in Time Series of Dynamical Model-based Systems, 学会名等: International Workshop on Database Technology on Life Science and Medicine (DBLM2010),発表年月日: 2010年8月31日,発表場所: ビルバオ(スペイン)

⑮発表者名: 山本めぐみ, 川越恭二,発表標題: 発育時制データの類似検索による育児ブログ推薦システム,学会名等: ヒューマンインタフェースシンポジウム 2010,発表年月日: 2010年9月10日,発表場所: 立命館大学（滋賀県）

⑯発表者名: Masafumi Suzuki and Kyoji Kawagoe,発表標題: A Music Retrieval System with Spelling Correction Technique, 学会名等: 10th International Society for Music Information Retrieval Conference (ISMIR 2009),発表年月日: 2010年10月30日,発表場所: 神戸ポートピアホテル(兵庫県)

⑰発表者名: Yu Suzuki,発表標題: An Assessment of Credibility for Message Streams on Microblogs,学会名等: SMDMS2010,発表年月日: 2010年11月5日,発表場所: 福岡工業大学(福岡県)

⑱発表者名: 見市高一, 川越恭二,発表標題: QAサイトにおけるS項目による回答検索システム,学会名等: 情報処理学会第151回データベースシステム研究会,発表年月日: 2010年11月13日,発表場所: 早稲田大学(東京都)

⑲発表者名: Akiyo Nadamoto, Yu Suzuki, Takeshi Abekawa,発表標題: Gist of a Thread in Social Network Services based on Credibility of Wikipedia,学会名等: HICSS 2011,発表年月日: 2011年1月8日,発表場所: ハワイ(アメリカ)

⑳発表者名: 鈴木督史,黄宏軒,川越恭二,発表標題: 旋律からの単語抽出による文書モデルベースの旋律検索,学会名等: 情報処理学会第89回音楽情報科学研究会,発表年月日: 2011年2月12日,発表場所: 福岡工業大学(福岡県)

㉑発表者名: 鈴木優, 吉川正俊,発表標題: Credibility Rank: 編集履歴と著者情報を用いたWikipediaの記事信頼度算出手法,学会名等: 第三回データ工学と情報マネジメントに関するフォーラム,発表年月日: 2011年2月27日,発表場所: ラフォーレ修善寺（静岡県）

㉒発表者名: 海江田隆博, 黄宏軒, 川越恭二,発表標題: contextHashtagによるTwitterユーザ向けイベント推薦システム,学会名等: 第3回データ工学と情報マネジメントに関するフォーラム,発表年月日: 2011年2月28日,発表場所: ラフォーレ修善寺（静岡県）

㉓発表者名: 嘉村亘太, 川村陸, 黄宏軒, 川越恭二,発表標題: ストリーム管理機構を用いたコミュニケーションストリームポータルC-SPOTの提案と試作,学会名等: 情報処理学会第73回全国大会,発表年月日: 2011年3月2日,発表場所: 東京工業大学（東京都）

㉔発表者名: スアンフイドー, 黄宏軒, 川越恭二,発表標題: リバースプロキシにおけるページランクを考慮したキャッシュ置換方式,学会名等: 情報処理学会第73回全国大会,発表年月日: 2011年3月2日,発表場所: 東京工業大学（東京都）

６．研究組織

(1)研究代表者

川越　恭二（KAWAGOE KYOJI）
立命館大学・情報理工学部・教授
研究者番号：４０２９８７２４

(2)研究分担者

鈴木　優（SUZUKI YU）
名古屋大学・情報基盤センター・研究員
研究者番号：４０３８８１１１