様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

# 科学研究費助成事業　　研究成果報告書

平成　２９　年　　６　月　１９　日現在

機関番号：　２３８０３
研究種目：　基盤研究(C)（一般）
研究期間：　2014～2016
課題番号：　２６３３０１３８

研究課題名（和文）一般化ピボットでのデータ構造化技術による類似検索の高速化

研究課題名（英文）Speeding up the similarity search by generalized pivots

研究代表者
池田　哲夫（IKEDA, Tetsuo）

静岡県立大学・経営情報学部・教授

研究者番号：６０３６３７２７

交付決定額（研究期間全体）：（直接経費）　　3,600,000円

研究成果の概要（和文）：本研究の目的は、画像や映像などのマルチメディアデータの効率的な類似検索方法を開発することである。1.ピボットベクトルの各要素を独立に高速に改善できることを特徴とするピボット生成方法を提案し、L1距離を用いたレンジ検索性能の向上を実現した。2.データ可視化方法として、属性分布のZスコアに基づき部分データ集合の特徴をアノテーションとして抽出する方法を提案し、有効性を確認した。3.クラスタリング方法として、前処理でピボットを構築してからクラスタ生成を行うことを特徴とする方法を提案し、有効性を確認した。
何れも類似検索及び関連する大量データ活用技術に関して新規性を有する技術であり、有意義な成果である。

研究成果の概要（英文）：The purpose of our research is to develop efficient similarity search methods of multimedia data such as picture images and movies. (1)We proposed a pivot generation method characterized in that each pivot vector element can be improved independently and at high speed, and confirmed its effectiveness using L1 distance. (2)As a data visualization method, we proposed a method to extract features of partial data sets as annotations based on Z scores of attribute distributions, and confirmed its effectiveness. (3)As a clustering method, we proposed a method characterized by constructing pivots in the preprocessing process and then generating clusters, and confirmed its effectiveness.
They are technologies having originality with respect to similarity search and related large amount data utilization.

研究分野：　データ工学

キーワード：　情報検索　類似検索　可視化　クラスタリング

## １．研究開始当初の背景

計算機と通信技術の飛躍的な向上に伴い、大規模なマルチメディアデータが蓄積され、今後さらに規模が増大する状況にあり、蓄積されたデータを有効利用するための優れた技術に対するニーズはますます高まっている。

本研究では、このニーズに応えるコア技術の一つとして、音声、画像、文書、蛋白質配列、DNA 配列など多様な要素から構成されるマルチメディアのオブジェクト集合から、ユーザがクエリとして与えるオブジェクト q と類似するオブジェクトを効率良く探索する類似検索問題に焦点を当てる。

本研究は、本研究参画者らの平成 23 年度からの基盤研究(C)での最重要研究成果「一般化ピボット構築技術」を土台に、より優れた類似検索性能を実現するための類似検索技術の確立を目的に展開する。

## ２．研究の目的

計算機と通信技術の飛躍的な向上に伴い、大規模なマルチメディアデータが蓄積され、今後さらに規模が増大する状況にあり、蓄積されたデータを有効利用するための優れた技術に対するニーズはますます高まっている。本研究では、このニーズに応えるコア技術の一つとして、音声、画像、文書、蛋白質配列、DNA 配列など多様な要素から構成されるマルチメディアの N 個のオブジェクト集合 X = {x, y, …} から、ユーザがクエリとして与えるオブジェクト q と最類似するオブジェクト x^* を効率良く探索する最類似検索など各種類似検索問題に焦点を当てる。

本研究課題では、ピボット集合の利用による縮小埋め込み（contractive mapping）に基づく類似検索の枠組みを採用する。縮小埋め込みとは、任意のオブジェクトペア x と y に対し、元空間での距離 d(x, y)、埋め込み関数 F(x)、及び、埋め込み後の距離 δ(F(x), F(y)) を適切に定義するとき、不等式 δ(F(x), F(y))≦ d(x, y) が成立することを言う。ピボット集合から適切な埋め込み関数 F(x) を構成すれば、最類似検索において、あるオブジェクト z ∈ X とクエリ q との元空間での距離 d(z, q) に対し、集合 S = {x ∈ X | d(z, q) < δ(F(x), F(q))} を求めることで、元空間での距離を計算することなく S に属すオブジェクトは最類似とならないことより枝刈りできる。

本研究参画者らの基盤研究(C)(23500128)(平成 23-25 年度)での重要な研究成果として、オブジェクト集合に限定された範囲でピボットを選択する既存技術とは異なり、機械学習アプローチに基づく反復改善法により、マンハッタン距離に基づくユークリッド空間の任意の点として適切な一般化ピボット（generalized pivot）を構築する新技術を確立し、極めて有望なコア技術となり得ることを実証した。

本研究課題では、一般化ピボット構築技術を土台に、より優れた最類似探索性能を実現する類似検索技術の構築を目的とする。

## ３．研究の方法

本研究の当初の目的は２に述べたとおりであるが、実際の研究の遂行においては、類似検索に関連する大量データ活用技術の研究も実施した。

具体的な研究方法を以下に説明する。

(1)ピボット生成方法の改良

一般化ピボット構築技術を土台により優れたピボット生成方法を構築する。

(2)大量データ可視化方法の構築

類似検索においては大量データを分かり易く可視化する方法の確立も重要な課題である。一般化ピボット構築技術をベースに大量データ可視化方法を構築する。

(3)クラスタリング方法の構築

大量データの活用の観点からは類似検索と並んで重要な技術にクラスタリング技術がある。一般化ピボット構築技術を活用して高速なクラスタリング方法を構築する。

## ４．研究成果

(1) ピボット生成方法の改良：ピボットベクトルの各要素を独立に高速に改善できることを特徴とするピボット生成方法の提案、およびクエリ分布あるいはクラスを考慮したピボット生成方法の提案を行い、いずれも類似検索性能向上に有効であることを確認した。

①PGM (Pivot Generation based on Manhattan distance) 法の提案

ピボットベクトルの各要素を独立に高速に改善できることを特徴とする新たなピボット生成方法（PGM 法と呼ぶ）を提案し、L1 距離（マンハッタン距離）を用いたレンジ検索技法の検索性能の向上を図った[学会発表⑪]。手書き文字 DB,新聞記事 DB、本のレビューDB を対象にして性能評価を行った。提案方法が、目的関数の値、ピボット生成・選択の計算時間、レンジ検索性能、および可視化結果のわかりやすさの点において、従来の代表的な方法（BNC 法）よりも優れていることを確認した。さらに、提案方法は、オブジェクトが低次元空間に密に存在する場合よりも、高次元空間に疎に存在する場合において、BNC 法よりも大きく優れることを示した。

表 1 にレンジ検索の性能比較結果を示す。PGM は提案手法、BNC は従来の代表的手法を表す。K はピボット数を表す。MNIST、Booklog、YahooNews は実験データの種類を表す。それぞれ手書き文字 DB, 本のレビューDB,新聞記事 DB である。r はレンジ検索における問合わせオブジェクトからの距離を表す。表の中の数値はピボットを用いることにより、問い合わせオブジェクトとの検索実行時における距離計算が不要になったオブジェクト数の

表1.レンジ検索の性能比較

| | MNIST | | Booklog | | YahooNews | |
|---|---|---|---|---|---|---|
| K=1 | PGM | BNC | PGM | BNC | PGM | BNC |
| r=1.0 | 9.0 | 1.1 | 39.6 | 0.0 | 28.3 | 0.0 |
| r=0.5 | 40.3 | 20.6 | 57.8 | 0.1 | 54.8 | 0.0 |
| r=0.25 | 67.4 | 52.7 | 72.4 | 1.1 | 73.5 | 2.6 |
| r=0.125 | 83.2 | 75.1 | 83.7 | 7.1 | 85.9 | 11.2 |
| r=0.0625 | 91.5 | 87.4 | 91.2 | 20.0 | 92.8 | 23.9 |
| K=5 | PGM | BNC | PGM | BNC | PGM | BNC |
| r=1.0 | 22.5 | 7.6 | 85.8 | 0.3 | 66.5 | 0.5 |
| r=0.5 | 77.7 | 53.6 | 97.3 | 1.1 | 95.2 | 2.2 |
| r=0.25 | 97.4 | 92.5 | 99.6 | 4.8 | 99.5 | 10.2 |
| r=0.125 | 99.8 | 99.4 | 99.9 | 20.6 | 100.0 | 34.4 |
| r=0.0625 | 100.0 | 100.0 | 100.0 | 43.9 | 100.0 | 64.9 |
| K=10 | PGM | BNC | PGM | BNC | PGM | BNC |
| r=1.0 | 28.9 | 9.6 | 92.8 | 0.7 | 77.2 | 0.6 |
| r=0.5 | 87.4 | 69.9 | 99.2 | 2.2 | 98.4 | 4.3 |
| r=0.25 | 99.2 | 97.8 | 99.9 | 9.2 | 99.9 | 17.6 |
| r=0.125 | 100.0 | 99.9 | 100.0 | 28.8 | 100.0 | 48.2 |
| r=0.0625 | 100.0 | 100.0 | 100.0 | 55.5 | 100.0 | 80.9 |

割合（枝刈り率と呼ぶ）を表す。この値が大きいほど検索性能が良いことを表す。表１より、ピボット数、レンジの大きさ、データ数、実験データ種類によらず、提案手法の方が従来手法よりも枝刈り率が大きい、すなわち検索の実行時性能が良いことが分かる。

②クエリ分布に基づく一般化ピボット法による類似検索の高速化

検索クエリ集合にはユーザの嗜好，トレンドなどによって何らかの偏り・分布が存在すると考えられる。そこで，ユーザのクエリ分布を有する学習データを用いてピボットを構築すれば，各ユーザに対応したピボットが生成され，さらなる類似検索の高速化が期待できる。この考えに基づき、クエリ分布を考慮した一般化ピボット法を用いた類似検索の高速化法を提案した。具体的には、オブジェクト集合とクエリ集合の和集合を考え、ピボット法で用いる目的関数の拡張を行い、その目的関数を最大化するアルゴリズムを反復することで最適なピボット集合を構築する方法を提案した[学会発表⑪]。

ニュース記事 DB を用いて実験を行った。ニュース記事は予め“国内”,“経済”,“エンタメ”,“生活”,“地域”,“サイエンス”,“スポーツ”,“世界”の８つのジャンルに分類されていることから、１種類のジャンルの記事集合をクエリ集合とみなして実験を行った。その結果、ジャンルを限定しない類似検索（レンジ検索）の場合と比べて、“ エンタメ”,“ 地域”ジャンル以外は、ジャンルを考慮する方が枝刈り率が高いこと、すなわち検索の実行時性能が優れることが確認できた。

ニュース記事 DB 以外に手書き文字 DB に関しても同様な実験を行い、短時間で，高い枝刈り性能を持つピボットを生成することを確認した[学会発表⑦]

(2) 大量データ可視化方法：類似検索においては大量データを分かり易く可視化する方法の確立も重要な課題である。

可視化結果へアノテーションを自動付与する手法を提案した[学会発表⑮]。ここで、アノテーションとは、注釈のことを言う。提案手法は，最小全域木とオブジェクトの属性情報を用いて，属性分布が大きく変化する部分で領域を分割することにより，与えられたオブジェクト集合をある特徴的な属性を持つオブジェクト部分集合へと分割し，Z-スコアを用いてその集合の特徴をアノテーションとして抽出することを特徴とする。実験により，PCA 法、MDS 法などの可視化法による低次元への埋め込み座標情報だけでなく，実際の位置情報に基づく座標情報を扱うデータセットに対しても，オブジェクト集合の特徴抽出が期待できることを確認した。

図１に、実際の位置情報に基づく座標情報を扱うデータセットである飲食店集合の可視化結果を示す。提案法による分割では”北海道”,”東北”,”関東”,”東京”,”横浜”,”中部”,”京都”,”関西・近畿”,”四国・中国・九州”,”沖縄”に分かれた。

それぞれの色が示す特徴を見てみると，北海道を示す茶色地域では”スープカレー”,”ジンギスカン”，関西・近畿地方を示す水色地域では”たこ焼き”,”串揚げ・串かつ”，沖縄を示す青地域では，”沖縄料理”といった各地域独特のジャンル，名物を抽出できているのが分かる。

また，東京付近では２つの集合が見られ，それぞれ，”居酒屋”,”ダイニングバー”といった他の地域では目立って出現しないジャンルが現れる地域と，”広東料理”,”飲茶・点心”,”肉まん・中華まん”が特徴的に出現する地域，中華街エリアを抽出されたのが分かる。よって，単純にその集合で多く見られた属性を抽出しているわけではなく，他では見られない，特徴的属性を抽出していることが確認できる。

また、[学会発表⑫]において、この方法をL1（マンハッタン距離）埋め込み手法に拡張した方法を提案した。この埋め込み手法は、一般化ピボット技術を用いる手法である。新聞記事データを用いた実験を行った。記事の属性ベクトルは各記事を形態素解析して得られた単語頻度ベクトルである。L2（ユークリッド距離）埋め込み手法と比べて、提案手法は外れ値的オブジェクトによる影響を受

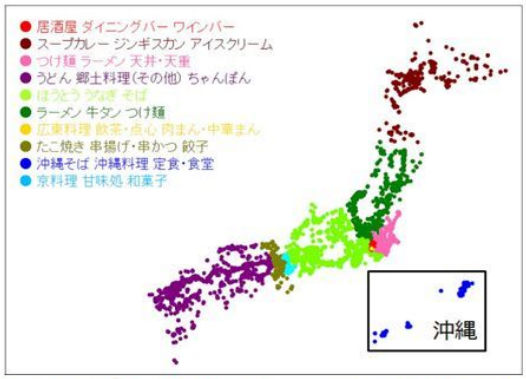

図1.飲食店集合のアノテーション可視化結果

けにくいという良い性質を有することを確認した。

②ピボットを用いた可視化方法の提案

（１）①で提案したピボット生成方法を用いた可視化方法を提案した。具体的には、生成するピボットを２つに限定し、各オブジェクトと２つのピボットとの距離より、オブジェクトを２次元空間に射影する方法を提案した[学会発表⑥]。

ピボット生成方法として従来方式を用いた場合との比較実験を行った。データは画像データ（CoPhIR）の４つの MPEG-7 属性を用いた。提案手法は、従来のピボット生成方法を用いた場合と比べて、射影結果の散らばりが大きい、すなわち優れた可視化結果であることを確認した。

また、可視化結果の評価方法の提案も行った[学会発表①]。具体的には、可視化前のオブジェクト間の距離順序が、可視化された 2 次元空間内の距離順序で保存される度合いで評価を行う。上記実験で用いた画像データの４つの MPEG-7 属性に関して、どの属性を用いるのが可視化に適しているかを評価した。可視方法は上記で提案した方法に加えて多次元尺度法、スペクトラル法を用いた。実験の結果、どの可視化方法でも 4 属性中で ColorStructure 属性が最も良く距離順序を保存することと、属性の種類によって可視化方法の優劣が異なることを明らかにした。

(3)クラスタリング方法：大量データの活用の観点からは類似検索と並んで重要な技術にクラスタリング技術がある。

大規模多クラス向けの k-means クラスタリグの、高性能なアルゴリズムを提案した[学会発表⑤]。不要な距離計算を回避することによって性能向上を実現している。具体的には、２つの工夫によって性能向上を実現している。

第１の工夫は、クラスタリング計算の初期段階において、オブジェクトとセントロイド間の距離の下限値を用いることである。この下限値は、ピボットを配置し、三角不等式を用いることにより求める。

第２の工夫は、オブジェクトが所属するクラスタが変化しない場合、すなわちセントロイドも変化しない場合には、セントロイドとオブジェクト間の距離計算を省略することである。さらに既存の有力な方法も組み込んだ。性能比較の結果、距離計算の削減率と計算時間の双方において、既存アルゴリズムよりも優れていることを示した。

SDM はデータマイニング分野のトップカンファレンスの一つであり（Microsoft の Top conferences in data mining では、41 の関連国際会議中の 5 位)、採択されたことは研究が高い評価を受けたことを示唆する。

従来手法との性能比較結果を図 2 に示す。Hamerly は従来の代表的手法を表し、

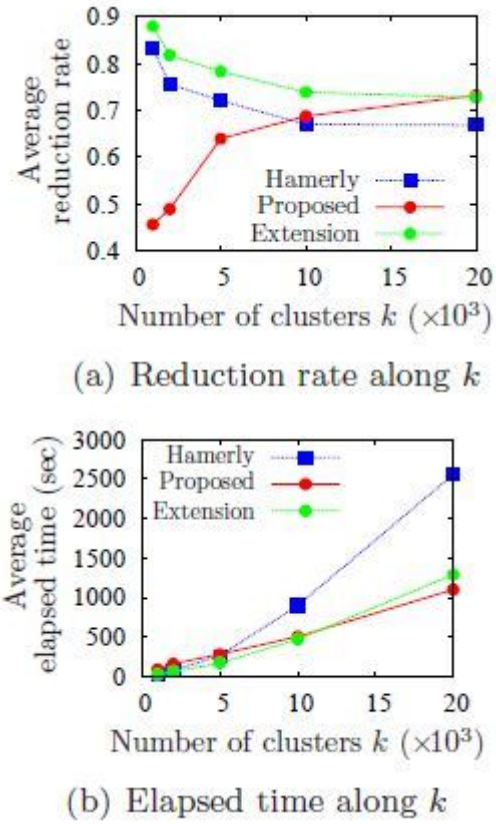


図 2.クラスタ数 1000～20000 での性能比較

Proposed が提案する手法を表す。Extension は Proposed に Hamerly の手法を組み込んだ手法である。(a)からクラスタリング生成において距離比較が不要となる割合（reduction rate）は Extension が最も優れ、(b)からクラスタ生成時間は Proposed よりはやや劣るものの Extension はほぼ程度の生成時間であることが分かる。すなわち、Extension は距離計算の削減率と計算時間の双方において、Hamerly の手法よりも優れる。

以上で説明した(1)(2)(3)のいずれの項目も、類似検索および関連する大量データの活用技術に関して新規性・有用性を有する技術を提案するものであり、意義の大きい成果と考える。

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔学会発表〕（計　15　件）


① 宋鵬, 斉藤 和巳, "画像データを用いた距離順位保存度による可視化結果の評価,"情報処理学会第 79 回全国大会 (IPSJ2017), 2017 年 3 月 16 日-3 月 16 日,名古屋大学東山キャンパス.

② 塚本竜太郎, 斉藤 和巳, "移動中心性と移動連結性による都市避難地の比較評価,"情報処理学会第 79 回全国大会 (IPSJ2017), 2017 年 3 月 16 日-3 月 16

日,名古屋大学東山キャンパス.

③ 宋 鵬, 斉藤 和巳, "画像データ間の距離定義の違いによるピボット可視化評価, "第 15 回情報科学技術フォーラム（FIT2016）,2016 年 9 月 7 日-9 月 7 日, 富山大学 五福キャンパス.

④ 宋 鵬, 斉藤 和巳, "ピボット選択法と距離定義の違いによる類似探索性能評価, "第 13 回 ネットワーク生態学シンポジウム（NETECO2016）,2016 年 8 月 22 日-8 月 23 日 木更津 オークラアカデミアパーク ホテル

⑤ Takashi Hattori, Kazuo Aoyama, Kazumi Saito, Tetsuo Ikeda and Eri Kobayashi, "Pivot-based k-means Algorithm for Numerous-class Data Sets," Proc. of the International Conference on Data Mining (SDM2016), May, 2016, Miami, Florida, USA.

⑥ 島崎 涼, 小林 えり, 斉藤 和巳, 池田 哲夫, "画像データを用いた L1 ピボット可視化法の評価, "情報処理学会第 78 回全国大会(IPSJ2016), 2016 年 03 月 10 日～2016 年 03 月 12 日, 慶應義塾大学矢上キャンパス.

⑦ 小林 えり, 斉藤 和巳, 池田 哲夫, 青山 一生, 服部 正嗣, "L1 距離上でのクラス割り当てピボットによる類似検索, "第 162 回 DBS 研究発表会 (SIG-DBS162), 2015 年 11 月 26 日, 芝浦工業大学・豊洲キャンパス.

⑧ 小林 えり, 斉藤 和巳, 池田 哲夫, 青山 一生, 服部 正嗣, "クエリ分布を考慮した一般化ピボット法の距離定義による特性評価, "第 14 回情報科学技術フォーラム（FIT2015）, 2015 年 09 月 15 日～2015 年 09 月 17 日, 愛媛大学 城北キャンパス.

⑨ 小林 えり, 斉藤 和巳, 池田 哲夫, "Mean-shift クラスタリングによる類似ユーザ分析法とその性能評価, "第 12 回観光情報学会全国大会（STI2015）, 2015 年 06 月 19 日, 石川県金沢市青草町 近江町交流プラザ.

⑩ 小林 えり, 斉藤 和巳, 池田 哲夫, 青山 一生, 服部 正嗣, "クエリ分布を考慮した類似検索の高速化, "第 29 回人工知能学会全国大会（JSAI2015）, 2015 年 05 月 31 日, 北海道函館市 公立はこだて未来大学.

⑪ Eri Kobayashi, Takayasu Fushimi, Kazumi Saito, and Tetsuo Ikeda, "Similarity Search by Generating Pivots based on Manhattan Distance, " Proc. of the 13th Pacific Rim International Conference on Artificial Intelligence (PRICAI2014), pp.435-446, 2014 年 12 月 04 日, Gold Coast, Queensland, Australia.

⑫ 小林 えり, 斉藤 和巳, 池田 哲夫, 大久保 誠也, "L1 埋め込みによるアノテーション付き可視化法, "第 7 回 Web とデータベースに関するフォーラム（WebDB2014）, 2014 年 11 月 20 日, 東京都江東区豊洲、芝浦工業大学.

⑬ 小林 えり, 伏見 卓恭, 斉藤 和巳, 池田 哲夫, " メディアンに基づく時系列データの変化点検出法, "第 13 回情報科学技術フォーラム（FIT2014）, 2014 年 09 月 04 日, 茨城県つくば市、筑波大学.

⑭ 小林 えり, 斉藤 和巳, 池田 哲夫, 大久保 誠也, "可視化結果へのツリー分割による アノテーション付与法, "日本ソフトウェア科学会ネットワークが創発する知能研究会（JWEIN2014）, 2014 年 08 月 21 日～2014 年 08 月 24 日, 東京都調布市調布ヶ丘、電気通信大学.

⑮ 小林 えり, 伏見 卓恭, 斉藤 和巳, 池田 哲夫, "ツリー分割によるアノテーションの可視化法, "第 102 回人工知能学会知識ベースシステム研究会 (SIG-KBS), 2014 年 7 月 24 日, 大阪市北区茶屋町、関西学院大学.

〔その他〕
特になし

## ６．研究組織

(1)研究代表者
池田 哲夫（IKEDA, Tetsuo）
静岡県立大学・経営情報学部・教授
研究者番号：60363727

(2)研究分担者
武藤 伸明（MUTOH, Nobuaki）
静岡県立大学・経営情報学部・教授
研究者番号： 40275102

斉藤 和巳（SAITO, Kazumi）
静岡県立大学・経営情報学部・教授
研究者番号： 80379544

大久保 誠也（OKUBO Seiya）
静岡県立大学・経営情報学部・助教
研究者番号：90422576

以上