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

令和　２ 年　 ５ 月 １２ 日現在

機関番号：２３８０３
研究種目：基盤研究(C)（一般）
研究期間：2017～2019
課題番号：１７Ｋ００１５９

研究課題名（和文）縮小写像での積算型下限値によるクラスタリング法の高速化

研究課題名（英文）Speeding up the clustering methods with summable lower bounds in contractive mappings

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）　　3,300,000円

研究成果の概要（和文）：本研究計画の目的は、大規模データの効率的なクラスタリング技法の確立である。1. クラスタリングの関連技術である類似検索技術として、CBT（完全2分木）索引を生成する索引生成技術と、このCBT索引を用いるオンライン類似検索技術、2. Lloyd型クラスタリング技法における、ユークリッド距離自乗の積算型下限値の導入を特徴とする効率的な加速手法、3. オブジェクトと平均特徴ベクトル双方の疎表現と、平均特徴ベクトルの集合の転置ファイルデータ構造を特徴とする、転置ファイルk-means技法等を考案した。
何れもクラスタリング技法および関連する類似検索技法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。

研究成果の学術的意義や社会的意義
画像、文書、DNA 配列などのマルチメディアデータは近年爆発的に増加している。これらのマルチメディアデータの集合の基本構造を把握し理解するための技法としてクラスタリング技法と類似検索技法がある。クラスタリングとは、データの集合をクラスタという互いに似ているデータからなる部分集合に分けることである。類似検索とは、入力となるデータと類似度の大きいデータを検索することである。
クラスタリングおよび類似検索ともに、一般にデータ量が大きいと処理時間を多く要することが知られており、高効率なクラスタリング技法及び類似検索技法の実現が強く求められている。本研究の成果はこの要望に応えるものである。

研究成果の概要（英文）：The purpose of this research project is to establish efficient clustering and similarity search technologies for large data:
(1) We proposed index construction algorithm that recursively builds a CBT (complete binary tree) index, and an online similarity search algorithm that efficiently prunes unnecessary branches and filters objects by using the CBT index. (2) We proposed an efficient acceleration algorithm for Lloyd-type k-means clustering, which employs a projection-based filter (PRJ) to avoid unnecessary distance calculations. The PRJ exploits a summable lower bound on a squared distance defined in a lower-dimensional space to which data points are projected. (3) We proposed an inverted-file k-means clustering algorithm (IVF). To achieve high performance, IVF exploits two distinct data representations. One is a sparse expression for both the object and mean feature vectors. The other is an inverted-file data structure for a set of the mean feature vectors.

研究分野：データ工学

キーワード：情報検索　クラスタリング　縮小写像

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

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、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る。

これらのマルチメディアデータの集合の基本構造を把握し理解するための技法としてクラスタリング技法がある。クラスタリングとは、データの集合をクラスタという互いに似ているデータからなる部分集合に分けることである。

クラスタリング技法は、一般にデータ量が大きいと処理時間を多く要することが知られており、高効率なクラスタリング技法の実現が強く求められている。

本研究では、効率的なクラスタリング技術の確立に焦点を当てる。

２．研究の目的

本研究は、クラスタリング問題において、縮小埋め込み技術を土台にして、オブジェクト間などの不要な距離計算を削減する技術の確立を目的とする。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)、埋め込み後の距離 δ(F(x), F(y)) で、不等式δ(F(x), F(y))≦ d(x, y) が成立することを言う。

本研究参画者らの基盤研究(C) (26330138)(平成 26-28 年度)での最重要施成果の一つに、ユーザがクエリとして与えるオブジェクト q との距離がレンジ r 以下となるオブジェクト部分集合を効率良く探索するレンジクエリ検索など各種類似検索問題において、機械学習アプローチに基づく反復改善法をベースに、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）を構築する技術がある。これも縮小埋め込みに基づく技術である。

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、クラスタリング技術に関連する技術である類似検索技術の研究も実施した。

具体的に研究方法を以下に説明する。

(1) ピボットを用いた効率的な類似検索技法の提案

類似検索では、一般にオフラインでの索引生成とオンラインでの類似検索の双方が高性能なことが要求される。この要求に応えるため、索引生成とオンライン類似検索の新技術を提案した。

(2) 類似検索を高速化するための新たなピボット生成法の提案

次いで、類似検索を高速化するための新たなピボット生成法を提案した。

(3) 積算型下限値を用いた Lloyd 型クラスタリングアルゴリズムの効率的な acceleration 手法の提案

大規模高次元データで多くのクラスを前提にした場合の、高速なクラスタリングアルゴリズムとして有名な Lloyd 型のアルゴリズムの効率的な acceleration 手法を提案した。

(4) 転置ファイル k-means クラスタリングアルゴリズムの性能分析

大規模高次元疎データに適した、転置ファイル k-means クラスタリングアルゴリズム（inverted-file k-means clustering algorithm）を提案し、優れた処理性能を有することを実証すると共に、高性能の要因の分析を行った。

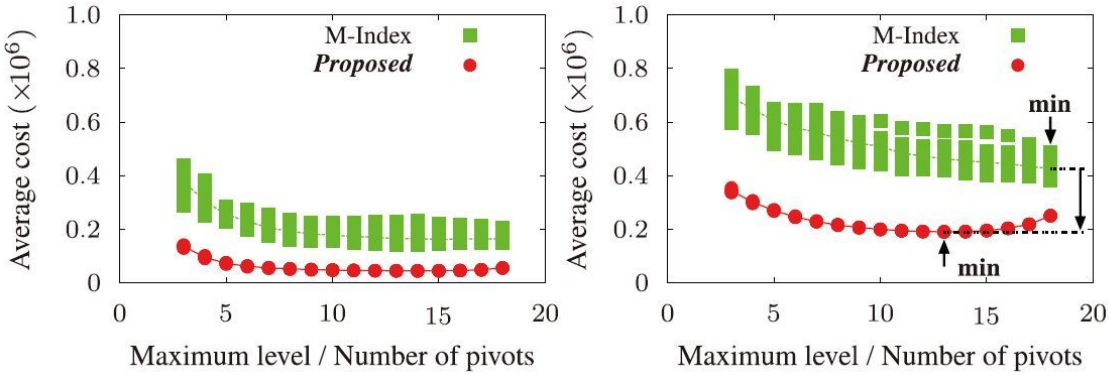
４．研究成果

(1)ピボットを用いた効率的な類似検索技法の提案[1]

マルチメディアデータのクラスタリング技術の関連技術である類似検索技術の研究を先ず行った。類似検索では、一般にオフラインでの索引生成とオンラインでの類似検索の双方が高性能なことが要求される。この要求に応えるため、索引生成とオンライン類似検索の新技術を提案した。前者は、ピボット生成関数と、ピボットを用いて各ノードでのオブジェクト集合をほぼ同じサイズの部分集合に分割する関数とを再帰的に実行して、CBT（完全 2 分木）索引を生成する。後者は、この CBT 索引を用いて、効率的に不必要な枝を枝刈りしオブジェクトをフィルタリングする。

大規模画像データを用いて実験を行った、実験対象データは、1 億枚の画像から抽出した MPEG-７記述子データから構成される CoPhIR (Content-based Photo Image Retrieval)を用いた。実験では、提案手法と代表的手法の M-index の類似検索コストを比較した。

SC（scalable color）記述子を用いた場合の類似検索性能比較および複数の記述子の混合記述子（MX）を用いた場合の類似検索性能比較の両方において、平均的に提案法の方が M-index よりも優れているが分かった（図 1(a)、図 1(b)）。また、SC 記述子の場合は提案法の検索コストが M-index の約 4 分の１であることが分かり（図 1(a)）MX の場合は、提案法の検索コストが M-index の約半分であることが分かった（図 1(b)）。すなわち提案方法は、類似検索の性能が代表的手法の M-index よりも優れていることが実証された。



(a)SC 記述子　　(b)MX 記述子

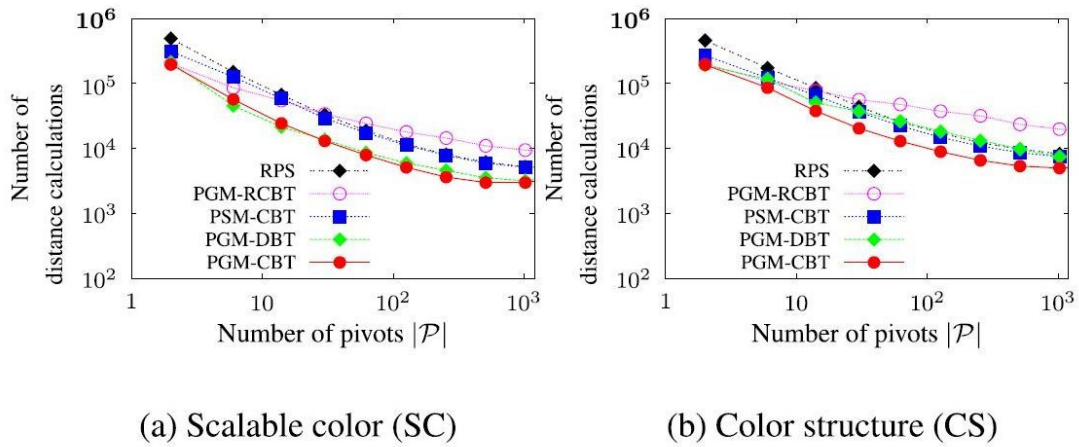
図１．提案法と M-index の類似検索コストの比較

(2) 類似検索を高速化するための新たなピボット生成法の提案[2]

次いで、類似検索を高速化するための新たなピボット生成法を提案した。提案法は、①オブジェクト集合を再帰的に同一サイズの二つの部分集合に分割し、結果的に完全 2 分木を生成する階層的オブジェクト集合分割法と、②少ない計算量での目的関数計算を特徴とする高速ピボット最適化技法、とを用いてピボットを生成する。提案法は、クエリとオブジェクト間の距離のタイトな下限を提供するという性質を持ち、類似検索での不必要な距離計算を効率的に回避できる。

大規模画像データを用いて実験を行った、実験対象データは、(1)と同様に 1 億枚の画像から抽出した MPEG-７記述子データから構成される CoPhIR (Content-based Photo Image Retrieval) を用いた。実験では、提案手法と比較手法で生成されたピボット集合を用いて、類似検索における距離計算回数の比較を行った。

SC 記述子、CS 記述子いずれの場合も提案法 (PGM-CBT) で生成されたピボット集合を用いる方が、比較手法で生成されるピボット集合を用いるよりも距離計算回数が少ないことが分かった (図 2(a)、図 2(b))。すなわち提案方法を用いることにより、類似検索での不必要な距離計算を効率的に回避できることが実証された。



(a) Scalable color (SC)　　(b) Color structure (CS)

図 2.提案手法と比較手法で生成されたピボット集合を用いての類似検索での距離計算回数比較

(3) 積算型下限値を用いた Lloyd 型クラスタリングアルゴリズムの効率的な acceleration 手法[3]

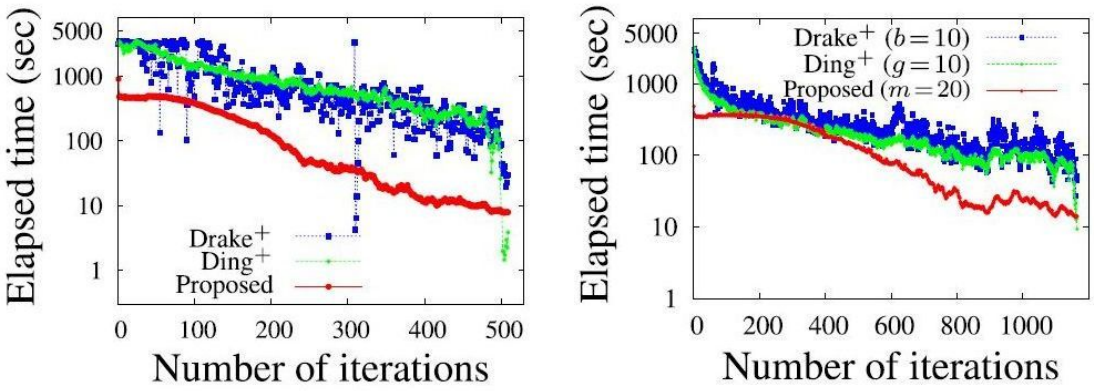
マルチメディアデータのクラスタリング技術に関する研究を行った。具体的には、大規模高次元データで多くのクラスを前提にした場合の、高速なクラスタリングアルゴリズムとして有名な Lloyd 型のアルゴリズムの効率的な acceleration 手法を提案した。本研究の貢献は以下の 3 点である。

1)効率的な acceleration のスキームを提案した。鍵となるのは、新たに導入したユークリッド距離自乗の積算型下限値である。研究計画書でも説明したように、距離自乗は自乗要素の和であ

るため、距離自乗要素の部分和は、距離自乗の下限値となる。従って、距離自乗要素を増やすことにより、下限値を改善できるという優れた性質を有する。
2)現実的なアルゴリズムを提案した。アルゴリズムは２つの filter を用いる。第１は、上述した積算型下限値に基づく射影フィルタである。第２は、我々が以前に提案したフィルタに類似したフィルタである invariant centroid-pair based filter(ICP)である。PRJ においては、低次元空間を生成する直交基底を得るため に、所与のデータポイント集合の特異値分解を用いる。
3)実験によって、現時点で最速と考えられている Drake アルゴリズムと Ding アルゴリズムとの比較を行った。

実験対象データは、大規模高次元の画像データセット 2 種類（TinyImages と Holidays）を用いた。TinyImages では、画像は 384 次元の GIST 特徴ベクトルで表現される。実験用に 10,219,916 個の特徴ベクトルを抽出した。Holidays では、画像は、128 次元の特徴ベクトルの集合で表現される。実験用に、20,964,516 個の特徴ベクトルを抽出した。実験では、各アルゴリズムの主要ループの１繰り返しごとの所要時間の比較を行った。

TinyImages においては、提案法が最も所要時間が少なく、Ding アルゴリズムの約 17%の平均所要時間であることが分かった（図 3(a)）、Holidays においては、所与時間は Drake アルゴリズムの約 17%、Ding アルゴリズムの約 47%であることが分かった（図 3(b)）。すなわち、提案法が効率的な acceleration 手法あることを確認した。



(a)Tinyimage　　(b)Holidays

図 3.主要ループの一繰り返しごとの所要時間の比較

(4) 転置ファイル k-means クラスタリングアルゴリズムの性能分析[4]

大規模高次元疎データに適した、転置ファイル k-means クラスタリングアルゴリズム（inverted-file k-means clustering algorithm）（以下、IVF と呼ぶ）を提案した。IVF は、大規模高次元疎データに対して、標準的な k-means アルゴリズムである Lloyd's アルゴリズムと同一の解を維持しつつ、高速かつ低メモリ消費量で効率的に動作する。高性能性は、２つの異なるデータ表現に起因する。１つは、オブジェクト特徴ベクトルと平均特徴ベクトル双方の疎表現である。もう一つは、平均特徴ベクトルの集合の転置ファイルデータ構造である。前者によって、消費メモリ量の削減を可能にし、後者によって、高速性を可能にした。

これらの表現の効果を確認するため、異なるデータ表現とデータ構造を有する３つのアルゴリズムを設計し、アウトオブオーダー実行が可能なスーパースカラープロセッサと深いメモリ階層を備えた最新の計算機システムを用いて実験を行った。大規模な実文書データセットに IVF を適用した場合に、設計されたアルゴリズムよりも優れた性能が得られることを実証した。
実験対象文書としては、医学分野の代表的な文献情報データベースである PubMed の要約文書のうち 1,000,000 文書と、The New York Times の記事のうち 1,285,944 記事を用いた。各アルゴリズムの主要ループの１繰り返しごとの所要時間の比較を行った。

IVF はクラスタ数 k が大きい領域で優れた性能を有することが分かった。ピボット数 k=2000 で PubMed データを用いた場合は CPU 時間が 2 番目に性能の良いアルゴリズムである IFN の 33.7%であった（図 4(b)）。一方、kの値が小さい場合は、IVF と IFN の性能差はほとんどなかった。((図 4(a) （図 4(b))

また、命令当たりクロックサイクル（CPI）モデルを用いて、最新の計算機システムにおける高速動作の要因を分析した。その結果、キャッシュミス数、分岐誤予測数、完了命令数（投機的実行において実際に必要であると証明された命令数）という３つの性能劣化要因を抑制できることを明らかにした。

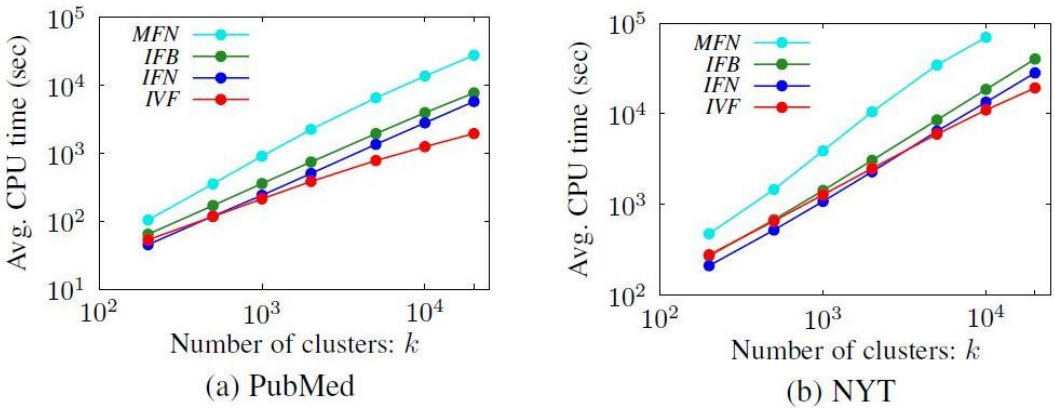


図 4．主要ループの一繰り返しごとの平均 CPU 時間の比

何れもクラスタリング技術および関連する類似検索技術の関して新規性を有する技術であって、有意義な成果である。

＜引用文献＞

[1]Yuki Yamagishi, Kazuo Aoyama, Kazumi Saito, and Tetsuo Ikeda,"Efficient Similarity Search with a Pivot-Based Complete Binary Tree," IEICE Transactions on Information and Systems,E100.D,pp.2526--2536,2018.

[2]Yuki Yamagishi, Kazuo Aoyama, Kazumi Saito, and Tetsuo Ikeda,"Pivot Generation Algorithm with a Complete Binary Tree for Efficient Exact Similarity Search," IEICE Transactions on Information and Systems,E101.D,pp.142--151,2018.

[3]AOYAMA Kazuo、SAITO Kazumi、IKEDA Tetsuo,"Accelerating a Lloyd-Type k-Means Clustering Algorithm with Summable Lower Bounds in a Lower-Dimensional Space," IEICE Transactions on Information and Systems,E101.D,pp.2733-2783,2018.

[4]Kazuo Aoyama, Kazumi Saito, Tetsuo Ikeda,"Inverted-File k-Means Clustering: Performance Analysis," arXiv:2002.09094.

５．主な発表論文等

〔雑誌論文〕　計4件（うち査読付論文　3件／うち国際共著　0件／うちオープンアクセス　3件）

１．著者名 Kazuo Aoyama, Kazumi Saito, Tetsuo Ikeda	４．巻 －
２．論文標題 Inverted-File k-Means Clustering: Performance Analysis	５．発行年 2020年
３．雑誌名 arXiv:2002.09094	６．最初と最後の頁 －
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 AOYAMA Kazuo、SAITO Kazumi、IKEDA Tetsuo	４．巻 E101.D
２．論文標題 Accelerating a Lloyd-Type k-Means Clustering Algorithm with Summable Lower Bounds in a Lower-Dimensional Space	５．発行年 2018年
３．雑誌名 IEICE Transactions on Information and Systems	６．最初と最後の頁 2773～2783
掲載論文のDOI（デジタルオブジェクト識別子） https://doi.org/10.1587/transinf.2017EDP7392	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Yuki Yamagishi, Kazuo Aoyama, Kazumi Saito, and Tetsuo Ikeda	４．巻 Volume E100.D
２．論文標題 Efficient Similarity Search with a Pivot-Based Complete Binary Tree	５．発行年 2017年
３．雑誌名 IEICE Transactions on Information and Systems	６．最初と最後の頁 2526--2536
掲載論文のDOI（デジタルオブジェクト識別子） https://doi.org/10.1587/transinf.2017EDP7100	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Yuki Yamagishi, Kazuo Aoyama, Kazumi Saito, and Tetsuo Ikeda	４．巻 Vol.E101-D
２．論文標題 Pivot Generation Algorithm with a Complete Binary Tree for Efficient Exact Similarity Search	５．発行年 2018年
３．雑誌名 IEICE Transactions on Information and Systems	６．最初と最後の頁 142--151
掲載論文のDOI（デジタルオブジェクト識別子） https://doi.org/10.1587/transinf.2017EDP7077	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

〔学会発表〕　計1件（うち招待講演　0件／うち国際学会　0件）

１．発表者名 宋鵬，斉藤 和巳，池田哲夫，青山一生
２．発表標題 貪欲到達中心性によるネットワーク探索性能の特徴付け
３．学会等名 第16回情報科学技術フォーラム（FIT2017）
４．発表年 2017年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究分担者	斉藤　和巳 (Saito Kazumi) (80379544)	静岡県立大学・経営情報学部・教授 (23803)	
研究分担者	青山　一生 (Aoyama Kazuo) (80447028)	日本電信電話株式会社ＮＴＴコミュニケーション科学基礎研究所・知能創発環境研究グループ・研究員 (94305)	