

平成 26 年 5 月 7 日現在

機関番号：23803

研究種目：基盤研究(C)

研究期間：2011～2013

課題番号：23500128

研究課題名(和文)メトリック空間オブジェクトの縮小埋め込みによる類似検索の高速化

研究課題名(英文)Speeding up the similarity search by contractive embedding methods

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) 3,900,000円、(間接経費) 1,170,000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。具体的にはBustosらのピボットによる類似検索方法を土台として類似検索方法を開発することである。類似検索方法として、Bustos法でのピボット集合要素の交換方法を改良した方法と、マンハッタン距離に基づく一般化ピボット法の2方法を提案した。2方法とも従来方法に比較して、類似検索性能、ピボット集合選択時間が優れていることを実験で確認した。

さらに、実験対象データの性質を効率的に解析可能とすることを目的として、ネットワーク内のコミュニティ抽出方法と、ネットワーク可視化方法を複数考案した。

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. Although Bustos and others have devised several pivot-based methods and have obtained the excellent results, in this field there is still room for further investigation.

We devised two pivot-based similarity search methods. The first method is characterized by improving Bustos's exchange algorithm of pivot set elements. The second method is a generalized pivot method based on Manhattan Distance. We devised it by extending one of our previously-developed methods. By experiments, we showed that both methods were better than the conventional methods in retrieval performance and creation time of pivot sets.

Furthermore, we devised extraction methods of the communities in networks and network visualization method in order to make clarifying the nature of the multimedia data more efficient.

研究分野：総合領域

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：情報検索 類似検索

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 と最類似するオブジェクトを効率良く探索する最類似検索問題に焦点を当てる。

本研究で対象とするような複雑データを構成要素とするオブジェクトでは、類似計算は複雑で、任意のオブジェクトペア間に距離公理を満たすメトリックが定義されたメトリック空間での最類似検索問題となる。

メトリック空間での最類似検索は近年活発に研究されているものの、改良の余地が大きい。

2. 研究の目的

本研究では、先ずメトリック空間のオブジェクト集合をユークリッド空間の固定次元ベクトル集合へ写像する、優れた埋め込み関数 $F(x)$ を構築し、次いで、ユークリッド空間での適当な最類似検索法の利用による、メトリック空間でのより高速な最類似検索手法を開発することで、問題全体にアプローチする。

研究を進めるにあたって、縮小埋め込みに着目する。縮小埋め込みに関しては、埋め込み先での容易に計算可能な距離を用いることにより、メトリック空間での複雑な距離計算の回数を削減できるという、最類似検索に有用な性質を有することが知られているためである。

本研究では、Bustos らが提案した、メトリック空間でのピボットによるレンジクエリ法を土台に検討を進める。レンジクエリとは、与えられたレンジ r に対して、クエリ q との距離 $d(x, q)$ が r 以下のオブジェクト集合を求める問題である。

この枠組みでの最類似検索性能は、どのようなピボット集合を採用するかで大きく影響される。Bustos らは、予めランダムに選定した候補ピボット集合と候補オブジェクト集合を利用し、候補オブジェクト同士のマンハッタン距離 $(F(x), F(y))$ の総和を最大にするように、ピボット集合を貪欲法に基づき選択する方法を提案している。そこでの目的関数は、任意のオブジェクトペア x と y に対して、 $d(x, y) = (F(x), F(y))$ となることのみを指向するが、最類似検索性能を直接向上させる別の目的関数の導出も可能と考えられる。また、このような選択問題の目的関数

はサブモジュラと呼ばれる数値構造を持つと想定できる。よって、遅延評価と呼ばれる手法の導入で貪欲法の大幅な高速化や性能改善が期待できる。本研究参画者は、ピボット選択と類似の性質を持つと考えられる施設配置問題において、サブモジュラ構造を利用した解法の高速化を実現している。

3. 研究の方法

本研究は、以下のような方法で、縮小埋め込みを用いた最類似検索の性能向上を目指す。

(1) 最類似検索性能向上に寄与するピボット選択目的関数の構築

Bustos らの基本目的関数を土台に、重みの導入などの工夫により、優れたピボット選択目的関数を構築する。

(2) サブモジュラ構造を利用したピボット選択の高速化法の構築

基本目的関数や重み付き目的関数はサブモジュラ関数となることに着目する。サブモジュラ構造を利用することにより、我々が提案済の遅延評価法の活用により高速なピボット選択方法の構築を目指す。

(3) メトリック空間オブジェクト集合をユークリッド空間へ写像する、優れた埋め込み関数の構築

最類似検索の性能向上のため、いくつかの埋め込み関数に関して検討する。

(4) 構築した埋め込み法に基づいた、ユークリッド空間での適当な最類似検索法の利用による、メトリック空間での高効率な最類似検索手法の開発

ピボット数 k の設定法とともに、構築した方法のさらなる改良をはかる。

4. 研究成果

(1) ピボットによるレンジクエリの高速化に向け、2 つの方法を新たに提案し、それぞれの特性を明らかにした。2 方法とも従来方法と比較して、類似検索性能が優れていることを、実データを用いた実験で確認した。

[雑誌論文] Bustos らのピボット集合選択方法には、目的関数を最適化するために、ピボット集合の要素の交換を用いる方法がある。

我々は、Bustos らの方法におけるピボット集合要素の交換方法を改良した方法、より具体的にはランダムに選択されたピボット候補をピボット集合に追加した後、最悪の要素を削除することを特徴とする方法を提案した。実データ (EMACS 辞書データ) を用いて、Bustos 法 (図では INC 法と略) と提案方法 (図では RIWO-INC 法と略) の類似検索性能 (実際のレンジクエリでの距離計算回数の期待値) を比較した (他のデータでの比較も実施したが紙面の都合上略す)。

図 1 において、提案方法は、Busots 法と比

較して縦長の長方形同士が十分離れており、有意に少ない距離計算回数の期待値が得られている。

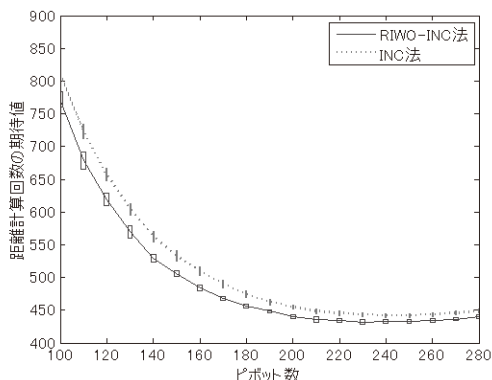


図1 類似検索性能の比較

[学会発表] Bustos の方法等の従来のピボット法ではピボットを与えられたオブジェクト集合から選択していた。これに対し、分担研究者らはピボットをオブジェクト集合の中から選択するのではなく、オブジェクト空間から最適なピボットを構成する一般化ピボット法が提案している[木村]。[木村]においては、ユークリッド距離での対応を可能としているが、我々は新たにマンハッタン定義に対応可能となるよう拡張した。これを、「マンハッタン距離に基づく一般化ピボット法」と呼ぶ。

[木村]木村学, 斉藤和巳, 上田修功: "効率的な類似検索のためのピボット学習法", 情報処理学会論文誌, Vol. 50, No. 8, 1883-1891 (2009)

MNIST データ (手書き数値文字データ) を用いて、Bustos 法 (図では bnc と略記) と提案方法 (図では pvl と略記) を比較した (他のデータでの比較も実施したが略す)。

横軸はレンジ r を表し、縦軸は枝刈り成功率を表す。縦軸の値が大きいくほど効果的な枝刈りが実現でき、結果的に少ない計算量で類似計算が可能なることを意味する。 k はピボット数を表す。

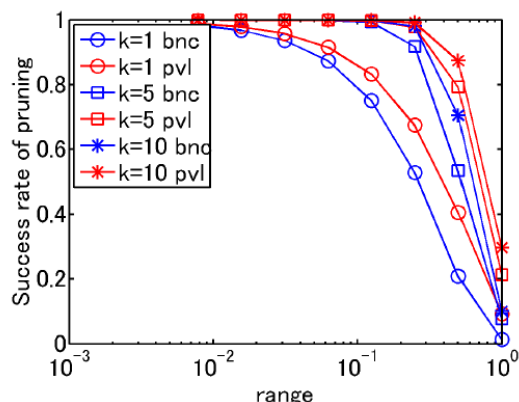


図2 レンジクエリ枝刈り比率の比較

図2のMNISTの結果からは、レンジ r が小

さい、または大きい ($r=1.0$) のとき、Bustos 法と提案方法に大きな違いは見られないが $r=0.1$ 、 $r=0.25$ 、 $r=0.5$ ではどのピボット数でも提案方法が有意に効率的な枝刈りができていることが分かる。

(2) ピボットによる類似検索ではピボット選択自体の高性能化も重要である。(1)で提案したそれぞれの方法に関して高速化方法を考案し、実験によってピボット選択の高性能化が実現できることを確認した。

[学会発表] (1) で提案した方法は、ピボット選択の目的関数がサブモジュラ性を有することから、遅延評価法を導入し、高速化をはかった。我々は、ピボット選択と類似の性質を持つと考えられる施設配置問題において、サブモジュラ構造を利用した解法の高速度を実現しており[武藤]、この従来研究の延長線上での工夫である。

[武藤]武藤, 斉藤, 池田, 他, 非線形施設配置問題における貪欲法の遅延評価による高速化-複数施設が確率的に貢献する場合-, 情報処理学会論文誌, 数理モデル化と応用, Vol. 4, No. 1, pp. 1-10, 2011.

MNIST データ (手書き数値文字データ) を用いて、Bustos 法と提案方法のピボット集合選択時間を比較した。具体的には、Bustos 法でのピボット集合選択時間と Bustos 法と同等の目的関数値を得るための提案方法での時間とを比較した (他のデータでの比較も実施したが略す)。

表1から、提案方法は約1/10の421.4秒で Bustos 法の目的関数値を得られている。

表1 ピボット集合選択時間の比較

| | Bustos 法 | RIWO 法 |
|-----------------|----------|---------|
| ピボット集合の選択にかかる時間 | 32899.6 | 5531.9 |
| 目的関数値 | 15497.8 | 15495.4 |
| クエリ毎の距離計算回数 | 18636.6 | 18714.1 |

[学会発表] (1) で提案したマンハッタン距離に基づく一般化ピボット法においては、高次元空間のデータを射影軸に射影後に、射影後値を降順ソートした順序リストを利用して高速化する工夫を考案した。

MNIST データ (手書き数値文字データ) を用いて、Bustos 法 (図では bnc と略記) と提案方法 (図では pvl と略記) を比較した他のデータでの比較も実施したが略す。 k はピボット数を表す。

図3の結果からは、どのピボット数、オブジェクト数でも提案方法のほうが有意に短い時間で結果を出力していることが分かる。

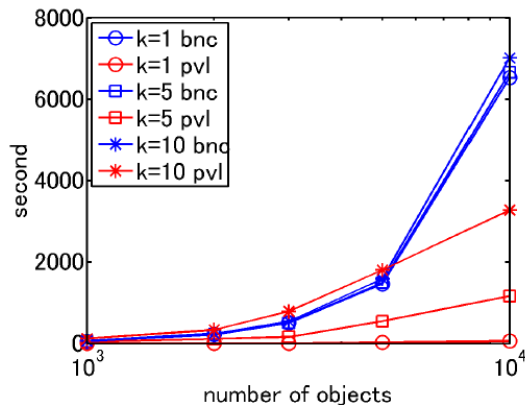


図3 ピボット集合選択時間の比較

(3) ネットワークでのコミュニティ抽出方法、可視化方法

実験対象データの性質を効率的に解析可能とすることを目的として、ネットワーク内のコミュニティ抽出方法と、ネットワーク可視化方法を考案した。抽出方法、可視化方法とも複数考案したが、紙数の都合で一例ずつ説明する。

コミュニティ抽出方法[雑誌論文] 同質の機能を有するノード群を同一のコミュニティとして抽出する手法を提案した。提案法では無向ネットワークを対象として、Web ページのランキングアルゴリズムとして有名な PageRank の収束過程の変化曲線によりノードをクラスタリングし、機能コミュニティを抽出する。ある大学のウェブページのハイパーリンクネットワークを用いて、提案方法によるコミュニティの抽出結果と、リンク密度に基づく代表的なコミュニティ抽出法である Newman クラスタリング法などの抽出結果との比較を行った。を比較する。(Newman クラスタリング法以外の方法との比較は略する)。

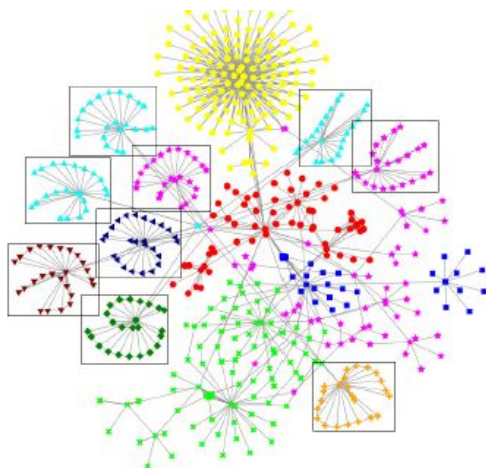


図4 提案法によるコミュニティ抽出

図4の提案法では可視化結果の四角で囲まれた部分は同じ色で塗り分けられている。この

ノード群は、対象大学の各年度の教員の研究成果報告ページであり、ノードの機能としては同質であると考えられ、同一のコミュニティとして抽出できている。一方図5のNewmanクラスタリング法は、年度ごとに異なるコミュニティとして抽出している。これは、異なる年度のノード同士が直接リンクされていないために、異なるコミュニティとして抽出されている。

提案方法では従来方法と異なる視点でのコミュニティ抽出が可能であり有効性が確認できた。

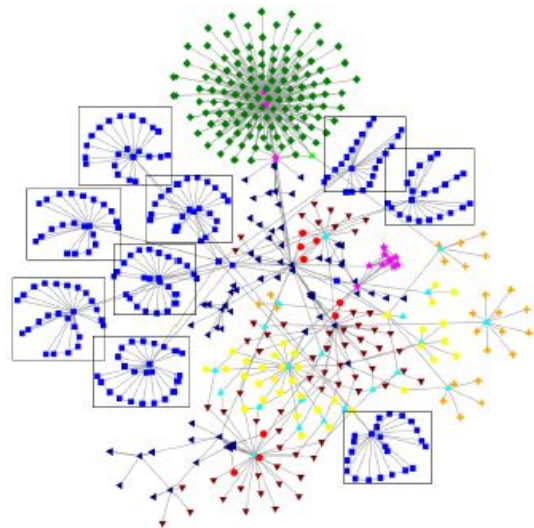


図5 Newman法にコミュニティ抽出

可視化方法[学会発表]

マンハッタン距離に基づく一般化ピボット法は、可視化手法として用いることもできる。ある大学のウェブページのハイパーリンクネットワークを用いて、提案方法での可視化結果と、高次元データの低次元空間への埋め込み法の代表例である主成分分析(PCA法)などでの可視化結果との比較を行った。(PCA以外の方法との比較は略する)。

図6のPCA法では、「ラッシュ」というブランドのアイテムがほかのブランドとは異なるユーザ層に支持されているため他のブランドとの距離が大きくなっている。そのため、「ラッシュ」とその他のブランド間の距離が支配的となり他のブランド間の距離が縮小され、多くのブランドの商品が水平軸の周囲に密集する結果となった。

図7の提案法では、オブジェクトが全体に満遍なく広がった結果となった。左上部分にはスキンケア、洗顔剤、ボディーケア商品が占め、左下にはクレンジングオイル、化粧落とし用品などがプロットされた。また、左下には「ケイト」ブランドといったティーンエイジャー向けメイクアップ商品を主力とするブランドのアイテムがプロットされ、左上には「オルビス」などのメイクアップ用品以外にも力をいれて市場展開しているブランドのメイクアップ商品がプロットされた。

提案法ではデータが万遍なく散らばったために、ボディケア・オーラルケア商品群によるオブジェクト集合の検出、価格やブランドを考慮した使用用途ごとのオブジェクト集合の検出など、より細かいオブジェクト間の関係を発見することができた。提案方法では従来方法と異なる特長を持つ可視化が可能であり有効性が確認できた。

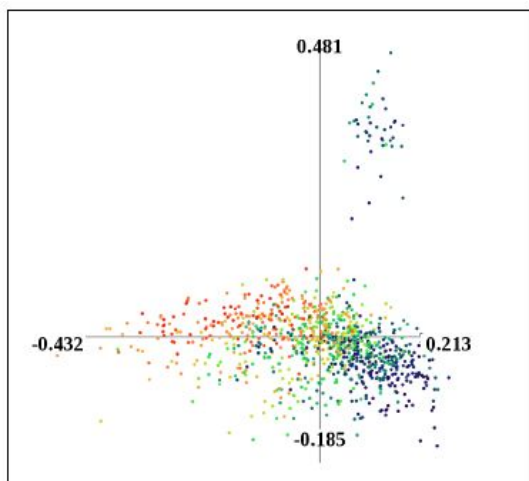


図 6 主成分分析でのオブジェクト分布可視化結果

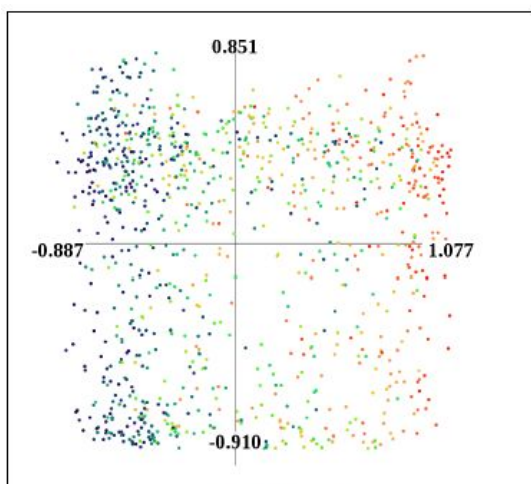


図 7 提案手法のオブジェクト分布可視化結果

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)

〔雑誌論文〕(計 7 件)

伏見卓恭, 齊藤和巳, 池田哲夫, 風間一洋, ノードの注目度に基づく機能コミュニティ抽出法, 情報処理学会論文誌 データベース, 査読有, Vol.6(4), 2013, 104-112

伏見卓恭, 齊藤和巳, 池田哲夫, 風間一洋, リンクの向きに着目した機能コミュ

ニティとモチーフの関係分析, 情報処理学会論文誌 数理モデル化と応用, 査読有 Vol.6(2), 2013, 137-146

伏見卓恭, 齊藤和巳, 池田哲夫, 武藤伸明, ノード群の協調的振舞いに着目した集合媒介中心性の提案と応用, 電子情報通信学会論文誌, 査読有, Vol. J96-D(5), 2013, 1158-1165

三津山雅規, 齊藤和巳, 池田哲夫, 大久保誠也, 武藤伸明, 類似検索における逐次入れ替えによるピボット改善選択法, 日本データベース学会論文誌, 査読有, Vol.11(3), 2013, 9-14.

伏見卓恭, 齊藤和巳, 池田哲夫, 風間一洋, 異なる視点からのノード機能に基づくコミュニティ抽出法, 日本データベース学会論文誌, 査読有, Vol.11(3), 2013, 27-32

伏見卓恭, 齊藤和巳, 風間一洋, 機能性に基づくコミュニティ抽出法の比較, 情報処理学会論文誌 データベース, 査読有, Vol.5(3), 2012, 26-35

伏見卓恭, 齊藤和巳, 風間一洋, ネットワーク機能コミュニティ抽出法, 日本データベース学会論文誌, 査読有, Vol.10(3), 2012, 13-18

〔学会発表〕(計 23 件)

小林えり, 伏見卓恭, 齊藤和巳, 池田哲夫, 一般化ピボット法の距離定義による特性評価, 第6回データ工学と情報マネジメントに関するフォーラム (DEIM2014), 2014年3月5日, 兵庫県, 淡路夢舞台&ウェスティン淡路

小林えり, 伏見卓恭, 齊藤和巳, 池田哲夫, マンハッタン距離に基づく一般化ピボット計算法, 第6回Webとデータベースに関するフォーラム (WebDB2013), 2013年11月28日(木), 京都大学・百周年時計台記念館

小林えり, 伏見卓恭, 齊藤和巳, 池田哲夫, マンハッタン距離に基づく一般化ピボットによるオブジェクト可視化, 第100回知識ベースシステム研究会 (SIG-KBS), 2013年10月25日(金), 慶応義塾大学 日吉キャンパス 来住舎

小林えり, 伏見卓恭, 齊藤和巳, 池田哲夫, ソートを利用したL1ノルム総和計算によるピボット選択の高速化, 第12回情報科学技術フォーラム (FIT2013), 2013年9月6日, 鳥取大学 鳥取

小林えり, 伏見卓恭, 齊藤和巳, 池田哲夫, "一般化ピボット計算の高速化と特性評価, 第10回ネットワーク生態学シンポジウム (NETECO2013), 2013年表9月3日, 兵庫県, かんぼの宿 有馬

伏見卓恭, 齊藤和巳, 池田哲夫, 風間一洋, "情報拡散影響度に基づく機能コミュニティ抽出法, 第27回人工知能学会全国大会 (JSAI2013), 2013年6月5日,

富山国際会議場, 富山市民プラザ, 富山商工会議所
小林えり, 伏見卓恭, 齋藤和巳, 池田哲夫, 頑健線形射影法の特性評価, 第 27 回人工知能学会全国大会 (JSAI2013), 2013 年 6 月 4 日, 富山国際会議場, 富山市民プラザ, 富山商工会議所
伏見卓恭, 齋藤和巳, 郷古浩道, "Zスコアを用いたオブジェクト間関係の効果的可視化法, 第 5 回データ工学と情報マネジメントに関するフォーラム (DEIM2013), 2013 年 3 月 5 日, 福島県郡山市磐梯熱海 ホテル華の湯
小林えり, 伏見卓恭, 齋藤和巳, 池田哲夫, オブジェクト可視化と類似検索性能による頑健線形射影法の評価, 第 5 回データ工学と情報マネジメントに関するフォーラム 2013 年 3 月 4 日, 福島県郡山市磐梯熱海 ホテル華の湯
小林えり, 伏見卓恭, 齋藤和巳, オブジェクト間の非類似度の違いによる全体類似度構造の分析, 第 9 回 ネットワーク生態学シンポジウム (NETECO2012), 2012 年 12 月 15 日, 沖縄国際大学
伏見卓恭, 齋藤和巳, 池田哲夫, 風間一洋, 道路ネットワークでの機能コミュニティ抽出による地域特性分析, 第 9 回 ネットワーク生態学シンポジウム (NETECO2012), 2012 年 12 月 15 日, 沖縄国際大学
伏見卓恭, 齋藤和巳, 池田哲夫, 風間一洋, 松尾真人, 個々のノードの視点に基づく機能コミュニティ抽出法, 第 5 回 Web とデータベースに関するフォーラム (WebDB2012), 2012 年 11 月 21 日, 秋葉原コンベンションホール (秋葉原ダイビル)
伏見卓恭, 齋藤和巳, 池田哲夫, 風間一洋, 有向ネットワークにおける機能コミュニティと頑健性, 第 97 回 知識ベースシステム研究会 (SIG-KBS), 2012 年 11 月 15 日, 慶応義塾大学 日吉キャンパス 来住舎
Takayasu Fushimi, Kazumi Saito and Kazuhiro Kazama "Extracting Communities in Networks based on Functional Properties of Nodes, " Proc. of the 2012 Pacific Rim Knowledge Acquisition Workshop (PKAW2012), pp.328--334, 2012 年 9 月 6 日, Sarawak, Malaysia
伏見卓恭, 齋藤和巳, 風間一洋, 池田哲夫, 有向ネットワークに対する機能コミュニティ抽出法の検証, 第 8 回ネットワークが創発する知能研究会 (JWEIN'12), 2012 年 8 月 31 日, 立正大学大崎キャンパス
伏見卓恭, 齋藤和巳, 池田哲夫, 武藤伸明, 中心性に基づく大規模ネットワーク粗視化法の提案, 第 24 回人工知能学会

全国大会 (JSAI2012), 2012 年 6 月 12 日, 山口県, 教育会館他
三津山雅規, 齋藤和巳, 池田哲夫, 大久保誠也, 武藤伸明, 大規模データにおける逐次入れ替えによるピボット改善選択法, 第 4 回データ工学と情報マネジメントに関するフォーラム (DEIM2012), 2012 年 3 月 4 日, シーサイドホテル舞子ピラ神戸
Takayasu Fushimi, Yamato, Kubota, Kazumi Saito, Masahiro Kimura, Hiroshi Motoda, and Kouzou Ohara, "Speeding up Bipartite Graph Visualization Method," Proc. of the 24th Australasian Joint Conference on Artificial Intelligence (AI2011), 2011 年 12 月 8 日, Australia
伏見卓恭, 齋藤和巳, 風間一洋, ネットワーク機能コミュニティ抽出法の提案, Web とデータベースに関するフォーラム (WebDB Forum2011), 2011 年 11 月 5 日, 工学院大学新宿キャンパス
伏見卓恭, 齋藤和巳, 風間一洋, PageRank 影響度曲線によるコミュニティ抽出法の提案, 第 7 回ネットワークが創発する知能研究会 (JWEIN'11), 2011 年 9 月 7 日, 信州大学松本キャンパス

〔図書〕(計 0 件)

〔産業財産権〕

出願状況 (計 0 件)

取得状況 (計 0 件)

〔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・経営情報部教授
研究者番号: 60363727

(2) 研究分担者

齋藤 和巳 (SAITO, Kazumi)
静岡県立大学・経営情報部教授
研究者番号: 80379544
武藤 伸明 (MUTOH, Nobuaki)
静岡県立大学・経営情報部准教授
研究者番号: 40275102

(3) 連携研究者

なし