様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成 ２８ 年　 ６ 月　 ７ 日現在

機関番号：１７１０４
研究種目：基盤研究(B)（一般）
研究期間：2012～2015
課題番号：２４３０００６０

研究課題名（和文）高次元特徴空間の埋め込みと次元縮小に基づく知識発見基盤の構築

研究課題名（英文）Foundations of knowledge discovery based on embedding and dimension reduction in high-dimensional feature space

研究代表者

平田　耕一（Hirata, Kouchi）

九州工業大学・大学院情報工学研究院・教授

研究者番号：２０２７４５５８

交付決定額（研究期間全体）：（直接経費）　　8,800,000円

研究成果の概要（和文）：高次元特徴空間の埋め込みと次元縮小に基づく知識発見基盤の構築として，本研究では，まず，最小コストが木編集距離の変種と対応するTaiマッピングがなす階層を数理的に特徴付け，木編集距離の変種の計算に対する時間計算量を解明し，その階層に基づきさまざまな成果を得た．また，カテゴリカルデータに対する無矛盾性に基づく世界最速の特徴選択アルゴリズムであるSuper-CWCおよびSuper-LCCを設計し実装した．さらに，高次元特徴空間の近似検索に有用なヒルベルト整列に基づく索引付けを考案した．

研究成果の概要（英文）：As the foundation of knowledge discovery based on embedding and dimension reduction in high-dimensional feature space, this research characterizes mathematically a Tai mapping hierarchy consisting of mappings that provide the variations of a tree edit distance, analyzes the time complexity of computing their variations and provides the several results concerned with the hierarchy. Also this research designs and implements the fastest feature selection algorithms Super-CWC and Super-LCC based on consistency in categorical data. Furthermore, this research proposes the similarity search method in high-dimensional feature space based on Hilbert sorting.

研究分野：知能情報学

キーワード：離散構造距離　埋め込み　次元縮小　木編集距離　Taiマッピング　特徴選択　ヒルベルト整列

１．研究開始当初の背景

（１）近年のコンピュータの発展により，データから何かしらの知識を発見するためには，構造を持たないテキスト(文字列)だけでなく，HTML や XML などの半構造テキストのような(ラベル付き)木やネットワークや化学組成式といった(ラベル付き)グラフのデータといった，より複雑な離散構造のデータを扱う必要がある．

（２）このような離散構造を構造の特徴空間の次元が高いデータと考えることで，データを高次元特徴空間の要素と捉えることができる．また，音声や画像のマルチメディアデータは多くの特徴量が並んだ本質的に高次元なデータであるので，同じく，高次元特徴空間の要素と捉えることができる．

（３）高次元特徴空間のデータは，その次元数の高さから，次元が上がると質問に対する検索などの効率が指数関数的に増大する，いわゆる次元の呪いがよく知られている．この次元の呪いを避けるために，高次元データを扱う際には，次元を下げて処理することが多い．特に，高次元特徴空間での情報検索では，埋め込みや次元縮小を用いた，ある程度の不一致を許容する近似検索や類似検索の手法が用いられる．

２．研究の目的

（１）そこで本研究では，高次元特徴空間における具体的な離散構造間距離の低次元への埋め込み，および，高次元特徴空間の次元を直接縮小する次元縮小，に着目する．

（２）離散構造の埋め込みとしては，文字列に対する距離やその埋め込みについて，まずは，文字列の次に複雑な離散構造である，根付き木(順序木, 無順序木)を特徴づけ，その成果を，根無し木，有向非巡回グラフ，グラフ，超グラフといった離散構造へと拡張することを目指す．また，高次元特徴空間の次元縮小としては，任意の距離空間から L_∞距離(チェビシェフ距離)空間への次元縮小法である SimpleMap 法，および，空間充填曲線に基づく索引付けによる効率よい高次元空間の近似検索に取り組む．

３．研究の方法

（１）離散構造間距離として，文字列の編集距離の木への拡張である木編集距離について，それを特徴づける Tai マッピングについてさまざまな変種を考案してその最小コストとしての編集距離の変種とその計算量について研究を進める．特に，Tai マッピングの階層を新たに特徴づける．また，離散構造のうち，最も一般的な超グラフから，それを比較するための距離やカーネルの基礎づけとして，木に対応する非巡回部分超グラフによる特徴づけに取り組む．

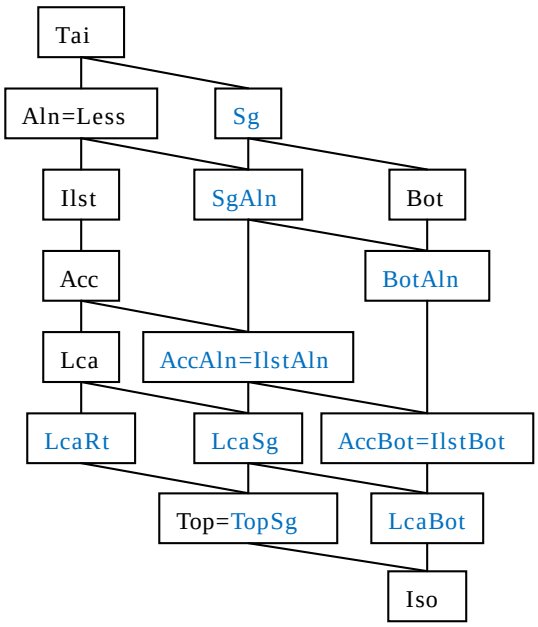
（２）離散構造を扱う際，そのデータを何らかの特徴に基づいたヒストグラムなどの高次元ベクトルで表現することで，カーネルを用いた機械学習に適用することができる．そこで，離散構造のカーネルについても研究を進める．

（３）高次元特徴空間のデータにおける一種の次元縮小として，データ全体を説明する次元軸を選択する特徴選択に対して，高速な手法を開発する．

（４）高次元特徴空間の近似検索における SimpleMap による精度向上の手法，および，空間充填曲線の中でもヒルベルト曲線に着目した高次元特徴空間の索引付け手法を開発する．

４．研究成果

（１）木編集距離を特徴づける木のノード間の Tai マッピングに対して，数理的に有意な特徴付けに基づく Tai マッピング階層を明確にし，その階層におけるマッピングの最小コストである木編集距離の変種の計算に対する時間計算量を解明した．以下がその階層であり，青字が新たに導入したマッピングである．



Tai マッピング階層の列は，マッピングによって対応するノードがなす木の部分森の形状を表しており，左の列が埋め込み部分森，中の列は誘導部分森，右の列は完全部分森となる．また，Tai マッピング階層の行は，その部分森における木の配置方法を表しており，1 行目(Tai，Sg，Bot)は任意の部分森，2 行目(Aln，SgAln，BotAln)はねじれ無し部分森，3 行目(Ilst,Acc,AccAln,AccBot)は並行部分森，4 行目(Lca，LcaSg，LcaBot)は部分木，5 行目(LcaRt，Top，Iso)が根保存部分木となる．

さらに，時間計算量についても，Tai マッピング階層と下記のように対応している．ここで，n を二つの木の最大ノード数，D を最大次数，d を最小次数とする．

①順序木では，Tai は $O(n^3)$時間，2 行目(Aln，SgAln，BotAln)は $O(n^2D)$時間，Iso は $O(n)$時間，それ以外はすべて $O(n^2)$時間となる．

②無順序木では，1 行目(Tai，Sg，Bot)は最小次数が 2 であっても MAX SNP 困難，2 行目(Aln，SgAln，BotAln)は一般には MAX SNP 困難だが次数制限の場合は多項式時間で計算可能，LcaBot が $O(n^2)$時間，Iso が $O(n)$時間で，それ以外はすべて $O(n^2d)$時間となる．

（2）離散構造間の距離として，主に木を比較する手法を中心に，（1）以外でさらに下記の研究成果を得た．

①Tai マッピング階層の Aln に対応するアライメント距離を計算する新たな手法として，マッピングをアンカーとして与えるアンカーアライメントを定式化し，順序木，無順序木と共に $O(H|M|^2+n)$時間で計算するアルゴリズムを設計した．ここで，H は木の高さの最大値，$|M|$はアンカーの要素数である．

②子のノードの巡回性を考慮した，順序木より一般的で無順序木に制限を入れた両順序木，巡回順序木，巡回両順序木の概念を導入し，両順序木のアライメント距離が $O(n^2D^2)$時間，巡回順序木と巡回両順序木のアライメント距離が $O(n^2D^4)$時間で計算できるアルゴリズムを開発した．なお，これらの木における編集距離の計算は MAX SNP 困難である．

③木包含問題とは，一方の木(テキスト木 T)からノードを削除することでもう一方の木(パターン木 P)になるか否かを判定する問題であり無順序木の場合は一般に NP 完全である．それに対して，Tai マッピングの階層における Lca に対応する次数 2 包含問題は $O(|T||P|^{3/2})$時間 $O(|T||P|)$領域で解ける．そこで，この問題を Tai マッピング階層の Ilst に対応する孤立部分木包含問題にまで拡張し，次数 2 包含問題と同一の計算量で解くアルゴリズムを設計した．

④統計的手法である PCA を木に適用した木 PCA の手法を開発し，その有用性を示した．

⑤木以外の離散構造として，超グラフにおける木に対応する非巡回超グラフに対して，超グラフに含まれるベルジュ非巡回部分超グラフの多項式時間遅延で列挙可能であることを示した．

（3）2 つの木に対する Tai マッピング(及びその変種)を数え上げることで，木カーネルの一種であるマッピングカーネルを設計することができるが，マッピングカーネルの計算は一般の無順序木では#P 完全となる．木カーネルについて，下記の研究成果を得た．

①生物種の進化を表す進化系統樹，すなわち，次数が 2 で葉にしかラベルが付いていない無順序二分葉ラベル木に対して，Lca の一種である合致部分木マッピングカーネルを設計し，それが正定値となること，および，$O(n^2)$時間で計算するアルゴリズムを設計した．この結果は次数が制限された葉ラベル木でも成り立つ．一方，次数が制限されていない場合は，葉ラベル木であっても合致部分木マッピングカーネルの計算は#P 完全となる．

②別の進化系統樹のカーネルとして，それに含まれている葉間パスをすべて数え上げる葉間パスカーネルを提案した．

③通常のラベル付き木に対して，（2）②の巡回的順序木における，Tai マッピング階層に対応する Top，LcaSg，Lca，Acc，Ilst のマッピングカーネルを $O(n^2dD)$時間で計算するアルゴリズムを設計した．一方，最下層となる Top でさえ無順序木マッピングカーネルの計算は#P 完全となる．

（4）カテゴリカルデータに対するフィルター型の特徴選択法として，世界最速のアルゴリズムである Super-CWC および Super-LCC を設計し，実装した．既存のベンチマークデータを用いて，速度と分類精度の両面で既存手法を凌ぐ性能が得られることを示した．そして，それらを大量の Twitter データからのトピック語抽出に適用し，有用性を示した．

（5）高次元特徴空間の近似検索における SimpleMap による精度向上手法として，っ最大値と最小値を利用する二値量子化に基づく SimpleMap の中心点探索手法を開発した．また，高次元特徴空間のオブジェクトを，曲線を生成せずに効率よくヒルベルト曲線順に整列する手法であるヒルベルト整列を開発し，それに基づく索引を利用した動画像データや音声データの近似検索に適用することでその効果を検証した．さらに，高次元近似検索における一括質問手法を設計し，その効率を検証した．

（6）高次元ベクトル空間からの知識発見の応用として以下のような研究成果を得た．

①100 以上の薬剤の感受性検査結果が報告されている薬剤感受性検査データから，菌の伝搬パターンを抽出するアルゴリズムを設計し，実装すると共に，医学的に有意なパターンを抽出した．

②インフルエンザウイルスの塩基配列に対

して，進化系統樹に基づいた塩基配列の位置の関係を比較する剪定距離を導入すると共に，これに加えて（３）①の合致部分木マッピングカーネルや②の葉間パスカーネルを利用し，インフルエンザウイルスの地域間遷移やパッケージングシグナル位置を解析した．

③単語の共起ネットワークが成す階層的コミュニティ構造の時系列変化を，木の編集距離を用いて追跡することにより，コミュニティ構造の遷移を抽出する手法を開発した．さらに，本手法をバズマーケティングサイトの掲示板データに適用し，利用者の関心の時系列遷移を抽出した．

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 21 件）

1. K. Nagayama, K. Hirata, S. Yokoyama, K. Matsuoka, Extracting Propagation Patterns from Bacterial Culture Data in Medical Facility, Lecture Notes in Artificial Intelligence, 2016 (採録決定), 査読有.

2. K.Hirata, T. Kuboyama, T. Yoshino: Mapping Kernels between Rooted Labeled Trees beyond Ordered Trees, Lecture Notes in Artificial Intelligence 9067, 317-330, 2015, 査読有, DOI 10.1007/978-3-662-48119-6_24.

3. Y. Ishizaka, T. Yoshino, K.Hirata: Anchored Alignment Problem for Rooted Labeled Trees, Lecture Notes in Artificial Intelligence 9067, 296-309, 2015, 査読有, DOI 10.1007/978-3-662-48119-6_22.

4. Q. Jin, M. Nakashima, T. Shinohara, K. Hirata, T. Kuboyama: Central Point Selection for Dimension Reduction Projection Simple-Map with Binary Quantization, Lecture Notes in Artificial Intelligence 9067, 310-316, 2015, 査読有, DOI 10.1007/978-3-662-48119-6_23.

5. T. Yamazaki, A. Yamamoto, T. Kuboyama: Tree PCA for Extracting Dominant Substructures from Labeled Trees, Lecture Notes in Artificial Intelligence 9356, 316-323, 2015, 査読有, DOI 10.1007/978-3-319-24282-8_27.

6. R. Saito, T. Kuboyama, H. Yasuda: User Behaviour Modelling by Abstracting Low-Level Window Transition Logs, International Journal of Computational Science and Engineering 11, 249-258, 2015, 査読有, DOI 10.1504/IJCSE.2015.072648.

7. K. Shin, T. Kuboyama, T. Hashimoto, D. Shepard: Super-CWC and Super-LCC: Super Fast Feature Selection Algorithms, IEEE Big Data, 1-7, 2015, 査読有, DOI 10.1109/BigData.2015.7363742

8. T. Yoshino, K. Hirata: Alignment of Cyclically Ordered Trees, Proc. 4th International Conference on Pattern Recognition Applications and Methods (ICPRAM2015), 263-270, 2015, 査読有, DOI 10.5220/0005207802630270.

9. I. Hamada, T. Shimada, D. Nakata, K. Hirata, T. Kuboyama: Classifying Nucleotide Sequences and Their Positions of Influenza A Viruses through Several Kernels, Proc. 4th International Conference on Pattern Recognition Applications and Methods (ICPRAM2015), 342-347, 2015, 査読有, DOI 10.5220/0005251103420347.

10. Y. Nasu, N. Kishikawa, K. Tashima, S. Kodama, Y. Imamura, T. Shinohara, K. Hirata, T. Kuboyama: High Dimensional Similarity Search with Bundled Query Processing on Hilbert R-Tree, 4th International Conference on Pattern Recognition Applications and Methods (ICPRAM2015), 354-359, 2015, 査読有, DOI 10.5220/0005279503540359.

11. T. Kan, S. Higuchi, K. Hirata: Segmental Mapping and Distance between Rooted Labeled Ordered Trees, Fundamenta Informaticae 132(4), 461--483, 2014, 査読有, DOI 10.3233/FI-2014-100.

12. Y. Yamamoto, K. Hirata, T. Kuboyama: Tractable and Intractable Variations of Unordered Tree Edit Distance, International Journal of Foundations of Computer Science 25(3), 307-329, 2014, 査読有, DOI 10.1142/S0129054114500154.

13. I. Hamada, T. Shimada, D. Nakata, K. Hirata, T. Kuboyama: Agreement Subtree Mapping Kernel for Phylogenetic Trees, Lecture Notes in Artificial Intelligence 8417, 321-336, 2014, 査読有, DOI 10.1007/978-3-319-10061_21.

14. K. Shin, T. Kuboyama: A Comprehensive Study of Tree Kernels, Lecture Notes in Artificial Intelligence 8417, 337-351, 2014, 査読有, DOI 10.1007/978-3-319-10061_22.

15. T. Miyahara, T. Kuboyama: Learning of Glycan Motifs Using Genetic Programming and Various Fitness Functions, Journal of Advanced Computational Intelligence and Intelligent Informatics 18, 401-408, 2014, 査読有, DOI 10.20965/jaciii.2014.p0401.

16. T. Hashimoto, B. Chakraborty, T. Kuboyama, T. Shirota: Temporal Awareness of Needs after East Japan Great Earthquake Using Latent Semantic Analysis, Information Modelling and Knowledge Bases 25, 200-214, 2014, 査読有, DOI 10.3233/978-1-61499-360-5.

17. K. Wasa, T. Uno, K. Hirata, H. Arimura: Polynomial Delay and Space Discovery of Connected and Acyclic Sub-Hypergraphs in a Hypergraph, Lecture Notes in Artificial Intelligence 8140, 308-323, 2013, 査読有, DOI 10.1007/978-3-642-40897-7.

18. K. Wasa, K. Hirata, T. Uno, H. Arimura: Faster Algorithms for Tree Similarity Based on Compressed Enumeration of Bounded-Sized Ordered Subtrees, Lecture Notes in Computer Science 8199, 73-84, 2013, 査読有, DOI 10.1007/978-3-642-41061-1.

19. S.Makino, T.Shimada, K.Hirata, K.Yonezawa, K.Ito; A Trim Distance between Positions in Nucleotide Sequences, Lecture Notes in Artificial Intelligence 7569, 81-92, 2012, 査読有, DOI 10.1007/978-3-642-33492-4.

20. T.Hashimoto, T.Kuboyama, Y.Shirota: Topic Detection about the East Japan Great Earthquake based on Emerging Modularity, Frontiers in Artificial Intelligence and Applications 215, 110-126, 2012, 査読有, DOI 10.3233/978-1-61499-177-9-110.

21. M. Nakahara, S. Maruyama, T. Kuboyama, H. Sakamoto: Scalable Detection of Frequent Substrings by Grammar-Based Compression, IEICE Transactions 96-D, 457-464, 2012, 査読有, http://search.ieice.org/bin/summary.php?id=e96-d_3_457

〔学会発表〕(計 62 件)

1. T. Hashimoto, D. Shepard, T. Kuboyama, K. Shin: Event Detection from Millions of Tweets Related to the Great East Japan Earthquake Using Feature Selection Technique, 1st International Workshop on Event Analytics Using Social Media Data, 2015 年 11 月 14 日, Atlantic City (USA).

2. T. Yoshino, K. Hirata: Alignment of II-Ordered Trees, Workshop on Graph-Based Algorithms for Big Data and Its Application (GABA2014), 2014 年 11 月 23 日, 慶應義塾大学 (神奈川県横浜市).

3. I. Hamada, K. Hirata, T. Kuboyama, T. Shimada: Agreement-Subtree Mapping Kernel and Leaf-Path Kernel for Phylogenetic Trees Reconstructed from Nucleotide Sequences, Workshop on Graph-Based Algorithms for Big Data and Its Application (GABA2014), 2014 年 11 月 23 日, 慶應義塾大学 (神奈川県横浜市).

4. K. Shin, T. Kuboyama: De Morgan Property of Bayes Risk as A Feature Selection Measure, Workshop on Graph-Based Algorithms for Big Data and Its Application (GABA2014), 2014 年 11 月 23 日, 慶應義塾大学 (神奈川県横浜市).

5. T. Yamazaki, K. Otaki, M. Ikeda, A. Yamamoto, T. Kuboyama: Local Similarity between Semi-Ordered Trees by Finding the Constrained Mapping, Workshop on Graph-Based Algorithms for Big Data and Its Application (GABA2014), 2014 年 11 月 23 日, 慶應義塾大学 (神奈川県横浜市).

6. T. Yoshino, K. Hirata: Hierarchy of Segmental and Alignable Mapping for Rooted Labeled Trees, Workshop on Data Discretization and Segmentations for Knowledge Discovery (DDS13), 2013 年 10 月 27 日, 慶應義塾大学 (神奈川県横浜市).

7. T. Takabatake, T. Kuboyama, A. Yasuhara, H. Sakamoto: Fast Computation of Invariant on Knot Theory, Workshop on Data Discretization and Segmentations for Knowledge Discovery (DDS13), 2013 年 10 月 27 日, 慶應義塾大学 (神奈川県横浜市).

8. Y. Li, T. Kuboyama, H. Sakamoto: An implementation of Truss Decomposition of Bipartite Graph, Workshop on Data Discretization and Segmentations for Knowledge Discovery (DDS13), 2013 年 10 月 27 日, 慶應義塾大学 (神奈川県横浜市).

9. Y. Li, T. Kuboyama, H. Sakamoto: Truss Decomposition for Extracting Communities in Bipartite Graph, 3rd International Conference on Advances in Information Mining and Management, 2013 年 11 月 17 日～22 日, Lisbon (Portgal).

10. T. Shimada, I. Hamada, K. Hirata, T.

Kuboyama, K. Yonezawa, K. Ito: Clustering of Positions in Nucleotide Sequences by Trim Distance, IIAI International Conference on Advanced Informatics (IIAI AAI 2013), 2013年8月31日～9月4日, くにびきメッセ (島根県松江市).

11. T. Yoshino, S. Higuchi, K. Hirata; A Dynamic Programming A* Algorithm for Computing Unordered Tree Edit Distance, IIAI International Conference on Advanced Informatics (IIAI AAI 2013), 2013年8月31日～9月4日, くにびきメッセ (島根県松江市).

12. S. Higuchi, T. Hashimoto, T. Kuboyama, K. Hirata: Exploring Social Context from Buzz Marketing Site - Community Mapping Based on Tree Edit Distance -4th International Workshop on Pervasive Collaboration and Social Networking (PerCol2013), 2013年3月13日, San Diego(USA).

13. S. Makino, T. Shimada, K. Hirata, K. Yonezawa, K. Ito: A Trim Distance between Positions as Packaging Signals in H3N2 Influenza Viruses, 6th International Conference on Soft Computing and Intelligent Systems & 13th International Symposium on Advanced Intelligent Systems (SCIS-ISIS 2012), 2012年11月20日～24日, 神戸コンベンションセンター (兵庫県神戸市)

14. T. Shimada, T. Hazemoto, S. Makino, K. Hirata, K. Ito: Finding Correlated Mutations among RNA Segments in H3N2 Influenza Viruses, 6th International Conference on Soft Computing and Intelligent Systems & 13th International Symposium on Advanced Intelligent Systems (SCIS-ISIS 2012), 2012年11月20日～24日, 神戸コンベンションセンター (兵庫県神戸市)

15. T. Miyahara, T. Kuboyama: Acquisition of glycan motifs using genetic programming and various fitness function, 6th International Conference on Soft Computing and Intelligent Systems & 13th International Symposium on Advanced Intelligent Systems (SCIS-ISIS 2012), 2012年11月20日～24日, 神戸コンベンションセンター (兵庫県神戸市)

16. K. Shin, T. Kuboyama: Dynamic labeling and tree kernels with gap penalties. 6th International Conference on Soft Computing and Intelligent Systems & 13th International Symposium on Advanced Intelligent Systems (SCIS-ISIS 2012), 2012年11月20日～24日, 神戸コンベンションセンター (兵庫県神戸市)

17. T. Hokazono, T. Kan, Y. Yamamoto, K. Hirata: An Isolated-Subtree Inclusion for Unordered Trees, IIAI International Conference on Advanced Applied Informatics (IIAI-AAI 2012), 2012年9月20日～22日, 九州大学 (福岡県福岡市).

18. T. Hashimoto, T. Kuboyama, B. Chakraborty, Y. Shirota: Discovering Topic Transition about the East Japan Great Earthquake in Dynamic Social Media, Global Humanitarian Technology Conference (IEEE GHTC), 2012年10月21日～24日, Seattle (USA)

19. T. Hashimoto, T. Kuboyama, B. Chakraborty, Y. Shirota: Discovering emerging topic about the East Japan Great Earthquake in video sharing website, 2012 IEEE Region 10 Conference (TENCON 2012), 2012年11月19日～22日, Cebu (Phillipines)

20. K. Shin, T. Kuboyama, H. Nishimura: A new consistency-based feature selection algorithm, International Conference on Soft Computing (MENDEL), 2012年7月27日～29日, Brno (Czech Republic)

〔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〕
ホームページ等

６．研究組織
平田　耕一 (HIRATA Kouichi)
九州工業大学・大学院情報工学研究院・教授
研究者番号：２０２７４５５８

(2)研究分担者
篠原　武 (SHINOHARA Takeshi)
九州工業大学・大学院情報工学研究院・教授
研究者番号：６０１５４２２５

久保山　哲二 (KUBOYAMA Tetsuji)
学習院大学・計算機センター・教授
研究者番号：８０３０２６６０