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　１３　日現在

機関番号：　１７４０１
研究種目：　若手研究(B)
研究期間：　2014～2016
課題番号：　２６７３００６０

研究課題名（和文）大規模時系列グラフデータのためのモデル学習と将来予測

研究課題名（英文）Mining and Forecasting of Big Time-evolving Events

研究代表者
松原　靖子（Matsubara, Yasuko）

熊本大学・大学院先端科学研究部（工）・助教

研究者番号：００７２１７３９

交付決定額（研究期間全体）：（直接経費）　　2,900,000円

研究成果の概要（和文）：本研究では，これらの多種多様な大規模時系列グラフビッグデータを対象とし，様々な現象，活動の時間的変遷とグラフ構造との関連性の発見，モデル化を行なうと同時に，将来の活動の予測を効果的，効率的に行なうことにより，様々な場面で利用可能な解析技術の研究開発に取り組んだ．具体的には，センサネットワーク，伝染病の拡散過程，ソーシャルネットワーク上のユーザ活動パターンを始めとする，高度な構造を持つ様々な時系列ビッグデータを対象とし，その特徴やダイナミクスをモデル化，予測するための基盤技術を開発した．

研究成果の概要（英文）：Time-evolving event analysis is becoming of increasingly high importance, thanks to the decreasing cost of hardware and the increasing on-line processing capability. In such a situation, the most fundamental requirement is an efficient modeling and mining of event streams. This research project addresses three classes of tasks for time-evolving event analysis, namely, (1) automatic mining, (2) non-linear modeling and (3) large-scale tensor analysis. We developed powerful algorithms that provide efficient and effective mining of large-scale time evolving events.

研究分野：　データベース，データマイニング

キーワード：　ソーシャルネットワーク　非線形解析　特徴自動抽出　テンソルデータ　将来予測

１．研究開始当初の背景

現在の高度に発達した情報化社会において，我々が取り扱うことのできるデータの量は飛躍的に増大している．そのような状況下において，現在特に注目されているのが，大規模時系列グラフデータである．時系列グラフデータは，Twitter に代表されるソーシャルネットワーク上におけるユーザ間の繋がり，コンピュータウィルスの伝染等の Web 上に発生する現象を始め，世界規模の伝染病の拡散過程，センサネットワークや IoT デバイスによる複合データストリーム等，様々な分野で大量に発生している．例えば，Web サービスの代表である Google の検索数は毎日 30 億クエリを超え，ソーシャルメディアサービス Twitter には，3 億近くのアクティブユーザから 5 億件以上の Tweet が日々生成され続けている．これらのオンライン活動データは，現実世界のニュース，季節性等を含む様々な実社会活動と連動し，リアルタイムに変化，推移している．このような大規模時系列グラフの時間発展の解析は，特定のビジネスのみならず，社会経済の活性化，行政，環境，防災など，重要な社会問題を解決するための効果的なアプローチとして期待されている．

２．研究の目的

本研究では，これらの多種多様な大規模時系列グラフビッグデータを対象とし，様々な現象，活動の時間的変遷とグラフ構造との関連性の発見，モデル化を行なうと同時に，将来の活動の予測を効果的，効率的に行なうことにより，様々な場面で利用可能な解析技術の研究開発に取り組む．具体的には，センサネットワーク，伝染病の拡散過程，ソーシャルネットワーク上のユーザ活動パターンを始めとする，高度な構造を持つ様々な時系列ビッグデータを対象とし，その特徴やダイナミクスをモデル化，予測するための基盤技術を開発することを目的とする．

３．研究の方法

本研究で扱う時系列グラフデータは，構造や現象が複雑であり，従来のデータ解析手法では表現できないダイナミクスが含まれている．そこで本研究では，高度な時系列ビッグデータを解析するための新たな基礎技術として，以下の解析手法を検討した．

(1) 自動特徴抽出

実用的な技術やシステムの開発を考えた際，最も重要である課題が特徴抽出の完全自動化である．例えば，自己回帰やフーリエ変換，特異値分解をはじめとする従来の時系列解析では，使用する係数の個数や閾値等のパラメータの設定が必要である．しかしセンサデータのように，解析するデータが膨大になるほど，専門のエンジニアによる細やかなパラメータチューニングには時間的，金銭的コストが多くなり，重大なボトルネックとなる．このようにビッグデータの解析において，解析技術の自動化は極めて重要な課題である．

(2) 非線形時系列解析

非線形モデルは，疫学，生物学，物理，経済など，様々な分野で用いられている．一方で，データマイニング分野においては，ソーシャルメディアやオンラインユーザアクティビティの分析が盛んであり，一部の研究ではソーシャルメディアのダイナミクスをとらえるために非線形モデルが用いられている．このような研究が行なわれている中で，非線形モデルとテンソルを統合した非線形テンソル解析技術を考案する．時系列ビッグデータを多角的に，かつ非線形性を考慮しながら分析することにより，各シーケンス間の潜在的な関係性をとらえることが可能となる．

(3) 大規模テンソル解析

時間発展の情報を含むデータの多くは，テンソルとして表現することができる．例えば，データセンタにおいて各計算機からいくつものセンサデータをモニタリングしている場合，(timestamp, sensor ID, machine ID, …)という m 個の属性からなるレコード群は，m 階のテンソルとして表現することができる．時間情報をともなう大規模なテンソル，すなわち時系列ビッグデータテンソルの解析技術は今後重要な要素技術の一つとなる．

４．研究成果

本研究ではセンサネットワーク，伝染病の拡散過程，ソーシャルネットワーク上のユーザ活動パターンを始めとする様々なドメインを対象として，情報抽出，パターン検出，モデル学習，将来予測の研究を行なった．以下において，各要素技術に対する研究成果の詳細を示す．

4.1 大規模時系列シーケンスからの特徴自動抽出

大規模な時系列シーケンスの中から，典型的なパターンや異常値を発見することは非常に重要な課題である．考案手法である AutoPlait [C8] は，大規模時系列データを対象とし，重要な時系列パターンの抽出を自動的に行なうことができる．特に，多次元時系列シーケンスを扱い，これらのデータ全体を表現する要約情報を抽出する．

4.1.1 提案手法とその特徴

図 1 はモーションキャプチャセンサデータに対する AutoPlait の出力結果例である．この時系列データは，「チキンダンス (chicken dance)」と呼ばれるダンスのステップを表現している．このモーションは，4 次元のシーケンスで構成され，それぞれの次元が左右の腕と足の加速度を表現している．チキンダンスは，beaks, wings, tail-feathers, claps の 4 つの代表的なステップ から構成されている．ここで，ステップのような部分シーケンスの特徴パターンをレジームと呼び，この時系列データは 4 つのレジームから構成され

る．図 1 の下段は，AutoPlait が自動抽出した 4 つのレジームを示している．提案手法は，ダンスに含まれる 4 つのレジームを抽出し，そして各レジームの切れ目も正しく発見することができる．AutoPlait は，これらレジームに関する事前知識を必要とせず，適切な数のレジームとその位置を自動的に把握することができる．

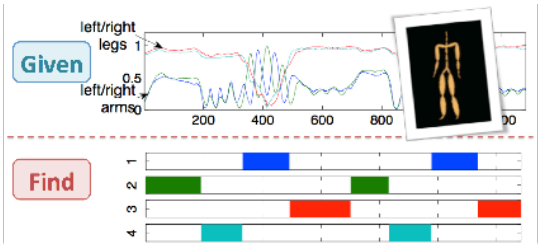


図 1 モーションキャプチャからの特徴抽出

AutoPlait の目的は，与えられた時系列シーケンス群 X の特徴を抽出し，すべての時系列パターンを表現するパラメータ集合 C = {m, r, S, Θ, F} を発見することである．ここで，m はセグメント数，r はレジーム数，S はセグメント，Θはレジーム，F セグメントメンバシップであり，これらを AutoPlait は自動的に抽出する．以下では，AutoPlait の 2 つの重要なアイディアについて説明する．

(1) 多階層連鎖モデル（MLCM: multi-level chain model）

複数のレジーム間の時系列パターンとその遷移を表現するために，多層的な連鎖モデル（MLCM）を提案する．提案モデルである MLCM は隠れマルコフモデル（HMM: Hidden Markov Model）を拡張しており，従来の HMM の遷移確率に加え，上位層の状態（super-state）の概念を導入することによって，パターンのグループ化を行なう．

(2) モデル表現コストとデータ圧縮

適切なセグメントとレジームの検出のため，最小記述長（MDL: minimum description length）の概念を用いる．MDL は情報理論に基づくモデル選択基準のひとつで，可逆圧縮を行なうことができる．本研究では，与えられた時系列シーケンス X を適切に表現するモデルを見つけるために，新しい符号体系を定義した．具体的には，(a) 最適なパラメータ集合 C を推定するためのコスト関数を定義し，(b) 最適解を発見するための効果的なアルゴリズムを提案する．

4.1.2 アルゴリズム

図 2 は，AutoPlait の最適化アルゴリズムの概要を示す．提案アルゴリズムは，次に挙げる 3 つの部分問題に分割される．

Algorithm 1, CutPointSearch: レジームの個数（r = 2）とモデルパラメータが与えられたとき，X を 2 つのレジームに分割し，それぞれのセグメントの分割位置を検出する．

Algorithm 2, RegimeSplit: レジームの個数 r = 2 が与えられたときに，2 つのレジームを表現するモデルパラメータ（θ_1, θ_2, Δ）を推定する．

Algorithm 3, AutoPlait: 最適なレジームの個数（r = 2, 3, 4, …）を求める．

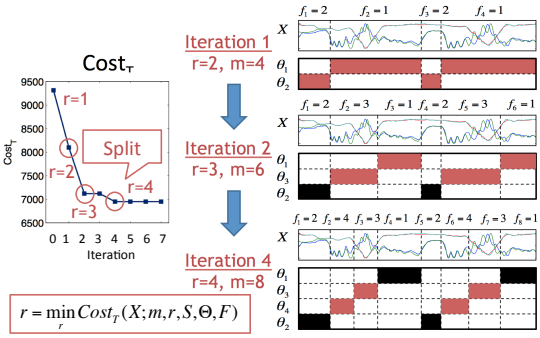


図 2 提案アルゴリズムの概要

4.2 非線形時系列解析

4.2.1 RegimeCast: 時系列データストリームのリアルタイム予測

RegimeCast[c2] は，センサデータをはじめとする大規模な時系列データストリームに対し，リアルタイムに将来のパターンを予測し続ける技術である．図 3 上段はオリジナルセンサストリーム，中段は 100 ステップ先の予測結果，下段は各時刻におけるリアルタイム予測のスナップショットを示す．

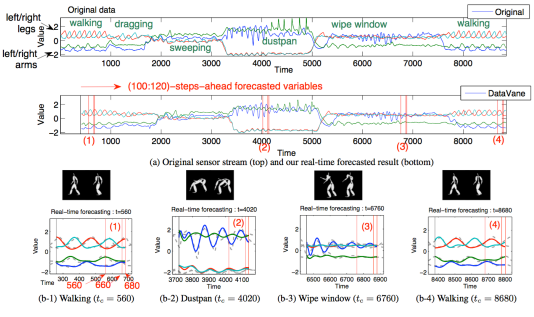


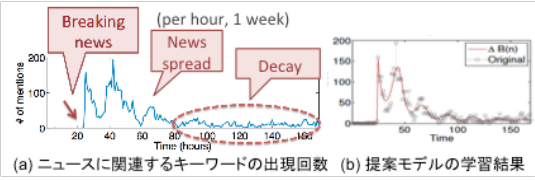
図 3 リアルタイム予測の概要

4.2.2 SpikeM: ソーシャルネットワーク上の情報拡散過程と非線形解析

SpikeM [J1] は，Blog や Twitter を始めとするソーシャルネットワーク上において，噂やニュース等の情報が，時間が経過するごとに，どのように拡散し減衰していくかを表現する（図 4 参照）．提案手法は，ユーザ間のグラフ構造に基づく情報拡散過程をパワー則に基づく非線形モデルで表現し，将来の情報拡散過程を予測することができる．

4.2.3 EcoWeb: 生態系モデルに基づくオンライン活動上の競合関係自動抽出

EcoWeb [C4] は生態系における種間競争モデルに基づき，Web 上のユーザ活動の中から潜在的な競合関係や季節性等の重要パターンを自動抽出する非線形モデルである．



(a) ニュースに関連するキーワードの出現回数 (b) 提案モデルの学習結果

図 4 Blog 上の情報拡散過程とモデル学習

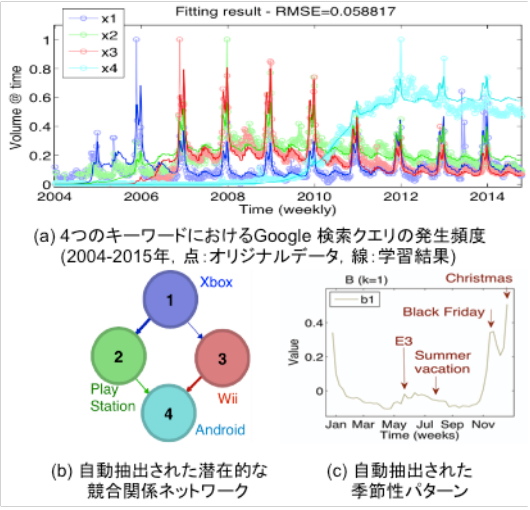
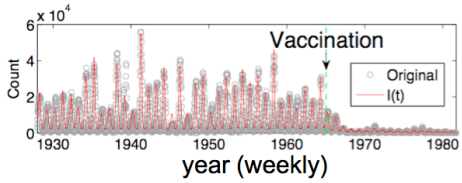


図 5 Google 上の活動パターンと特徴抽出

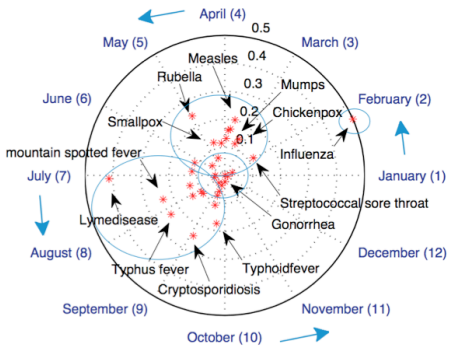
4.3 大規模テンソル解析

4.3.1 FUNNEL：大規模疫病テンソルデータのための非線形解析モデル

FUNNEL［C8］は大規模な疫病感染データ（timestamp, disease, location）に対し，非線形モデル学習とテンソル解析技術を融合し，疫病の感染力，季節性，地域性，ワクチン効果等の重要なパターンを自動抽出する統合モデルである．



(a) はしかの患者数の推移とFUNNELの学習結果



(b) 主要な疫病の季節性

図 6 大規模疫病テンソルの非線形解析

4.3.2 CompCube：複合テンソルデータに基づくオンライン活動自動解析

CompCube[c3]は，Web 上のユーザの地域別の活動データ（time, activity, location）の中から基本情報，競合関係，季節性，外れ値の 4 つの情報を，Global/Local 両視点から抽出する統合モデルである．例えば，各国における Google 検索件数のデータが与えられたとき，図のように提案手法は，競合関係（Kindle vs. Nexus 等），地域別季節性（Christmas, Chinese New Year）等の情報を完全自動で抽出し，各地域における今後のユーザの行動の予測を行うことができる．

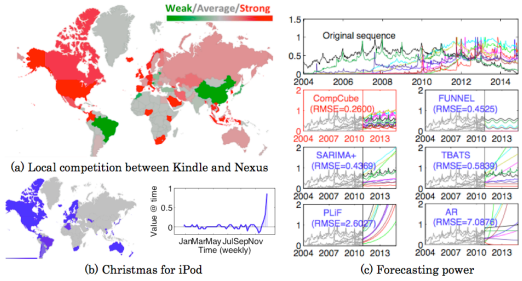


図 6 オンライン活動における特徴自動抽出

本取り組みでは，多種多様な大規模時系列グラフビッグデータを対象とし，様々な現象，活動の時間的変遷とグラフ構造との関連性の発見，モデル化するための研究を行った．実データを用いた実験により，センサネットワーク，伝染病の拡散過程，ソーシャルネットワーク上のユーザ活動パターンを始めとする，高度な構造を持つ様々な時系列ビッグデータに対し，その特徴を柔軟にモデル化，解析するための技術を開発した．

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　8 件）

j1. Yasuko Matsubara, Yasushi Sakurai, B. Aditya Prakash, Lei Li, Christos Faloutsos, "Non-linear Dynamics of Information Diffusion in Social Networks", ACM Transactions on the Web (TWEB), Volume 11 Issue 2, 10.1145/3057741, May 2017 **査読有.**

j2. Y. Matsubara, Y. Sakurai, C. Faloutsos: ``Ecosystem on the Web: non-linear mining and forecasting of co-evolving online activities", World Wide Web Journal, Springer, Volume 20, Issue 3, pp439-465, 10.1007/s11280-016-0389-x, 2016, **(invited paper) 査読有.**

j3. 松原靖子，櫻井保志，Christos Faloutsos：``生態系モデルに基づくオンライン活動データの非線形解析″，**電**

気情報通信学会論文誌 D, Vol. J100-D, No. 4, pp. 457-471, 2017 **（研究会推薦論文）査読有**.

j4. 松原靖子, 櫻井保志:``大規模データストリームのリアルタイム予測", **情報処理学会論文誌：データベース(TOD)** Vol.9 No.4, pp. 32-45, 2016, 12 月 22 日, **査読有**.

j5. 松原靖子, 櫻井保志, Willem G. van Panhuis, Christos Faloutsos:``大規模疫病データのための非線形モデル解析", **情報処理学会論文誌：データベース(TOD)** Vol.9 No.4, pp. 17-31, 2016, 12 月 22 日, **査読有**.

j6. 本田崇人, 松原靖子, 根山亮, 櫻井保志:``車両走行センサデータからの自動パターン検出", **情報処理学会論文誌：データベース(TOD)**, 9(3), 1-13 (2016) **研究会推薦論文**, **査読有**.

j7. T. M. Do, Y. Matsubara and Y. Sakurai: ``Non-linear Time-series Analysis of Social Influence", **情報処理学会論文誌：データベース(TOD)**, 24(6), 9 pages (2016), **査読有**.

j8. 松原靖子, 櫻井保志, Christos Faloutsos: ``大規模時系列データの特徴自動抽出", **情報処理学会論文誌：データベース**, Vol. 7, No. 2, pp. 37-50, 2014 年 6 月, **査読有**.

〔学会発表〕（計　8+3+11=22 件）

○国際会議

c1. Thinh Minh Do, Yasuko Matsubara, Yasushi Sakurai, "Automatic and Effective Mining of Coevolving Online Activities", Pacific-Asia Conference on Knowledge Discovery and Data Mining **(PAKDD)**, Jeju, South Korea, May 23-26 , 2017, **査読有**.

c2. Y. Matsubara, Y. Sakurai: ``Regime Shifts in Streams: Real-time Forecasting of Co-evolving Time Sequences", ACM SIGKDD International Conference on Knowledge Discovery and Data Mining **(KDD)**, San Francisco, California, August 13-17, 2016, acceptance Rate: 70/784, **8.9%, 査読有**.

c3. Y. Matsubara, Y. Sakurai, C. Faloutsos: ``Non-Linear Mining of Competing Local Activities", International World Wide Web Conference **(WWW)** Montreal, Canada, April 11-15, 2016, acceptance Rate: 115/727, **15.8%, 査読有**.

c4. Y. Matsubara, Y. Sakurai, C. Faloutsos: ``The Web as a Jungle: Non-Linear Dynamical Systems for Co-evolving Online Activities", International World Wide Web Conference **(WWW)**, Florence, Italy, May 18-22, 2015., acceptance Rate: 131/929, **14.1%, 査読有**.

c5. Y. Matsubara, Y. Sakurai, N. Ueda, M. Yoshikawa: ``Fast and Exact Monitoring of Co-evolving Data Streams", IEEE International Conference on Data Mining **(ICDM)**, Shenzhen, China, December 14-17, 2014, acceptance rate **9.7%, 査読有**.

c6. F. Figueiredo, J. M. Almeida, Y. Matsubara, B. Ribeiro, C. Faloutsos, "Revisit Behavior in Social Media: The Phoenix-R Model and Discoveries", **ECML/PKDD2014**, pp. 386-401, Nancy, France, September 15-19, 2014 (full presentation) 115/483, **23.8%, 査読有**.

c7. Y. Matsubara, Y. Sakurai, W. G. van Panhuis, C. Faloutsos: ``FUNNEL: Automatic Mining of Spatially Coevolving Epidemics", ACM SIGKDD International Conference on Knowledge Discovery and Data Mining **(KDD)**, pp.105-114, New York City, August 24-27, 2014, acceptance rate 151/1036,**14.5%査読有**.

c8. Y. Matsubara, Y. Sakurai, C. Faloutsos: ``AutoPlait: Automatic Mining of Co-evolving Time Sequences", **ACM SIGMOD** Conference, pp. 193-204, Snowbird, Utah, June 22-27, 2014, acceptance rate 107/421, **25.4%, 査読有**.

○チュートリアル講演

t1. Y. Sakurai, Y. Matsubara, C. Faloutsos: `` Smart Analytics for Big Time-series Data", ACM SIGKDD International Conference on Knowledge Discovery and Data Mining **(KDD)** Halifax, Nova Scotia, Canada, August 13-17, 2017, **査読有（チュートリアル講演）**.

t2. Y. Sakurai, Y. Matsubara, C. Faloutsos: ``Mining Big Time-series Data on the Web", International World Wide Web Conference **(WWW)** Montreal, Canada, April 11-15, 2016, **査読有（チュートリアル講演）**.

t3. Y. Sakurai, Y. Matsubara, C. Faloutsos: ``Mining and Forecasting of Big Time-series data", **ACM SIGMOD** Conference, Melbourne, AU, May 31-June 4, 2015, **査読有，（チュートリアル講演）**.

○招待講演

i1. 松原靖子, “時系列ビッグデータの特徴自動抽出とリアルタイム将来予測”, ステアラボ人工知能セミナー, 千葉工業大学, 4 月 21 日, 2017 **(招待講演)**.
i2. 松原靖子, “時系列ビッグデータのリアルタイム将来予測と知的社会サービスへの展開”,電子情報通信学会総合大会, スマート無線(SR)研究会,スマート無線における機械学習応用, 3 月 22 日, 2017 **(招待講演)**.
i3. 松原靖子, “時系列ビッグデータのための非線形解析と将来予測”, 電子情報通信学会 コミュニケーションクオリティ研究専門委員会, 大阪大学, 2017 年 1 月 21 日 **(招待講演)**.
i4. 松原靖子, “時系列ビッグデータのためのリアルタイム将来予測”, 電子情報通信学会 インターネットアーキテクチャ研究専門委員会, 広島市立大学, 2016 年 12 月 15 日**(招待講演)**.
i5. Y. Matsubara, 日本学術振興会 第 15 回日米先端科学**(JAFoS)**シンポジウム, (数学・応用数学・情報科学分野, 参加研究者),米国, 2016 年 12 月 2 日-4 日.
i6. Y. Matsubara, “Ecosystem on the Web: Non-Linear Dynamical Systems for Co-evolving Online Activities”, Korea-Japan Database Workshop 2015 (KJDB2015),**(招待講演)**.
i7. Y. Matsubara, “FUNNEL: Automatic Mining of Spatially Coevolving Epidemics”, IMAID2015, Sapporo, Japan, Ocboter 29, 2015, **(招待講演)**.
i8. 松原靖子, “時系列ビッグデータのための特徴自動抽出と将来予測”, 電子情報通信学会 知的環境とセンサネットワーク研究会, 東京, 2015 年 10 月 28 日, **(招待講演)**.
i9. 松原靖子, “時系列ビッグデータのための非線形解析とその応用”, 第 1 回 IPSJ-ONE (DBS 研究会推薦), 情報処理学会全国大会, 京都大学, 2015 年 3 月 17 日, **(招待講演)**.
i10. 松原靖子, “大規模時系列データのための特徴自動抽出と将来予測”, 知的環境とセンサネットワーク研究会, 電気情報通信学会総合大会, 立命館大学, 2015 年 3 月 12 日, **(招待講演)**.
i11. Y. Matsubara, “FUNNEL: Automatic Mining of Spatially Coevolving Epidemics”, Public Health Dynamics Seminar, University of Pittsburgh, USA, September 9, 2014 **(招待講演)**.

〔図書〕(計 0 件)

〔産業財産権〕

○出願状況 (計 1 件)

名称： 予測装置、パラメータ集合生産方法及びプログラム
発明者：松原靖子, 櫻井保志
権利者：国立大学法人熊本大学
種類：特許
番号：特願 2016-138075 (登録番号)
出願年月日：2016 年 7 月 12 日
国内外の別： 国内

○取得状況 (計 0 件)

〔その他〕

○受賞

a1. 2016 年度 情報処理学会 山下記念研究賞 (2017 年 3 月 16 日)
a2. 第 8 回データ工学と情報マネジメントに関するフォーラム (DEIM2016) 最優秀論文賞 (2016 年 6 月 18 日)
a3. 第 8 回データ工学と情報マネジメントに関するフォーラム (DEIM2016) 優秀論文賞 (2016 年 6 月 18 日)
a4. 日本データベース学会上林奨励賞(2016 年 3 月 1 日)
a5. 第 8 回 Web とデータベースに関するフォーラム(WebDB Forum 2015) 最優秀論文賞(2015 年 11 月 24 日)
a6. 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) 最優秀論文賞(2014 年 3 月)

○受賞

ソフトウェアの公開など
http://www.cs.kumamoto-u.ac.jp/~yasuko/

６．研究組織
(1)研究代表者
松原靖子 (MATSUBARA Yasuko)
熊本大学・先端科学研究部・助教
研究者番号：00721739

(2)研究協力者
櫻井保志 (SAKURAI Yasushi)
熊本大学・先端科学研究部・教授
研究者番号：30466411

Christos Faloutsos
Carnegie Mellon University
Dept. of Computer Science
Professor