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　元 年　 ５ 月 ２３ 日現在

機関番号：１４４０１
研究種目：基盤研究(A)（一般）
研究期間：2014～2018
課題番号：２６２４００１３

研究課題名（和文）モバイルユーザが生成する「人」センサデータの共有基盤システムの構築

研究課題名（英文）Development of Platform Systems for Sharing Human-Sensor Data Generated by Mobile Users

研究代表者
西尾　章治郎（Nishio, Shojiro）

大阪大学・その他部局等・総長

研究者番号：５０１３５５３９

交付決定額（研究期間全体）：（直接経費）　　32,100,000円

研究成果の概要（和文）：本研究課題では、スマートフォンなどモバイル端末の利用履歴やセンサデータおよびTwitterを始めとするマイクロブログなど「人」が生成する情報を有効活用して、新たなサービスを創出するための、データ収集・解析・管理・共有技術の研究開発を行った。具体的には、(1) センサデータの収集・解析・管理のための基盤技術、(2)マイクロブログデータの収集・解析・管理のための基盤技術、(3)ユーザの行動履歴データの収集・解析・管理のための基盤技術について、有効な技術群を開発した。

研究成果の学術的意義や社会的意義
本研究の成果は、データベースや自然言語処理などの分野で世界的に最も権威のある国際論文誌や、データベースや分散処理分野で最難関の国際会議に複数採択されるなど、国際的に高く評価されており、学術的に非常に大きな業績を達成した。特に、新しいデータ解析技術や、人が生成した短文に対する複合語の重要度を示す新しい指標を提案できたことなどは、学術的な貢献が大きいと考える。また、構築した技術群やデータおよび解析結果の共有プラットフォームは、今後、益々増加する「人」が生成するビッグデータを解析・管理・共有するための重要な基盤になるものと期待され、社会的な意義が大きいと考える。

研究成果の概要（英文）：In this research project, we aimed to make a full use of big data generated by "human" including sensor data collected by mobile devices such as smartphones and microblog data such as Twitter posts, and developed a variety of techniques for collecting, analyzing, managing, and sharing such data to create new services. Specifically, we developed (1) fundamental techniques for collecting, analyzing, and managing sensor data, (2) fundamental techniques for collecting, analyzing, and managing microblog data, and (3) fundamental techniques for collecting, analyzing, and managing historical data on users' activities.

研究分野：データ工学

キーワード：ビッグデータ分析・活用　参加型センシング　ソーシャルメディア　マイクロブログ　センサデータ収集・解析

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９、ＣＫ－１９（共通）

１．研究開始当初の背景

近年、スマートフォンなどモバイル端末の高機能化が進み、爆発的に普及している。このような高機能なモバイル端末は、GPS や温度、加速度、方位などの各種センサを搭載しており、大規模数のユーザが所持していることから、従来のセンサネットワークでは実現不可能であった超高密度・超広域なセンシングを可能とする新たなアプローチとして期待されている。このようなアイデアは「参加型センシング」や「都市センシング」と呼ばれ、ビッグデータの中心的な役割を担う重要な要素として、近年、様々な研究開発が進められている。一方、スマートフォンでは様々なアプリケーションやサービスが利用可能であり、単純な電話としてではなく、ユーザの生活に密着した高機能な PC としての役割を果たしている。そのため、旅行やショッピングなどユーザの実社会での活動や、アプリケーション利用履歴などサイバー社会での活動の両方を把握するためのデバイスとしても期待されている。さらに、Twitter などのソーシャルメディアを通じて、ユーザ自身が情報を発信するためにも用いられる。このような活動記録およびユーザが発信した情報を収集し、ユーザの行動をモデル化することで、状況に応じた情報推薦や、行動予測、ナビゲーションなど様々な応用が期待されている。

従来の研究や商用システムのほぼ全てにおいて、単独のアプリケーションとしてデータの収集・解析・管理を行っており、複数のアプリケーション間でのデータや解析結果の共有は想定していない。しかし、これらを共有可能にすることにより、データ量・種類の増加によって解析精度が向上するとともに、既に保存されている結果の再利用による解析時間の短縮および消費電力の低下などが期待できる。

２．研究目的

本研究課題では、世界で初めての試みとして、「人」（モバイル端末）が生成する多岐に亘るデータを収集・解析・管理するための統合的な基盤技術の確立を目指す。特に、スマートフォンなどモバイル端末の利用履歴やセンサデータおよび Twitter を始めとするマイクロブログなど「人」が生成する情報を有効活用して、新たなサービスを創出するための、データ収集・解析・管理・共有技術の研究開発を行う。スマートフォンの爆発的な普及により、モバイル端末を人に密着した一種のセンサと見なすことがトレンドとなっており、既にいくつかのサービスが開発されている。しかし、これらのほとんどは単独のアプリケーションとして、データの収集・解析・管理を行っており、複数のアプリケーション間でのデータや解析結果の共有は想定していない。本研究では、世界で初めての試みとして、一般のユーザが生成する多岐に亘るデータを収集・解析・管理するための統合的な基盤技術を確立し、その効果的な共有を目指す。

３．研究の方法

本研究課題では、上記の研究目的を達成するために、以下のような方法で研究開発を進めた。

(1) **センサデータの収集・解析のための基盤技術の確立**

「人」や「モノ」が生成した多種多様なデータに対して、データマイニングや学習技術などを用いて特徴抽出を行い、その結果を以降の検索の効率化のためにインデックスなどのデータ構造としてデータベースに蓄積する機構を実現する。さらに、そのデータ構造上での検索機構を実現する。

(2) **マイクロブログ・ソーシャルメディアデータの収集・解析**

Web などから収集したマイクロブログ・ソーシャルメディアデータに対して、データの内容解析・特徴抽出などを行う機構を実現する。例えば、 Twitter データを解析し、ツイートされた場所やイベントの特定や、話題になっている事象や その時間変化などの検出を試みる。

(3) **スマートフォンユーザの行動履歴データの収集・解析**

研究代表者らがこれまでに開発した育成ゲームを拡張し、効率的にユーザの状況および行動に関するログを収集する機構を構築する。また、収集したデータを解析し、ユーザのおかれている状況と、情報アクセスおよびアプリケーション利用の関係を明らかにする。

(4) **「人」が生成するデータを「社会センサ」として収集・解析・共有する基盤の確立**

本研究の 2 年目からは、多岐に亘るデータを効率的に収集・解析・共有するために、新たな共有基盤システムを構築する。

４．研究成果

上記の研究を推進した結果、下記のような成果を達成した。

(1) **センサデータの収集・解析のための基盤技術の確立**

IoT 技術により大量のデータが獲得できるため、データ集合のサイズは基本的に大きく、1,000,000 以上のサイズとなることもしばしばである。そのため、大量のデータの中からアプリケーションが求めるデータや知識を効率的に獲得する技術は、データ解析に不可欠である。本項目では、主に「1 次元センシングデータ」、「多次元データ」、および「トランザクション（集合）データ」を対象とし、それぞれのデータ集合の中から、有用なデータ・知識（例えば、重要なデータが集中している場所、多くのデータに勝っているデータ、および多くのデータに現れているパターンなど）を効率的（高速）に検索・獲得する技術（アルゴリズム）を開発した。

高速アルゴリズムを設計するためには、対象となるデータおよび解析ツール（オペレータ）の特徴を考慮する必要がある。そのため、本研究では特定のデータモデルおよびオペレータに特化したデータ構造とそれに基づくアルゴリズムを設計した。設計したアルゴリズムは実データを用いた実験により評価し、既存技術よりも大幅に性能が良いことを確認した。以下では、取り組んだ問題の中で成果が顕著であった 2 つのものについて具体的に紹介する。

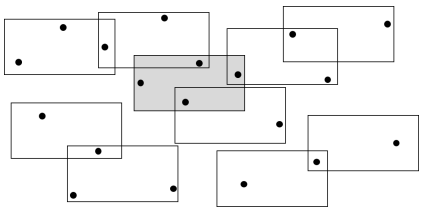

図 1：MaxRS monitoring 問題

MaxRS and MaxCRS monitoring：本問題では、各データはある重み（値）を持った 2 次元点で表されることを想定したとき、アプリケーションが確保したい長方形領域のサイズを指定した際、その長方形に含まれる点の重みの和が最大となる長方形の場所（図 1 の灰色の位置）を計算する。これは周辺のデータに対する影響力を最大化する重要な場所を計算する問題である。また、センシング環境では、古いデータは不要（削除）となり、新しいデータを追加し、常に結果をモニタリングすることが要件として挙げられる。つまり、解が時々刻々と変わる中で、それを効率的にトラッキングする技術が求められる。本問題では、点が次々に追加され、重要な位置はそれに応じて変化していく。これを高速に更新するアルゴリズムを開発した。

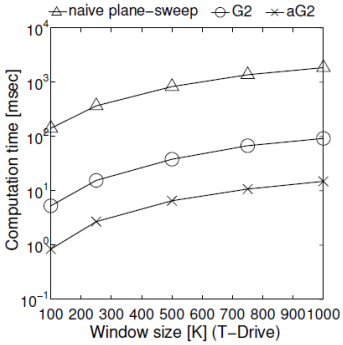

図 2：評価結果 1

本問題で開発したアルゴリズムは、新たな点が追加されたり古い点が削除されたりする場合においても解を一から再計算することなく更新できる。点が存在する空間をグリッドで分割し、点を中心とした長方形の交差関係を示したグラフ構造を統合した新たなデータ構造（G2 および aG2）を設計したことにより、新たな点が追加された場合に解になり得る空間を大幅に限定することに成功した。つまり、一から再計算するよりも計算コストが大幅に削減される。さらに長方形だけでなく、円も矩形として指定できるようにアルゴリズムを拡張した。実世界データを用いて評価実験を行った結果、既存技術を使ったアルゴリズムに対して、約 100 倍の高速化を達成したことを確認した（図 2）。本研究の成果はデータベース分野の最難関の国際会議、および地理情報システムに関する権威のある論文誌に採択されている。

Top-k dominating query processing：1 つのデータが複数の d 個の属性を持つとき、そのデータは d 次元のベクトルで表すことができる。2 つのデータが与えられたとき、d 個の各次元の値をそれぞれ比較した際、一方のデータ o がもう一方のデータ o’ に対して全ての次元で値が勝っている（または対等）であるとき、o は o’ を支配していると言う。本問題は、データ集合から最も多くのデータを支配している k 個のデータを検索する。これまで考えられてきたスカイライン検索や Top-k 検索に代表されるユーザ嗜好型検索は、検索結果の数をコントロールできないことやパラメータの指定が困難といった欠点があった。Top-k dominating 検索は検索結果の数（k）だけ指定すれば良いものであり、既存の検索方法の欠点を克服している。しかし、あるデータが支配している全てのデータを計算するコストは大きく、全てのデータに対してこの処理を行うことは望ましくない。

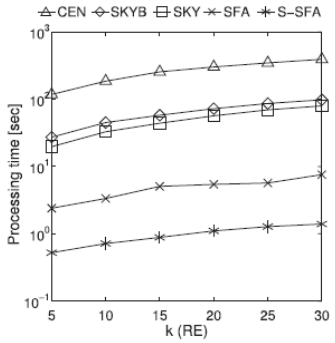

図 3：評価結果 2

そこで、この処理をできるだけ実行しない（検索結果に入り得るデータのみ実行する）分散処理アルゴリズム SFA を開発した。このアルゴリズムでは、各データの支配しているデータ数を推定し、解になりえないデータをフィルタリングしていく。また、計算するマシンを増やすほど実行時間は短縮される。さらに、少し解の精度を犠牲にするだけで数倍以上の実行時間短縮を達成する近似アルゴリズムを開発した。本研究の成果は、データベース分野で最高権威の論文誌に採択されている。

(2) マイクロブログ・ソーシャルメディアデータの収集・解析

Twitter などの SNS データから実社会のトレンドや、ユーザの興味やライフイベント等を抽出するマイニング・機械学習技術の開発を行った。さらに、これらのタスクに利用可能な汎用技術として、ツィートなどの投稿文に位置情報を付与する技術や、ツィートなどの短文に出現する複合語の重要性を見積もる指標を開発した。

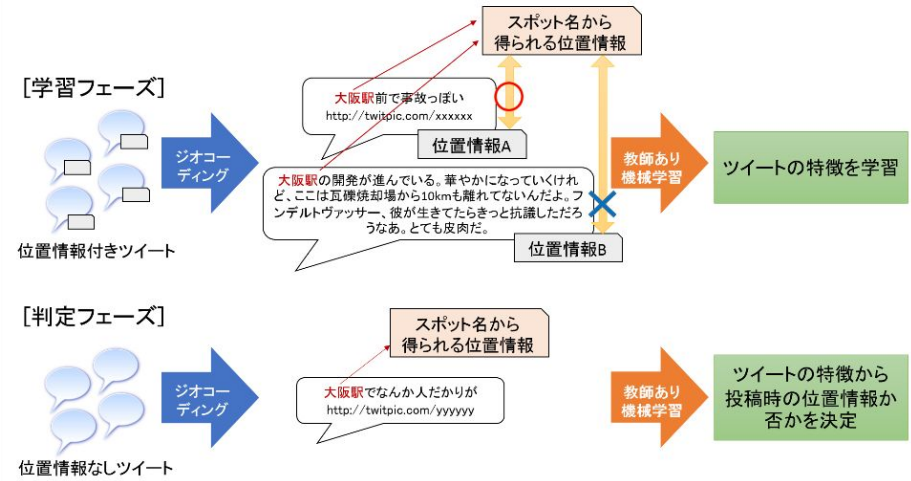

図 4：ツィートへの位置情報の付与手法

トレンド抽出の技術としては、Twitter 全体において爆発的に使用頻度が増加した用語をトレンドとしてとらえる手法を考案した。この際、ツイートがノイズの多い短文であることを考慮して、用語（特に複合語）の曖昧性を排除しつつ、重要性を正しく見積もるために、後述する指標を利用した。ユーザの興味を抽出する技術としては、各ユーザのツィートのタイムラインを解析し、ユーザの長期の興味と短期の興味の両方を考慮して、興味の対象とその度合いを適切に抽出する手法を考案した。この手法はツィートの情報量が少ないことを考慮して、Wikipedia を用いて語彙拡張を行っている。ユーザのライフイベントを抽出する技術としては、対象とするイベントを経験したユーザの過去のツィートのタイムラインを解析し、ツィート中に高い頻度で抽出する語や、その変化、文体の特徴等のパターンを学習する手法を考案した。この手法を用いて、他のユーザ（テストユーザ）が将来、そのイベントを経験するかどうかを予測することが可能となる。

さらに、汎用技術として、ツィートに位置情報を付与する手法を考案した。この手法では、まず位置情報付きツィートを収集し、その中で位置（大阪など）に関する言及があるものを選択し、その位置が付与されている位置情報と一致する場合と、そうでない場合の特徴を学習する。その後、一般の位置情報が付与されていないツィート（テストデータ）に対して、その特徴を考慮して、ツィート内に出現する位置と、投稿位置が一致するかどうかを判定し、一致する場合はその位置をツィートの投稿位置として付与する（図 4）。次に、複合語の重要性を見積もる指標としては、従来の IDF 指標を複合語に適するように拡張するために、情報距離等の数学理論に基づいた枠組みを考案し、それを実現するためのアルゴリズムを開発した。

上記の一連の技術は、実データを用いた詳細な評価実験により、その有効性を検証した。これらの成果は国内外で高く評価され、Web や自然言語処理の分野で最難関の国際会議や、情報システム分野で最高権威の国際会議に論文が採択されている。

(3) スマートフォンユーザの行動履歴データの収集・解析

研究代表者らがこれまでに開発したモンスター育成ゲーム「こんてきすとモンスター」を拡張し、ユーザが育成ゲームを楽しみながら、体調・気分・忙しさ・同伴者の人数などのセンサでは検出できないハイレベルなコンテキストを入力可能なシステムを開発した。このシステムでは、従来のゲームの機能に加えて、ユーザの状況に応じて適切なアプリケーション（アプリ）を推薦する機能や、入力したコンテキスト情報の統計等をライフログとして提供する機能を実装した（図 5）。このシステムを実運用した結果、約 400 名のユーザから約 70 万件のアプリ利用ログを収集した。収集したログを解析した結果、ユーザのハイレベルなコンテキストとアプリ利用の特性には強い関係があることを確認した。

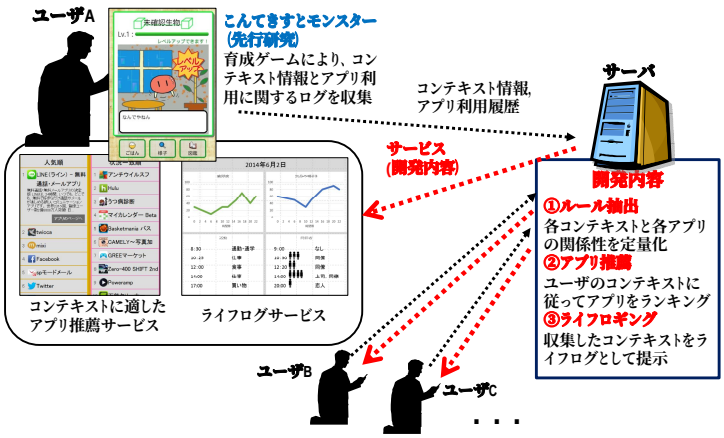

図 5：開発システムの概要

この知見に基づいて、ユーザのコンテキスト情報と過去のアプリ利用履歴に基づいて、ユー

ザが次に利用するアプリを予測する問題に取り組み、コンテキストを利用しない場合と比較して予測精度が大幅に向上することを確認した。さらに、ユーザのアプリ利用履歴から、逆に、ユーザが自宅にいるかどうかなど、ユーザのコンテキストを予測する問題にも取り組んだ。

(4) 「人」が生成するデータを「社会センサ」として収集・解析・共有する基盤の確立

本研究の 2 年目から、多岐に亘るデータを効率的に収集・解析・共有するために新たな共有基盤システム S³ (Social Sensor Sharing) システムを構築した。

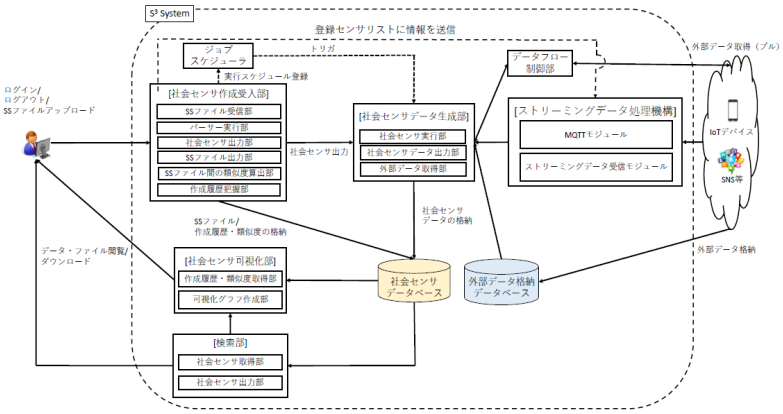

図 6：S³ システムのシステム構成

S³ システムの全体像と狙い：近年の携帯型端末の普及により、SNS (Social Networking Service) が広く利用されている。SNS に投稿された短文や写真から、単語の出現頻度や写真撮影地点を解析することで、実社会の情報を取得できる。本研究では、このような実社会の情報など、SNS への投稿等を解析して得られるデータを**社会センサデータ**と呼ぶ。社会センサデータや解析プログラムといった社会センサデータの生成に関する記述を共有することで、新たな社会センサデータを生成する際に参考にするなど、再利用することが可能となる。そこで本研究では、社会センサデータを様々な応用に再利用しやすくするため、社会センサデータを生成・共有するための S³ システムを構築した。

S³ システムの機能：社会センサデータを生成するユーザは、生成のための関数群、社会センサデータの型を定義するファイル、社会センサデータの出力方法に関する設定ファイルの三つのファイルを作成する。本研究では、これらのファイルをそれぞれ SSFD (Social Sensor Function Description)、SSTD (Social Sensor Type Definition)、SSOC (Social Sensor Output Configuration) と呼ぶ。社会センサデータの生成に関する記述をこれらのファイルに分けることで、新たな社会センサデータの生成に有用な部分のみ再利用できる。S³ システムのシステム構成を図 6 に示す。社会センサ作成受入部にこれらのファイルがアップロードされると、パーサ実行部で構文解析し、社会センサデータの生成プログラム SSCP (Social Sensor Data Creation Program) を作成して社会センサデータ生成部に送信する。社会センサデータ生成部は社会センサの実行状態を管理しており、ジョブスケジューラが社会センサを実行するための信号（トリガ）を社会センサデータ生成部の社会センサ実行部に送信すると、社会センサ実行部で社会センサが実行される。社会センサは、外部データ格納データベースもしくはデータフロー制御部を介して、SNS 等から投稿等を取得して解析し、結果を社会センサデータ出力部へ送り、社会センサデータベースに保存する。アップロードされた三つのファイルや生成された社会センサデータは、S³ システムの Web インタフェースを介して検索、ダウンロードできる。

S³ システムの実装：CentOS6.7 搭載のワークステーションを用いて S³ システムを構築した。SSFD の記述にはプログラム言語として広く用いられている Java を採用し、SSTD、SSOCS の記述には構造型文書の記述に適した XML を採用した。Web サーバソフトウェアには Apache/2.2.15、Java プログラムを動作させるサーブレットには Apache Tomcat/8.0.30、データベースソフトウェアには mysql5.7.10 を用いた。社会センサデータの更新のためのジョブスケジューラには、UNIX 系 OS に標準で搭載されている crontab を用いた。SSCP の作成、crontab ファイルの編集、データベースへの問合せなど、サーバ側の処理は Java1.8.0 66 で実装した。SSCP の作成および crontab ファイルの編集時の、ユーザが入力する XML ファイルの構文解析には、Java で XML を扱うための API である JAXP を用いた。

S³ システムの評価：S³ システムと、社会センサデータの共有に関連する機能を有する mTrend、Github、Climbi、SeRAVi、X-Sensor と機能比較した結果を表 1 に示す。S³ システムでは、SSFD に解析処理を記述することで、SNS のデータ等を解析できる。また、S³ システムは、SSTD、SSFD、SSOC の共有機能や Web インタフェースを介して社会センサデータを共有する機能を提供している。mTrend では、ツイートを解析して時空間的動向を可視化する機能を提供しており、SeRAVi や X-Sensor では、モバイルセンサデータの分布を可視化する機能は提供している。しかし、S³ システムでは、社会センサデータを表形式で表示するのみで可視化できないため、今後拡張を予定している。

表 1：機能比較

	データの解析	社会センサデータ生成に関する記述の共有	社会センサデータの共有	社会センサデータの可視化
提案プラットフォーム	○	○	○	×
mTrend	△	×	×	△
Github	×	△	×	×
Climbi	×	△	×	×
SeRaVi	×	×	△	△
X-Sensor	×	×	△	△

５．主な発表論文等

〔雑誌論文〕（計 27 件）

① Tatsuya Nakamura, Masumi Shirakawa, Takahiro Hara, Shojiro Nishio, Wikipedia-Based Relatedness Measurements for Multilingual Short Text Clustering, ACM Transactions on Asian and Low-Resource Language Information Processing, volume 18, number 2, pages 16:1-16:25, 2018 (DOI: 10.1145/3276473)（査読あり）.

② Daichi Amagata, Takahiro Hara, Makoto Onizuka, Space Filling Approach for Distributed Processing of Top-k Dominating Queries, IEEE Transactions on Knowledge and Data Engineering, volume 30, number 6, pages 1150-1163, 2018 (DOI: 10.1109/TKDE.2018.2790387)（査読あり）.

③ Daichi Amagata, Takahiro Hara, Mining Top-k Co-Occurrence Patterns across Multiple Streams, IEEE Transaction on Knowledge and Data Engineering, volume 29, number 10, pages 2249-2262, 2017（査読あり）.

④ Daichi Amagata, Takahiro Hara, A General Framework for MaxRS and MaxCRS Monitoring in Spatial Data Streams, ACM Transactions on Spatial Algorithms and Systems, volume 3, number 1, pages 1-34, 2017 (DOI: 10.1145/3080554)（査読あり）.

⑤ Masumi Shirakawa, Takahiro Hara, Shojiro Nishio, IDF for Word N-Grams, ACM Transactions on Information Systems, volume 36, number 1, Article 5, 2017 (DOI: 10.1145/3052775)（査読あり）.

⑥ Daichi Amagata, Yuya Sasaki, Takahiro Hara, Shojiro Nishio, Efficient Processing of Top-k Dominating Queries in Distributed Environments, World Wide Web, volume 19, number 4, pages 5454-577, 2016 (DOI: 10.1007/s11280-015-0640-6)（査読あり）.

〔学会発表〕（計 83 件）

① Daichi Amagata, Takahiro Hara, Chuan Xiao, Dynamic Set kNN Self-Join, IEEE International Conference on Data Engineering (ICDE), 2019.

② Zennosuke Aiko, Keisuke Nakashima, Tomoki Yoshihisa, Takahiro Hara, A Social Sensor Visualization System for a Platform to Generate and Share Social Sensor Data, IEEE 42nd Annual Computer Software and Applications Conference (COMPSAC), 2018.

③ Masumi Shirakawa, Takahiro Hara, Takuya Maekawa, Never Abandon Minorities: Exhaustive Extraction of Bursty Phrases on Microblogs Using Set Cover Problem, Conference on Empirical Methods in Natural Language Processing (EMNLP), 2017.

④ Masahiro Yokoyama, Takahiro Hara, Efficient Top-k Result Diversification for Mobile Sensor Data, International Conference on Distributed Computing Systems (ICDCS), 2016.

⑤ Daichi Amagata, Takahiro Hara, Monitoring MaxRS in Spatial Data Streams, Int. Conf. on Extending Database Technology (EDBT), 2016.

⑥ Masumi Shirakawa, Takahiro Hara, Shojiro Nishio, N-Gram IDF: a Global Term Weighting Scheme Based on Information Distance, International World Wide Web Conference (WWW), 2015.

⑦ Daichi Amagata, Takahiro Hara, Shojiro. Nishio, Distributed Top-k Query Processing on Multi-Dimensional Data with Keywords, Int. Conf. on Scientific and Statistical Database Management (SSDBM), 2015.

〔図書〕（計 2 件）

① Takahiro Hara, Elsevier, Adaptive Mobile Computing: Advances in Processing Mobile Data Sets, 全 262 ページ (第 3 章: Fusion of Heterogeneous Mobile Data, Challenges and Solutions (pages 47-63)を担当), 2017.

② Takahiro Hara, Jun Osawa, Institution of Engineering and Technology (IET), Big Data Recommender Systems: Recent Trends and Advances, 全 600 ページ (第 25 章: Investigation of Relationships between High-level User Contexts and Mobile Application Usage を担当), 発行予定, 2019.

６．研究組織

(1)研究分担者

研究分担者氏名：原　隆浩

ローマ字氏名：Takahiro Hara

所属研究機関名：大阪大学

部局名：大学院情報科学研究科

職名：教授

研究者番号（8 桁）：20294043

(2)研究協力者

研究協力者氏名：Sanjay Madria、Peter Scheuermann、Wang-Chien Lee、Xing Xie

ローマ字氏名：Sanjay Madria、Peter Scheuermann、Wang-Chien Lee、Xing Xie