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 ２７ 年　 ６ 月 ２９ 日現在

機関番号：３５４０３
研究種目：基盤研究(C)（一般）
研究期間：2012～2014
課題番号：２４５００３０１

研究課題名（和文）ソーシャルメディアを用いた地域情報の分析および位置情報システムの構築

研究課題名（英文）Local Area Information Analysis on Social Media and Location Based System

研究代表者
石田　和成（Kazunari, Ishida）

広島工業大学・情報学部・准教授

研究者番号：２０３０３０２６

交付決定額（研究期間全体）：（直接経費）　　2,400,000円

研究成果の概要（和文）：位置情報システムの基礎的技術を開発するために，ソーシャルメディアを用いた地域情報の収集システムおよび分析手法を開発した．地域情報収集についてマイクロブログ(Twitter)の情報を収集，蓄積するシステムの開発を行うとともに，地域情報分析についてジオタグ付きデータにもとづくユーザ位置推定，各地域の話題抽出を行った．また，地域情報システムにおけるオープンデータの活用について検討した．さらに，無線モジュールを用いた地域情報流通システム，および，定点センサ，ウェアラブルセンサを用いた位置情報や行動データの収集，分析について検討を行った．

研究成果の概要（英文）：We developed analysis methods to organize local information based on social media, in order to propose fundamental technologies for developing location information systems. In order to collect local information, we developed data collection systems concerning social media and public open data. We also developed an estimation method of users' locations so that we could extract local topics on administrative divisions. In addition, we discussed an application of open data on a local information system. Moreover, we also discussed local information distribution systems, in order to collect and analyze location information and behavioral data with radio frequency modules and sensor devices.

研究分野：情報システム

キーワード：ソーシャルメディア　マイクロブログ　位置情報　位置推定　地域固有表現　オープンデータ　センサ　GPS

１．研究開始当初の背景

(1)位置情報の収集分析のためにソーシャルメディアの利用した研究として，Sankaranarayanan(2009)らは，Twitter における話題を抽出する TwitterStand を開発した．このシステムでは単語にもとづき Tweets をトピックごとにクラスタリングし，クラスタ毎に地名に関する単語を抽出し，地名から位置を特定する公開サービス GeoNames を用いて，トピックの位置特定を行う．このシステムによりソーシャルメディア上の話題と位置との関係を大まかに把握できる．しかし，同じ地名で地理的にまったく異なる場所も多数存在するため，地名に関する単語抽出，地名から位置の特定という 2 つの段階で，位置特定の精度が低下する可能性がある．

(2)Kinsella(2011)らは，Geo-tag 付き Tweets を用いて，各局所的位置において特有の単語を調べ，単語と位置のモデルを構築することにより，Tweets の位置を特定する方法を開発した．そして，この方法と Yahoo! Placemaker について位置特定精度の比較を行い，提案手法の精度が高いことを示した．この言語モデルにもとづくこの手法は，各地域において変化の少ない継続的な話題については有効であると考えられが，モデル作成後に生じた新規の話題についての位置特定は精度が低下すると考えられる．

(3)オープンデータにもとづく位置情報システムとして，松村ら(2011)は Linked Open Data (LOD)による博物館情報および地域情報の連携活用システムを開発した．このシステムは補完的な複数データの連携により，利用者に対して幅広い情報提供を行うものである．オープンデータの有効活用のためには，このようなデータ連携事例の充実が必要である．

(4)位置情報の収集，分析のために，ＧＰＳ，カメラ付き携帯端末を利用した研究として，佐藤ら(2009)は，ＧＰＳ付き携帯電話を用いた写真投稿共有サイトを開発し，実証実験を行うことにより，地域やコミュニティに関する関心の変化について分析した．また，Papliatseyeu(2008)らは，GPS に加え，Wi-Fi，3G ネットワークにより位置を特定することにより，屋内外で利用できる，位置情報記録システムを開発した．Xie (2009)は，GPS を用いた移動経路の軌跡を用いて，徒歩，自転車，バスなどの移動手段を検出するとともに，ウェブ検索のログを用いて，検索者の位置にもとづく共起検索パターンの分類を行い，地域毎の頻出パターンが観察されることを示した．これらの研究では GPS が利用可能なインターネット携帯端末を持つ多数の利用者の存在が前提条件となっている．しかし，人口密度の低い地域では，この条件充足が困難な場合がある．

これらの国内，国外の研究動向を踏まえ，本研究では，ソーシャルメディアやオープンデータを用いた地域情報の分析および位置情報システムを構築する．このシステムの構成要素として，マイクロブログ(Twitter)の情報を収集，蓄積するシステムの開発を行うとともに，ジオタグ付きデータにもとづくユーザ位置推定，各地域の話題抽出を行う．また，地域情報システムにおけるオープンデータの活用について検討する．さらに，無線モジュールを用いた地域情報流通システム，および，定点センサ，ウェアラブルセンサを用いた位置情報や行動データの収集，分析について検討を行う．

２．研究の目的

本研究の目的は，ソーシャルメディアおよびオープンデータを用いた地域情報の分析のための位置情報システムの構築である．携帯端末の高機能化と普及により，ソーシャルメディアにおいて位置，時間の特定できる情報が増大している．この情報にもとづき，地域，時間毎のデータの量的，質的分析手法を構築し，地域情報分析システムを開発する．また，無線通信機器を用いた簡易型のデータ通信システムを開発し，既存の情報通信網の利用が困難な環境における情報共有基盤の試作を行う．加えて，定点センサを用いた地域における行動マイニングシステムの開発および実験を行う．さらに，ウェアラブルセンサを用いた行動マイニングシステムの開発および実験を行う．

３．研究の方法

(1)位置情報データの収集とデータベース構築を行う．ソーシャルメディアにおける情報収集蓄積システムを開発し，時間，位置情報にもとづきデータ管理する．

(2)複数オープンデータを用いた異種データの位置的，時間的な相関にもとづく地域情報の組織化手法を開発する．

(3)無線モジュールを用いた地域情報流通システムの開発を行う．安価，低電力な無線モジュールの活用により，既存の通信網の有無にかかわらず，データの収集，配信を行うための基盤を構築する．

(4)行動マイニングのための定点センサを用いた分析手法の開発を行う．定点センサを用いることにより，通信環境，情報端末の普及率にかかわらず，行動マイニングのためのデータ収集を行うことができる．

(5)行動マイニングのためのウェアラブルセンサを用いた分析手法の開発を行う．GPS で同期された複数センサを用いることにより，詳細な行動データ取得を行うことができる．

(6)地域特有の単語共起にもとづき，マイクロブログ利用者の位置推定と地域トピックの抽出を行う．単語共起により単語意味の曖昧性を削減することにより，位置推定精度を改善する．また，利用者位置推定にもとづき地域毎の話題抽出を行う．

## ４．研究成果

(1)オープンなソーシャルメディアである Twitter の情報を蓄積し，キーワード，位置情報を抽出，蓄積するシステムの開発を行った．大規模データを扱うため，Hadoop および Hive を用いてシステムの実装を行った．このシステムを用いて，全国的，および，各地域の話題抽出を行った．2011 年 8 月から 2012 年 3 月の間に収集したデータにおいて，位置情報付きの Tweets の割合を調査したところ，データ全体 56,376,438 件のうち，134,226 件，およそ 0.24%であった．また，位置情報付き Tweet を投稿するユーザは，全体 4,826,951 件のうち，13,720 件，およそ 0.28%であった．つまり，Twitter のデータにおける位置情報付きのデータは，非常に少ないことが分かった．位置情報付きデータの少ない原因の１つは，正確な場所と時間の情報の公開による，プライバシー問題によるものと思われる．この問題を克服し，地域毎の話題の抽出を行うためには，ユーザの大まかな位置の特定を行う必要がある．大まかな位置の特定に，総務省統計局の定める地域メッシュコードを用い，位置情報付きの Tweet にもとづき，地域特有のキーワード集合を特定した．この集合にもとづき，Twitter ユーザの大まかな位置の推定を行い，各地域のユーザが用いたキーワードにもとづき地域毎の話題の抽出を行った．特定の地域として広島西部を用い，地域のスポーツ，観光地の話題が抽出できることを確認した．また，特定のテーマとして原発を用い，Twitter ユーザが近隣原発の報道に関心を示すことを確認した．

(2) オープンデータにもとづく地域情報システムの具体例として，放射性物質の拡散をテーマとし，拡散推定手法を開発した．これは，Twitter データにもとづくデータ分析により，東日本大震災により生じた福島第一原発事故以降，近隣原発の報道に Twitter ユーザが高い関心を示しており，今後必要性の高まる地域情報となると考えたためである．そのため，オープンデータとして，原子力規制委員会が提供する全国のモニタリングポストの空間線量，気象庁が公開する降雨量，風速風向のデータを用いて，放射性物質の拡散源を特定する手法を提案する．

大気中の放射性微粒子は，降雨時に落下しモニタリングポストの値を増大させる．そのため，本手法では，降雨量のデータから，連続した降雨期間を特定し，それぞれの期間における有効降雨期間を求める．降雨初期には，大気中の放射性微粒子が落下し，線量増大が観察されるが，降雨が続くと，降雨に大気が清浄化され，新たな微粒子の落下が無くなるため，線量は通常のレベルまで減少する．ここで，降雨開始後から継続的降雨により線量が通常レベルに戻るまでの期間を，有効降雨期間とする．また，風向風速係数 w は，風速を s, モニタリングポストでの風向きを θ (北風，東風，南風，西風は，それぞれ，0，90，180，270 度)，モニタリングポイントから拡散源への方位角を α とするとき，以下の式(1)で定める．

$$w = \frac{1}{2}\left(1 - \frac{1}{1+s}\right)\left(1 + \cos\left(\frac{\pi(\theta - \alpha)}{180}\right)\right) \quad \ldots(1)$$

具体的事例として，北九州市八幡西区の八幡総合庁舎のモニタリングポストと，最寄りの気象庁の八幡観測所について，2012 年 11 月初旬から 12 月中旬の 1 カ月強のデータから，有効降雨期間として 8 ケースが抽出された．全 8 ケースにおいて，モニタリングポイントから拡散源への方位角 α について 0～359 まで 1 度刻みで変化させた場合の風向風速係数 w をそれぞれ求め，単位線量増大量との相関係数を計算したところ，図１の結果が得られた．母相関係数の検定にもとづき，有意水準１%において，母相関係数が 0 ではない棄却域は，F>13.7450 となり，相関係数が 0.8340 以上の範囲となる．方位角 α の限界の角度は 76 度と 113 度である．また，最大の相関係数が得られる，モニタリングポストから拡散源への方位角 α は 92 度である（相関係数は 0.8881）．図２に，拡散源の方位角を 92 度とした場合の，風光風速係数と放射線増大量との散布図を示す．有意水準１%の方位角範囲 76～113 度にはごみ焼却場が存在し，その方位角は，90.0865 と，最大相関係数が得られる方位角 92 度と非常に近く，ごみに含まれる放射性物質の再拡散の可能性が示唆された．

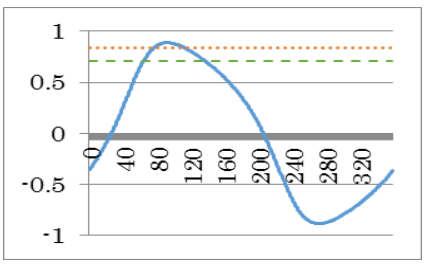


図１：拡散源方位角と相関係数

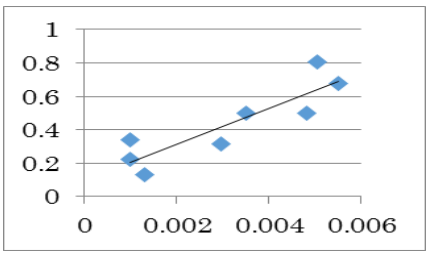


図２：風向風速係数と放射線増大量

(3)地域情報システムで利用するデータの流通基盤として，近距離無線通信を用いた階層的な地域情報化システムの試作を行った．本システムは，各種センサと機器間通信，そして人や乗り物の移動にもとづき，地域情報の収集配信を行う．通信料金の必要な 3G，LTE

などの移動体通信の代わりに，安価な近距離無線通信を用いるため，時間的な遅延が許容される情報について安価な転送方法となる．このシステムは 4 階層で構成される(図 3)．第 1 層はデータ収集を目的とした Sensor Node を配置する．単純なセンサ値の収集は，近距離無線通信機器（ZigBee）およびマイコン（Arduino）を用いる．第 2 層は，第 1 層の複数の Sensor Node から得られたデータを記録する Recording Node を配置する．第 3 層は，複数の第 2 層の Recording Node からデータを収集する Transporting Node を配置する．Transporting Node は，移動する人や乗り物に設置し，データ発信源(Recording Node)周辺で自動的なデータ収集を行い，第 4 層へのデータ転送を行う．第 4 層はインターネットへのゲートウェイを配置し，複数の第 3 層の Transporting Node が収集したデータをインターネットに転送する．インターネットへのゲートウェイは，常時接続のある施設に設置する．このデータ流通基盤を利用し，地域における人々の位置，活動情報など，行動マイニングに必要なデータを収集，配信する．

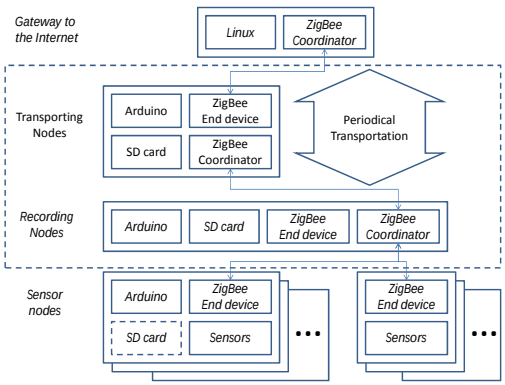


図 3：階層的地域情報化システム

(4)地域における位置情報の収集基盤として，赤外線人感センサを利用したシステムの試作を行った．このセンサを商店街や公園などに設置し，継続的に人の動きを計測することにより，平常時と催し物開催時との人の動きの変化にもとづき，催し物の効果の定量的な比較を行う．事例として島根県江津市商店街で定期的に実施されている「手つなぎ市」(春分の日, 2014 年 3 月 21 日金曜日 9:00～16:00)におけるデータ収集を行った．主会場および，主会場から離れた幹線道路沿いの周辺部における時系列データを図 4，5 に示す．催し物当日(21 日)の天候は曇り時々雪，翌日(22 日)は晴れであったが，主会場において，当日は翌日と比べピーク時は 5 倍程度の人の動きがあり，催し物による集客効果があったことが分かる(図 4)．幹線道路沿いの周辺部では，催し物当日と翌日を比較すると，ピーク時で半分程度の人の動きであった(図 5)．このように，催し物の開催による人の動きの差異を定量的に比較できることを確認した．

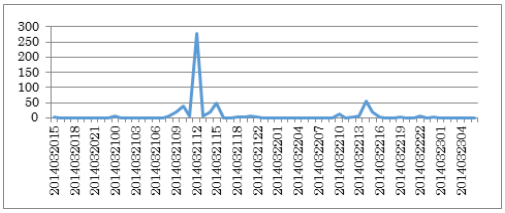


図 4: 主会場における人の動き

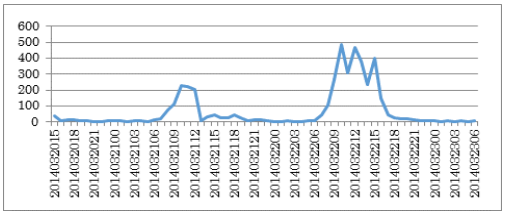


図 5: 周辺部における人の動き

(5)地域における人々の活動情報の収集基盤として，GPS と 3 軸加速度，3 軸角速度，3 軸磁気で構成される 9 軸センサを用いた動作計測システムを開発した．このシステムを利用し，運動中の動作分析を行った．動作計測機器は，頭部，腹部，両足首に装着する．複数の計測機器から得られたデータは，前処理ソフトウェアにより，GPS タイムスタンプを用いて同期される．この前処理データにもとづき，運動の種別を判別するために，自己相関関数を用いて，各時系列データ間の距離を定義した．距離の評価のため，距離にもとづく運動の類似度ランキングについて精度を定義し，階層的クラスタリングを行った(図 6)．これら精度および樹形図の観点から，動作，行動を分類する手法として本研究の提案手法が有効であることを確認した．

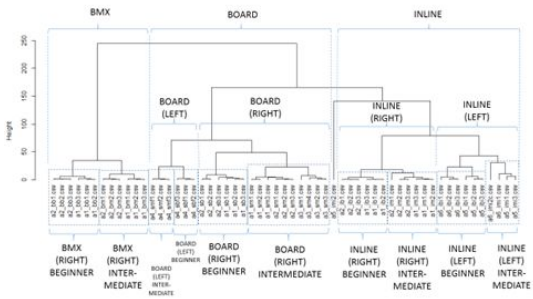


図 6：提案手法によるクラスタ

(6)マイクロブログにおける地域特有の単語および単語共起にもとづく情報発信者の位置推定手法を開発し，地域トピックの考察を行った．そのため，位置情報付きデータに含まれる単語について，出現頻度，緯度経度の平均，標準偏差を求め，単語や単語共起の地域特定スコアを定義し，位置推定を行った．この情報発信者の位置推定にもとづき地域トピックの考察を行った．用いるデータセットは，Twitter Streaming API を用いて，2011

年 3 月から 2014 年 5 月までに収集したツイートである．データセットにおけるツイート数は347,742,872件，単語数は4,124,568,983個，単語の種類は 58,994,705 種類，ユーザ数は 17,251,905 人である．また，位置情報付きツイートは1,132,580 件と全ツイートの0.33%，位置情報付きツイートを発信したユーザ数は311,812 人と全ユーザの1.8%である．

手法 1「単語を用いた位置推定」では，位置情報付きツイートから単語を抽出し，単語の出現した緯度経度の平均，分散を求める．ここで得られた単語の平均緯度経度について，国土地理院のデータにもとづき作成した緯度経度と住所のデータベースを用いて単語と住所の対応関係を抽出する．さらに単語の地域特定スコアを式(2)で定義する．

$$Score = tf \times \exp\left(-\sqrt{sx^2 + sy^2}\right) \quad \cdots (2)$$

各単語についての位置情報付き単語の頻度(tf)，経度の標準偏差(sx)，緯度の標準偏差(sy)を用いた式(2)の定義により，地理的分散が小さく出現頻度の単語は，地域を特定する単語として高いスコアを得る．このスコアにもとづき，全ツイートに含まれる単語を用いて，ユーザの位置推定を行う．各ユーザについて，ツイートから抽出した単語に対応する住所のスコアを加算する．これをこのユーザの全単語について行うことにより，ユーザの推定住所のランキングが得られる．このランキングでトップの地域をユーザの推定位置とする．

手法 2「地理的散らばり，頻度を制限した位置推定」では，地理的散らばりと単語頻度に閾値を設け，位置推定に用いる単語を制限し，地域特定スコアを定義する．これらの閾値の設定により，出現頻度が高く緯度経度の散らばりの大きい単語による，位置推定精度の低下を防ぐ．

手法 3「単語共起を用いた位置推定」では，単語共起における 2 つの単語のうち，一方の単語のみ，地理的散らばりや出現頻度の閾値を用いた単語共起について，方法 1 で定義した単語のスコアと同様に，単語共起にもとづく地域特定スコアを定義する．

定義した 3 種類の手法について，位置推定精度の評価を行うため，単語や単語共起にもとづく位置推定結果について，ツイートに付与された実際の位置情報との比較により評価を行う．ユーザの位置推定結果は，地域特定単語や単語共起による，位置スコアの合計にもとづき，推定された地域が順位付けされる．そのうち，一番得点の高い地域をユーザの推定地域とする．この推定地域と，実際にユーザが滞在した地域との距離にもとづき，位置推定結果を評価する．この評価方法にもとづき，3 種類の位置推定手法を比較する．図 7 は，推定された位置と実際の位置との誤差について，ユーザの度数分布の推移を表す．手法 1 （Method 1）では，誤差 250～300km のユーザ頻度が高い．手法 2（Method 2）では，誤差 50～100km のユーザ頻度が高い．位置推定に用いる単語に制限を加えることにより，誤った位置推定が低減していることがわかる．手法 3（Method 3）では，誤差 50km 以下のユーザ頻度が高い．単語(手法 1，2)の代わりに単語共起(手法 3)を用いることにより位置推定の精度向上が見られた．

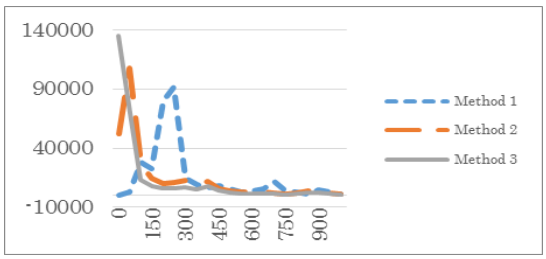


図 7：推定誤差とユーザ数

3 つの手法の中で最も高い精度の得られた手法 3 で構築した単語共起と位置の対応データベースにもとづき，データセット内の全ユーザの位置を推定し，ユーザのつぶやきを地域毎に集計することにより，地域別話題の推定を行った．手法 3 を全データに適用して位置推定できたユーザ数は，約 8 千万人(8,575,766)であった．これは全ユーザ数の約 50%に達する．データセットにおいて位置情報付きツイートを発信したユーザ数は，約 30 万人(311,812)であるため，約 28 倍の地域特定ユーザが得られたこととなる．

位置推定による地域情報の増大と確認するために，位置情報付きツイートを発信しているジオユーザと，データセットから位置推定されたユーザについて，発信された単語の種類，頻度について比較する．データセットにおいて位置情報付きツイートを発信したユーザのツイートのみを用いた単語の集計では，約 5 百万(5,032,683)種類の単語が抽出された．各単語の平均頻度は 30.45 であった．これに対して，位置推定ユーザのツイートを用いた場合，平均頻度は 613.51 と，20 倍以上の頻度であった．さらに，全位置推定ユーザのツイートのみに現れた単語数は 49,309,065 個で，単語の種類は 10 倍程度拡大した．このように，位置推定ユーザの利用により，得られる地域情報は飛躍的に増大したことが分かる．ただし，ユーザの位置は推定値であるため，完全に正確であるとは限らない．そのため，ユーザの推定位置の精度を定義し，その精度を単語頻度の重みづけとして用いることにより，精度の低いユーザの影響を低減する．そのため，各ユーザについて住所ランキングにある地域のスコアを重みとして緯度経度の重み付き平均と標準偏差を求める．ここで得られた平均緯度経度の地点を推定位置として用いる．また，緯度経度の標準偏差を sx，sy とし，住所ランキングの平均スコアを scave としたとき，ユーザ推定位置での重み(Weight)を式(3)で定義する．

この重みにもとづき，ユーザ推定位置での重み付き単語頻度(WTF)を式(4)で定義する．

$$Weight = scave \times \exp\left(-\sqrt{sx^2 + sy^2}\right) \quad \cdots(3)$$

$$\mathrm{W}TF = \left(1 - \exp\left(-\left(1 + Weight\right)\right)\right) \quad \cdots(4)$$

これらの定式化により，位置推定の平均スコアが高く，標準偏差が小さい場合，単語頻度の重みは重くなり，逆に，スコアが低く標準偏差が大きい場合は軽くなる．つまり，位置推定の精度の低いユーザについては，単語頻度を割り引いて集計することとする．WTF を用いて地域ごとに抽出した単語の時系列にもとづき話題の考察を行った．

都道府県レベルで WTF 集計し，各地域で単語頻度ランキングを作成したところ，都道府県名，市区町村名が単語頻度ランキング上位を占め，その他の項目としては地域特有の食文化や方言などが観察された．例えば，東京都の単語頻度ランキングにおいては，3 位「東京」(1,135,150)，11 位「新宿」(381,421)，12 位「渋谷」(360,807)であった．大阪府では，3 位「大阪」(1,525,890)，38 位「梅田」(135,327)といった地名に加え，2 位「ほんま」(2,413,362)，48 位「やから」(101,397)，58 位「やけど」(90,295)といった方言や，115 位「マクド」(39,296)，137 位「たこ焼き」(32,216)といった地域特有の食文化に関するキーワードが観察された．このように本手法において構築した地域特有の単語共起データベースにもとづくユーザの位置推定により，地域毎の話題抽出や情報組織化の基盤技術を構築することができたと考えられる．

## ５．主な発表論文等

〔雑誌論文〕（計　2 件）


Kazunari Ishida, "Identifying Existence Range of Diffusion Sources of Radioactive Small Particles," Global Journal of Human Social Science, peer reviewed journal, Vol 14, No 3-B, 2014, pp. 26 - 32.

Kazunari Ishida, "On a Field Investigation and Open Data Analysis to Identify Diffusion Sources of Radioactive Substance," International Journal of Environmental Science and Development, peer reviewed journal, Vol. 4, no. 3, 2013, pp. 291-295.


〔学会発表〕（計　10 件）


Kazunari Ishida, "Estimation of User Location and Local Topics Based on Geo-tagged Text Data on Social Media," 6th International Conference on E-Service and Knowledge Management (ESKM 2015), July 12 - 16, 2015, Okayama Convention Center, Okayama, Japan.

石田和成，"複数ウェアラブルセンサを用いたアクションスポーツの種目・技能レベルの分類"，電気学会　制御研究会，2015 年 3 月 28 日，香川高等専門学校(香川県高松市)．

Kazunari Ishida, "Estimating user location to identify area specific topics based on geo-tagged term co-occurrence," International Conference on Location-Based Social Media, March 14 – 15, 2015, Georgia Conference Center and Hotel, Athens, GA, USA.

石田和成，"地域特有の単語共起にもとづく位置推定と地域トピックの考察"，第 6 回テキストマイニング・シンポジウム，2015 年 2 月 5～6 日，ティーオージー会議室(大阪府大阪市)．

石田和成，"マイクロブログにおける地域固有表現にもとづく位置推定"，第 21 回社会情報システム学シンポジウム，2015 年 1 月 22 日，電気通信大学（東京都調布市)．

石田和成，"9 軸センサと GPS による動作計測システムの開発とアクションスポーツの動作解析"，第 35 回バイオメカニズム学術講演会，2014 年 11 月 8～9 日，岡山大学 鹿田キャンパス(岡山県岡山市)．

石田和成，"近距離無線通信を利用した階層的地域情報化システムの検討"，第 20 回社会情報システム学シンポジウム，2014 年 2 月 12 日，立正大学(東京都品川区)．

石田和成，"放射性物質拡散源検出のための統計的なオープンデータの分析"，第 20 回リモートセンシングフォーラム，2013 年 3 月 1 日，首都大学東京秋葉原サテライトキャンパス(東京都千代田区)．

石田和成，"オープンデータを用いた地域密着型放射性物質拡散予測システムのための検討"，第 19 回社会情報システム学シンポジウム，2013 年 1 月 23 日，電気通信大学(東京都調布市)．

Kazunari Ishida, "Extracting Geo-Social Information based on Geo-Tagged Social Media," 4th World Congress on Social Simulation (WCSS 2012), September 4 – 7, 2012, National Chengchi University, Taipei, Taiwan.


## ６．研究組織

(1)研究代表者


石田　和成（ISHIDA, Kazunari）
広島工業大学・情報学部・准教授
研究者番号：20700222