Vol.2013-MBL-69 No.3
2013/12/19

イベント参加者のフォロー関係に基づくイベント分類手法の提案

河野 慎[1]　米澤 拓郎[2]　中澤 仁[1,2]　川崎 仁嗣[3]　太田 賢[3]　稲村 浩[3]　徳田 英幸[1,2]

概要：近年 GPS が搭載されたスマートフォンや SNS の普及によって，ユーザがリアルタイムに位置情報を付加させた情報を発信できるようになってきた．これらの位置情報付き発言を収集・解析することで，人々が集まって形成されるソーシャルイベントを検出することが可能となる．ソーシャルイベントを検出するには発見と分類の２段階の過程があり，本研究ではイベントが発見された後の分類手法を提案する．イベントには特徴・性質として内容・規模・大衆性の３つがあると考え、分類軸として大衆性に着目する．位置情報を付与させて発言しているイベント参加者のフォロー関係を解析することで大衆性の推定をし，イベントの分類を目指す．本研究ではリアルタイムに解析を行えるツールを設計・実装し，大衆性の推定手法について考察を行った．

1. はじめに

近年 Twitter や Facebook，Foursquare といったソーシャルネットワークサービス（SNS）が普及し，ユーザ一人ひとりが容易にかつリアルタイムに情報を発信することが可能になってきた．このユーザによって発信された情報には，ブログやウェブサイトに関する情報などの拡散だけではなく，ユーザが見たり，経験したことについてなどが含まれている．IT 技術の発展によってセンサが開発され，様々な事象についてセンシングすることが可能になってきたが，未だそれができていない事象も存在する．しかしそういった事象をユーザが経験し，SNS に情報を発信することで，従来では得ることができなかった情報を取得することが可能となる．このようにユーザをセンサとみなして情報を収集し，何らかの知見を得ようとする参加型センシングと呼ばれる研究が注目されるようになってきている．

また IT 技術の発展によって，GPS センサが普及されるようになってきた．GPS を用いることで，バスなど車両の現在位置などをリアルタイムに特定することが可能になった．東日本大震災のときも自動車の GPS センサを用いて道路の混雑状況などを把握した例も存在する．GPS センサはユーザ一人ひとりが持つ携帯端末にも搭載され，端末をなくしてしまった時に見つけたりするサービスやチェックインと呼ばれるサービスが提供されるようになってきた．

1　慶應義塾大学 環境情報学部
2　慶應義塾大学大学院 政策メディア研究科
3　株式会社 NTT docomo

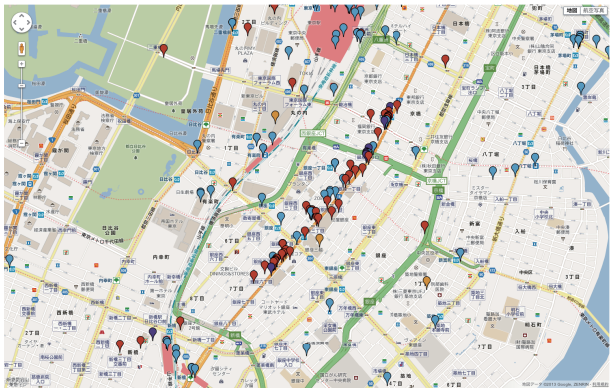

図 1　iPhone5 発売日の様子

チェックインは特に SNS と共に利用されることが増え，またチェックインだけではなく，発信される情報にそのユーザの位置情報を付与させることも多く見受けられる．

この位置情報が付与され，SNS を用いて発信された情報（位置情報付き発言）を解析することで，人々が集まって形成されるイベント（ソーシャルイベント）を発見することが可能となる．ソーシャルイベントの一例として図 1 に Apple iPhone5 発売日の AppleStore 銀座店付近の様子を示す．ピンが位置情報付き発言を表しており，この図から行列の様子を知ることができる．

このように位置情報つき発言から地震など特定のソーシャルイベントの検出は多く試みられている．しかしイベントの種類はたくさんあり，これらを対象に検出するためにはイベントを検知した後，さらに分類する必要がある．そこで本研究では，ソーシャルイベントを検知した後に分

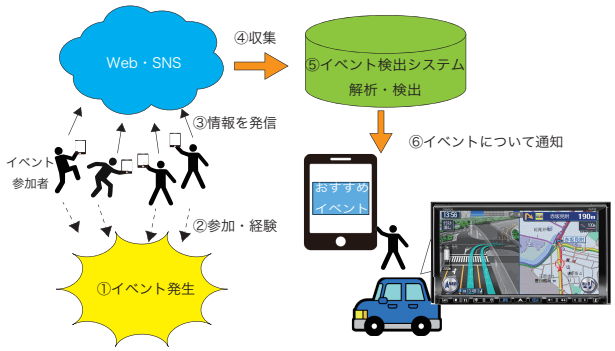

図 2 本研究応用システム例

類するための手法を提案する．大衆性という新しい分類軸を定義し，イベント参加者のフォロー関係を分析することでイベントの分類を可能にする．本研究では提案した手法を実装し，ユーザがインタラクティブに操作可能なインタフェースを実現した．その結果イベント参加者のフォロー関係から大衆性だけでなく，イベントの内容を推定できる可能性もあることがわかり，今後の可能性を示唆することができた．

2. 社会イベントの検出と分類

2.1 動機

近年ソーシャルイベントに対するニーズは高まってきている．ゲリラ豪雨や地震など現在の技術を用いても予測が難しいものが増えてきていることや，人の属性に合わせたマーケティング広告，さらに都市計画などに利用することができるからである．本研究が実現し，ソーシャルイベントの性質をもとに分類ができることでこういったニーズに応えることが可能になる．図 2 に本研究を応用した推薦システムの例を示す．この例では，以下の手順で推薦が行われる．

(1) ソーシャルイベントが発生する．
(2) イベントにユーザが参加・目撃などをする．
(3) 参加したことなどについて SNS で発言する．
(4) SNS 上の情報を収集する．
(5) 本研究を応用したシステムがイベントを検知し，そのイベントの性質から分類をする．
(6) その性質とイベントの近くを通ったユーザの興味を比較し，合致している場合にそのイベントをユーザにプッシュ通知などを通じて推薦をする．

上記のシステムを実現するためには．ソーシャルイベントのリアルタイムな検知と分類手法の提案をする必要がある．

2.2 関連研究

ソーシャルイベント発見をする研究は数多く存在する．Sakaki ら [1] は Twitter から地震の検出に，James ら [2] は Twitter からスポーツ観戦におけるイベントの検出に成功している．これらの研究では，予めキーワードを指定し，SVM を用いることでツイートがキーワードに関するものかどうか解析し，その結果をもとに検出を行っている．また群集行動を解析してイベントを検出している Lee R[4] らによる研究もある．これは位置情報をもとに局所的なイベントとしてある地域の祭りや花火大会を検出している．これらの研究は予め指定した特定のイベントを検出することに成功しているが，複数の種類のイベント検出は行っていない．複数のイベントの検知を行っている研究は存在する [5] が，分類にまでは至っていない．

2.3 目的

本研究は特定のソーシャルイベントだけではなく，複数のソーシャルイベントをリアルタイムに分類する手法の提案を目的とする．本研究が実現することで分類したイベントに合わせたリアルタイムな推薦システムやマーケティング広告，都市計画に利用することが可能になる．

3. イベント分類手法の提案

3.1 分類軸

テキストマイニングはメールのスパムフィルタなど様々なところで用いられている手法である．しかし Twitter はメールなどの文面とは違いその表現方法も多種多様になっていること，テキストの量が制限されていることがあるため，テキストマイニングを行っても有益な結果を得ることが難しい．これらの問題を解決するため，Nishida ら [8] のデータ圧縮や Sriram らの発言者情報の追加 [9] など様々な手法が提案されている．しかしテキストマイニングによってイベント名を特定できただけでは，ユーザに推薦するかどうか評価するのは難しい．ゆえにテキストマイニングによるイベント名以外の評価・分類軸が必要となる．分類軸の一つとしてイベントの規模が考えられる．しかしイベントに参加者しているユーザ全員が位置情報付き発言をするわけではない．したがって位置情報付き発言のみでそのイベントの参加者数や規模を推定することはできない．

そこで本研究ではイベントの新しい分類軸として**大衆性**という属性を提案する．大衆性とは大衆に受け入れられる性質を意味する．本研究では大衆を様々な属性をもつ人々と捉え，参加者集団の属性の均等性，すなわちイベント参加者の多様性を示すものとして大衆性を扱う．イベント参加者集団の属性が偏っていれば大衆性は低く，偏りが少なければ大衆性が高いと考えられる．大衆性の高いものとしては花火大会などが挙げられ，逆に低いものとしてはサークルの飲み会などが挙げれられる．一般に集団の属性を分析する手法としてクラスタリングがある．イベント参加者集団のクラスタリングをした際に，クラスタ数が多ければ多いほど大衆性が高いといえる．大衆性が高いイベントの場合はあらゆるユーザに推薦をすることができ，また大衆性が低いイベントの場合でも，もしそのイベント参加

者の属性とあるユーザの属性が一致していれば推薦をすることができる．このように大衆性を評価することでユーザに推薦するかどうかの評価が容易になる．

3.2 既存手法・問題点

Twitter などで相互フォロー関係に注目し，クラスタリングによるコミュニティ分類の研究は数多く存在する．多くの研究で用いられているグラフ理論における手法の一つに Clique Percolation Method（CPM）[6][7] が存在する．しかし Twitter で CPM を用いようとした場合，２つの問題がある．１つ目は CPM の計算困難性である．クリーク間で共通ノード数を調べるため，計算量が膨大になってしまうためである．２つ目はユーザのフォロー情報を取得するまでの時間である．CPM はクリークを利用してコミュニティを推定するものであるが，推定するためにはノード間のリンクを少なくとも２ホップ先まで取得する必要がある．イベント参加者 n 人に対して，CPM を使う場合は，参加者一人あたり平均 100 人のユーザをフォローしているとすると，Twitter API を少なくとも $100n$ 回呼ぶ必要がある．2013 年 6 月 14 日の Twitter API 制限の変更に伴い，Twitter API は 15 分間で 15 回しか呼ぶことができなくなり，2 ホップ先のユーザのフォロー関係を取得するまでの時間が相当要してしまう．以上のことから，Twitter に CPM を用いてイベント参加者間におけるつながりによる大衆性を推定することが難しい．

3.3 アプローチ

大衆性を推定するためにイベント参加者のコミュニティ分類以外の手法で考える必要がある．本研究ではイベント参加者の興味に着目する．イベント参加者の興味は Twitter においてそのフォローしているユーザ（フレンド）に現れているとし，イベント参加者のフレンドを解析する．取得するフレンドの情報を 1 ホップ先のみにすることで，TwitterAPI を呼ぶ回数を n 回にすることが可能になる．イベント参加者 a がフォローしているフレンドの集合を F_a，イベント参加者全員を $A = \{a|$ イベント参加者 $\}$ とすると，フレンドの集合 P は

$$P = \bigcap_{i \in k} F_i (k \subseteq A) \qquad (1)$$

と表せる．図 3 にフォロー関係の様子を示す．図 3 左のようにイベント参加者の多くが特定のフレンドをフォローしている場合，このイベント参加者はある共通の興味・関心（特定のフレンド）をもつといえることから，このイベントは大衆性が低いといえる．逆に図 3 右のようにイベント参加者のフォローしているフレンドが特定のフレンドに集中せずに分散している場合，イベント参加者の共通の興味・関心はないため，このイベントの大衆性は高いといえる．このようにイベント参加者のフォロー関係を解析することでイベントの大衆性を推定することが可能となる．

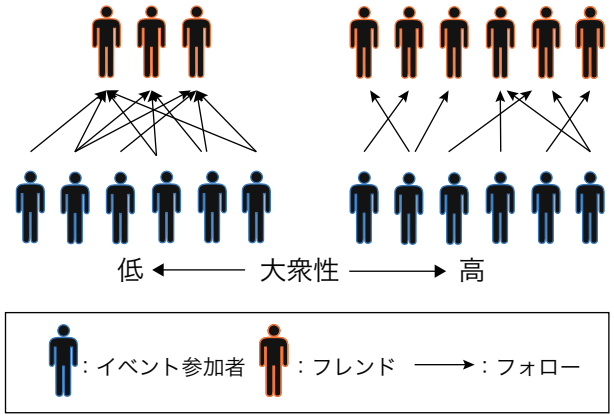

図 3 イベント参加者のフォローの様子

4. 設計・実装

本研究では，リアルタイムにデータの取得・解析が可能かつユーザがインタラクティブに操作可能なインタフェースをもつツールの設計と実装を行った．

4.1 解析ツール

本研究ではイベントの分類が目的であり，検出は既存手法の利用を想定する．そこで図 4 のような直感的にイベントを発見することを支援するツールを実装し，手動でイベントの発見・ブックマークを可能にさせる．発見・ブックマークされたイベント名は図 5，6 の左側に一覧表示される．一覧の中から選択されたイベントを解析してグラフ（図 5）で表示し，また特定のフレンド集合 P をイベント参加者からのフォロー獲得数によるランキング（図 6）で表示する．

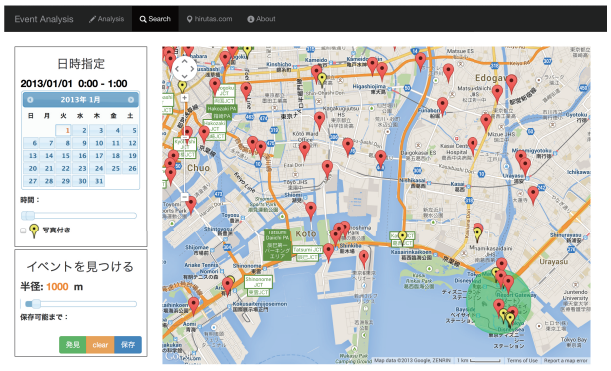

図 4 イベント発見ツール

4.2 システム構成

図 7 に解析ツールのシステム構成図を示す．Twitter から Streaming API を利用して日本国内の位置情報つき発言を取得し，整形した後，TweetDB に保存していく．次

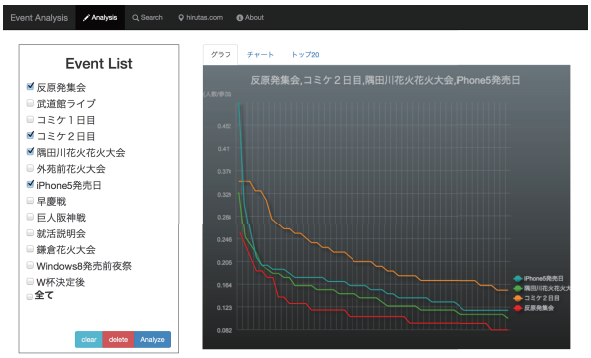

図 5 解析ツール 1

図 6 解析ツール 2

に図 4 のように地図上に発言をピンで表示し，イベントを発見する．発見ツールで囲まれた発言をしたユーザがフォローしているフレンドの情報を Twitter から Rest API を利用して取得する．その後取得したフレンドの情報を解析し，その結果を表示する．

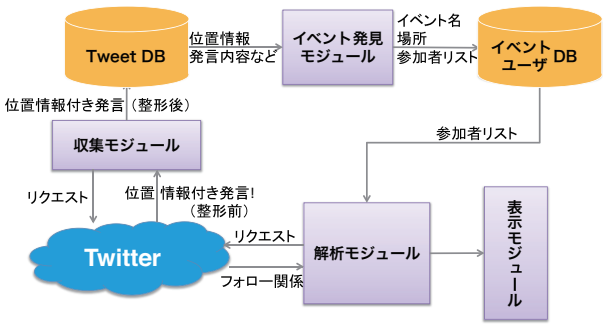

図 7 本システム構成図

5. 解析結果・考察

解析結果の一部を図 8，表 1，2 に示す．これらは 2012/12/29 のコミケ初日と 2012/7/27 の FUJI ROCK FESTIVAL を解析したものである．図 8 のグラフはイベント参加者のフレンドの集合 P をイベント参加者にフォローされている割合から降順に並べたものであり，横軸はフレンドの集合 P を，縦軸はその割合を意味している．図 8 に示されるようにべき乗分布に従っていることがわかる．べき乗分布には「パレートの法則（80：20 の法則）」と呼ばれる法則がある．これは主要な一部（上位 20 %）が全体の大部分（残り 80 %）に影響を持っていることが多いというものである．パレートの法則をこの結果に当てはめると，「イベント参加者に多くフォローされている上位 20 %のフレンドがイベント全体を表している」という仮説がたつ．

表 1，2 はこの仮説をもとに多くのイベント参加者にフォローされているフレンド 20 %のうち上位 5 人とそのフォローされている割合を示している．この上位 5 人のフレンドを見てみるとコミケでは 1 位に声優の田村ゆかり，2 位に艦これのアカウントが来ている．FUJI ROCK FESTIVAL では 1 位，2 位ともに FUJI ROCK FESTIVAL に関連するアカウントになっている．これらを仮説に当てはめると，このアカウントがイベントを表していることになり，それぞれのイベントを考えると妥当であるといえる．このようにテキストマイニングをしなくてもイベント参加者のフレンドを解析することでイベントがどんな内容・性質をもつのかがわかるといえる．

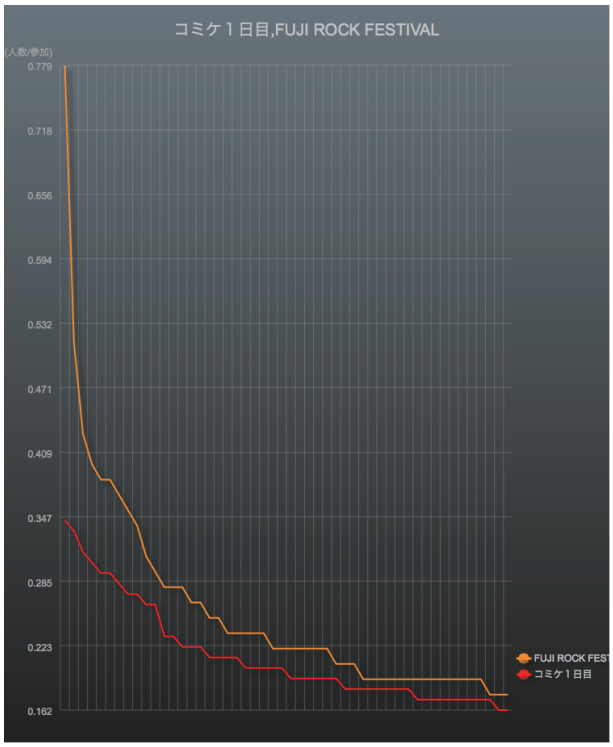

図 8 解析結果

表 1 コミケ 1 日目

順位	アカウント	フォロー数/参加者
1 位	田村ゆかり@ 11/20 新アルバム発売	0.343
2 位	「艦これ」開発/運営	0.333
3 位	geek@akibablog	0.313
4 位	竹達 彩奈	0.303
5 位	NHK 広報局（ユル〜く会話しますよ）	0.292

表 2 FUJI ROCK FESTIVAL

順位	アカウント	フォロー数/参加者
1 位	FUJI ROCK FESTIVAL	0.779
2 位	fujirockers.org	0.514
3 位	孫正義	0.426
4 位	Creativeman	0.397
5 位	Radiohead	0.382

表 3 回帰分析（昇順）

イベント名	α
隅田川花火大会	-0.49115
外苑前花火大会	-0.4872
コミケ 1 日目	-0.45234
東京モーターショー	-0.43046
鎌倉花火大会	-0.41675
東大五月祭	-0.38829
早慶戦	-0.33782

表 4 ジニ係数（降順）

イベント名	ジニ係数
東京モーターショー	0.41458
コミケ 1 日目	0.19097
隅田川花火大会	0.167438
外苑前花火大会	0.15869
日吉セレモニー	0.08023
東大五月祭	0.08017
早慶戦	0.0359

また，べき乗分布の曲線を数式で表すため，回帰分析を用いた [10]．図 8 の両軸を対数にとって回帰分析を行い，曲線の式 $y = x^{\alpha}$ の α を求める．α の値の昇順に並べ替えた一部が表 3 である．α の意味は曲線の曲がり具合を示すものであり，値が大きければ大きいほどべき乗曲線が緩やかになり，割合も全体的に値が大きいことになる．昇順に並べたことで一般的に大衆性が高いと考えられるものが上位に，低いと考えられるものが下位に来ている．すなわちイベントの参加者たちがフォローしている共通のフレンドが多いほど一般的に大衆性が高いことを意味している．

また経済学においてべき乗分布で用いられるジニ係数を解析に用いた．ジニ係数はある集団において所得分配の不平等性を示すものであり，係数の値が 0 に近いと格差が小さく，逆に 1 に近いと格差が大きいことを意味している．このジニ係数を図 8 に当てはめた場合，フレンド集合 P のイベント参加者からのフォローが分配されているかを表すことになる．そして係数の降順にイベントを並べると表 4 のようになる．これも回帰分析と同様に，一般的に大衆性が高いと考えられるもの，低いと考えられるものが対比的に並んでいる．大衆性が高いと考えられるイベントのジニ係数が高いということはフレンドのイベント参加者からのフォローが偏っていることを意味する．

図 9 は回帰分析とジニ係数を軸にした散布図である．相関係数は -0.654 となり，負の相関関係が回帰分析とジニ係数には存在する．散布された各点を見ると左上に早慶戦や日吉セレモニーといった大衆性が低いと考えられるイベントが来ており，右下に行くに連れて，徐々に大衆性が高いと考えられるイベントが来ている．また K-means 法を適用した場合図 9 のように以下の 4 つのクラスタに分類することができた．

- 規模が大きく大衆性の高いイベントクラスタ
  隅田川花火大会，コミケ
- 規模はそこまで大きくないが，大衆性が高く地域に根付いているイベントクラスタ
  厚木鮎まつり，東大五月祭
- 目的が明確かつ大衆性が高いイベントクラスタ
  反原発集会，プロ野球巨人阪神戦
- ある集団内輪にむけた大衆性の低いイベントクラスタ
  早慶戦，日吉セレモニー

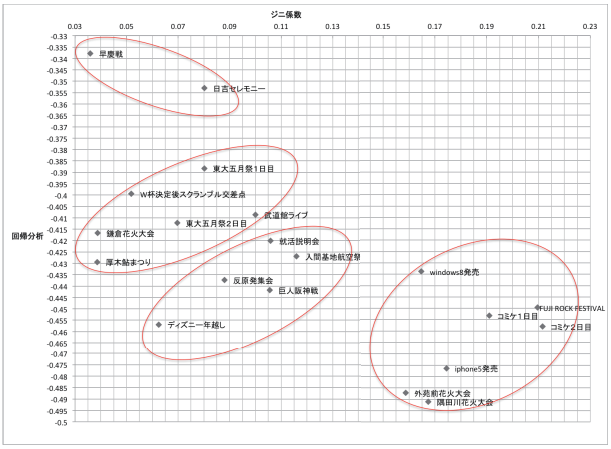

図 9 回帰分析とジニ係数の散布図
k=4 の K-means 法によるクラスタリング

6. 議論

イベントの大衆性はジニ係数と回帰分析によって推定ができることがわかるが，大衆性の定義に曖昧さが残っている．あるイベントの大衆性を考えたときに，その考える人の年齢や住んでいる地域，周りの環境によって変わってくるからである．また複数のイベントの大衆性を考えた場合に，相対的であることも多い．そのため，本研究によって推定される大衆性とユーザが考える大衆性との誤差を評価する必要がある．二項比較法などを用いて，実験協力者に予め用意したソーシャルイベントのリストを大衆性の高い順に並べてもらい，これを正解データとする．そして本研究で推定された大衆性によるイベントの並び順とこの正解データの比較を行うことが必要となる．

また大衆性を推定するためのパラメータに本研究では回帰分析とジニ係数を用いたが，これらのパラメータの選択が妥当であるか，あるいはパラメータの数が十分であるかの判断も同様に評価をするべきである．他の分析手法 [11] を利用して推定した大衆性と，回帰分析やジニ係数によって推定された大衆性あるいはいくつかのパラメータを組み合わせて推定された大衆性の比較を行うべきである．K-means 法を用いる際もパラメータを増やして 3 軸，4 軸と評価する軸を増やして評価をするべきである．本研究では Twitter API の制限もあり，手法を既存手法である CPM などに比べて情報量，計算量ともに少ないが，これも妥当であるかどうかを定量的に計算時間や精度を比較・評価をするべきである．

今回実装したツールを用いてイベント参加者の発言を収集した．しかしその中には参加していないユーザの発言も含まれている可能性があり，結果的に解析の精度を下げていることが考えられる．情報の信頼度を上げるために蛭田ら [12] や Carlos ら [13] はフィルタリングなどの手法を用いている．今後の解析においてこの信頼度についても考慮する必要がある．

7. まとめ

近年 Twitter などの SNS や GPS が搭載された携帯端末が普及し，ユーザがリアルタイムに位置情報付き発言を発信できるようになってきた．この位置情報付き発言を解析することでソーシャルイベントの検出が可能となる．本研究は，イベント参加者のフォロー関係を解析することでソーシャルイベントの分類ができる手法を提案した．分類する際の評価軸として参加者の多様性を意味する大衆性を提案し，推定するためのツールを実装した．実装したツールを用いてイベントを発見・解析し，その解析結果について考察を行った．本研究によって推定された大衆性とユーザが考える大衆性を比較して評価をすることが今後の課題として挙げられる．

参考文献

[1] Sakaki, Takeshi, Makoto Okazaki, and Yutaka Matsuo. "Earthquake shakes Twitter users: real-time event detection by social sensors." Proceedings of the 19th international conference on World wide web. ACM, 2010.
[2] Lanagan, James, and Alan F. Smeaton. "Using twitter to detect and tag important events in live sports." Artificial Intelligence (2011): 542-545.
[3] Thelwall, Mike, Kevan Buckley, and Georgios Paltoglou. "Sentiment in Twitter events." Journal of the American Society for Information Science and Technology 62.2 (2011): 406-418.
[4] Lee, R., Sumiya, K. Measuring geographical regularities of crowd behaviors for Twitter-based geo-social event detection, Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Location Based Social Networks(2010).
[5] Becker, Hila, Mor Naaman, and Luis Gravano. "Beyond Trending Topics: Real-World Event Identification on Twitter." ICWSM. 2011.
[6] Palla, Gergely, et al. "Uncovering the overlapping community structure of complex networks in nature and society." Nature 435.7043 (2005): 814-818.
[7] Palla, Gergely, Albert-Lszl Barabsi, and Tams Vicsek. "Quantifying social group evolution." Nature 446.7136 (2007): 664-667.
[8] Tweet-Topic Classication using Data Compression, Kyosuke NISHIDA, Ryohei BANNO, Ko FUJIMURA, and Takashide HOSHIDE,NTT Cyber Solutions Laboratories, NTT Corporation, 2011
[9] B. Sriram, D. Fuhry, and M. Demirbas, "Short text classification in twitter to improve information filtering," Proceedings of 33rd Inter- national ACM SIGIR Conference on Research and Development in Information Retrieval, pp.841—842, 2010.
[10] Si, Si. "ベキ乗分布に基づく環境情報の統計学的分析とその応用に関する研究."
[11] 栗原一貴, and 土谷洋平. "ロングテール時代のための中心極限定理によらない統計分析手法." 情報処理学会論文誌 52.2 (2011): 477-487.
[12] 蛭田慎也, 米澤拓郎, and 徳田英幸. "場所誘因型位置情報付き発言の検出と可視化." 情報処理学会論文誌 54.2 (2013): 710-720.
[13] Castillo, Carlos, Marcelo Mendoza, and Barbara Poblete. "Information credibility on twitter." Proceedings of the 20th international conference on World wide web. ACM, 2011.