3ZA-3

Twitterタイムライン解析による存在感の抽出

松田 有史† 杉山 治‡
慶應義塾大学 理工学部 株式会社 国際電気通信基礎技術研究所
大澤 博隆§ 今井 倫太¶
独立行政法人 科学技術振興機構 慶應義塾大学 理工学部

yushi@ayu.ics.keio.ac.jp, sugiyama@atr.jp, osawa@nii.ac.jp, michita@ayu.ics.keio.ac.jp

1 はじめに

本研究では，Twitter[1] タイムラインから個人の特徴を抽出し，特徴モデルを作成する手法を提案する．Twitter には，個人のライフログとしての側面があり，Twitter タイムラインは個人の社会的な写像であると考えられる．そのため，Twitter タイムラインから個人の特徴を捉えることで，従来では明らかにならなかった存在感の一面を明らかにすることができると考えられる．

ここで，個人の存在感とは，行動パターン，人間関係といった Twitter から分析できる個人の特徴のパターンである，それら個人に特有のパターンを抽出することにより，個人同定や個人の状態の検出をすることができると考えられる．また，得られた特徴は個人のアバターを創り出すといったアプリケーションや個人に適切な情報を発信するといった推薦システムなどを実装するための基本モデルになると考えられる．

従来でも，パーソナリティを推定するシステムは存在する．しかし，性別，年齢など一般的な分類に関するものがほとんどで，細かな個人差まで抽出することができなかった．そこで，本研究では，細かな個人差まで抽出するため，個人の特徴や存在感といったものがどのような点に現れるか検証し，特徴モデルを形成する方法を提案する．

2 関連研究

Twitter から特徴抽出をするシステムは，これまでも存在した．例えば，KDDI 研究所 [2] では個人のプロフィールを推定することを目的とした推定システムを開発した．いくつかの特徴語を用いることで，ユーザの年齢，性別，地域を推定することができる．一方，富永らは [3]Twitter エージェントに地域活性を目的とした存在感を持つボットを開発した．しかしながら，これらの研究は人一般に焦点を当てていて，個人差といったものに踏み込んでいない．したがって，本研究では個人差や社会的な存在感に焦点を当てる．

Clustering ofTwitter Time Line
†Yushi MATSUDA
Faculty of Science and Technology, Keio University
‡Osamu SUGIYAMA
Advanced Telecommunications Research Institute International
§Osawa HIROTAKA
Japan Science and Technoloogy Agency
¶Michita IMAI
Faculty of Science and Technology, Keio University

3 社会的存在感

本研究では，社会的存在感を Twitter に現われる個人の「その人らしさ」と定義する.「その人らしさ」とは, 行動パターン，人間関係といった twitter から分析できる個人の特徴のパターンであり，これらの特徴は以下の条件を満たす必要がある，

1. 個人の tweet の中で繰り返されるパターンであること
   個人の tweet の中で一定間隔, 条件で繰り返される共通する特徴である．これらを抽出するためには，個人の一定期間（本稿では仮にその間隔を一週間と定めた）のツイートの集合から，共通するパラメータを抽出すればよい．
2. 上記のパターンのうち，他者と異なるものであること
   その人らしさとは, 他者との比較によって生じる．例え, 個人の tweet の中で頻出するパターンであったとしても, 他者も似たようなパターンを持つ場合, そのパターンは個人特有の特徴とはいえない

4 社会的存在感抽出システムの提案

4.1 抽出手法

前章で定義した社会的存在感を抽出するためには，上記の 2 つの条件を満たす個人の特徴パターンを探索する必要がある.. 本研究では，以下のような手順で個人の特徴を抽出することを試みる．

1. Twitter から一定人数（本研究では 100 名）のツイートを取得し, その種類ごとにタグ付けする．
2. タグ付けられた情報から, あるユーザの一週間分のつぶやきをまとめて, 一定間隔 (6H) 毎にその比率, つぶやき頻度 (tweet/h) を計算する． これをマトリックスの 1 エントリとして捉える（従って, 同じユーザの違う週のデータは別のエントリとして捉えられる）
3. 計算したパラメータから構成できるマトリックスの組み合わせを全て作成する
4. 全てのマトリックスに対して，クラスター分析を行う．
5. クラスタリングの結果から, 同ユーザのエントリが同じクラスターにあり，かつその距離が最も近い，特徴の組み合わせを探索する

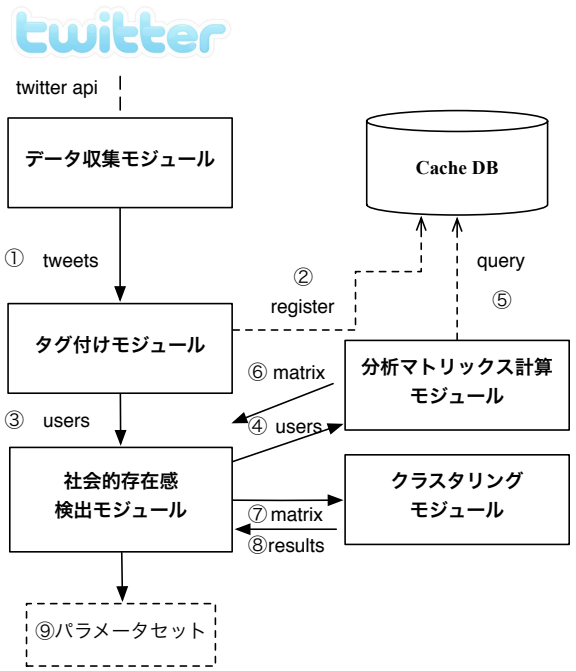

図 1　システム構成図

6. 得られた特徴の組み合わせをそのユーザの社会的存在感として捉える

4.2　提案システム

前節で提案した手法を実現するために, 本研究では図 1 に示されるシステムを提案する.

システムは, まず, twitter の API を用いて, 100 名のユーザの最新のつぶやき 200 エントリを収集する. その後, タグ付けしたデータをキャッシュデータベースに格納し, ユーザのリストを社会的存在感検出モジュールに出力する. 社会的存在感検出モジュールは, 前節で提案された手順に従い, パラメータマトリックスを計算, そのマトリックスをクラスタリングモジュールに入力し, クラスタリング結果に基づいて, 最終的な個人の特徴パターンを検出する. 下記にそれぞれのモジュールで行われる処理を簡単に示す.

4.3　モジュール

- データ収集モジュール

  TwitterAPI からツイートを取得する.

- タグ付けモジュール

  ここでは, ツイートにタグを付けていく. 付けるタグは「ツイート」「返信」「公式リツイート」「非公式リツイート」「なう」「写真」「ＵＲＬ」の７つである. これらは Twitter の仕様に関わるものとユーザが決めたルールのようなものがあり, Twitter 上で顕著に見られる特徴パターンである.

- 分析マトリックス計算モジュール

  本研究では, クラスター分析のため, データマトリックスを形成した. まず, ツイートの種類として, ツイート, 返信, 公式リツイート, 非公式リツイート, なう, URL, 写真の 7 つをタグを利用した.

  また, 時間軸を朝（6 時〜12 時）昼（12 時〜18 時）夕（18 時〜24 時）夜（0 時〜6 時）の 4 つに分けた. 本研究ではこの 4×7 の 28 次元を個人のパラメータとして使用した.

  また, パラメータを形成する数値はツイート回数ではなく, 比率とし, 以下のように定義した.

  あるタグの一定間隔のつぶやきの総数を t_i とするとその比率マトリックス M は以下の式で表される.

$$M = \left\{ \frac{t_i}{\sum_{k=1}^{n} t_k} \right\}_{i=1 \ldots n}$$

  ツイート回数を使用すると, Twitter をよく使用するユーザの数値が大きくなりすぎ, ツイートが多いか少ないかといった違いしか見られないため, 各パラメタのツイート比率を使用した.

- クラスタリングモジュール

  本研究では, データマイニングにクラスター分析を使用した. クラスター分析とは教師なし学習の手法の一つである. 個体間の類似度により, 似ているもの同士を集めて集合（クラスター）に分類する方法である. 本研究では階層的クラスター分析の一種であるウォード法を使用してクラスター分析を行った.

  その人らしさとは, 未知のデータであるためクラスター分析が適当であった. また, 分かれるクラスタの個数も未知であったため, 階層的クラスター分析を適用した.

- アウトプット

  クラスター分析によって得られた各クラスターのパラメータセットが出力される.

5　まとめ

本稿では, Twitter タイムラインから個人の社会的存在感を抽出する方法を提案した. 今後は, 提案したシステムによって得られた特徴セットが個人のどんな特徴を示しているのかを解析し, 特徴パターンの活用法についてより詳細に分析する予定である.

参考文献

[1] Twitter: ”http://twitter.com/”

[2] KDDI 研究所: “「つぶやき」から投稿者のプロフィールを自動推定する技術の開発に成功〜口コミ投稿者の年齢や性別, 趣味などが推定可能に〜”, 2010.

[3] 富永祐衣, 山本吉伸, 椎尾一郎: “おんせんはいったー〜外湯巡りシステムと連動した Twitter エージェント〜”. 情報処理学会創立５０周年記念（第７２回）全国大会, 2010.