4A1-1

文字 n-gram 法を利用した Twitter からの行動抽出

Behavior Retrieval from Tweets using Character N-gram Models

矢野　裕司　Yuji Yano
橋山　智訓　Tomonori Hashiyama
市野　順子　Junko Ichino
田野　俊一　Shun'ichi Tano

電気通信大学大学院情報システム学研究科
Graduate School of Information Systems, The University of Electro-Communications

This paper focused on retrieving human behavior from the tweets. When using Twitter, we may usually use domain-specific terms and post incorrect sentences. These perspective on Twitter make us hard to analyze tweets within grammatical manner or existing dictionaries. To tackle them, we are applying character n-gram tokenization and naive Bayes classifier to extract appropriate behavioral information from tweets. Using n-gram tokenizer, domain-specific words can be identified and incorrect grammar can be handled. Some experiments are carried out using actual tweets to show the feasibility of our approach.

1. はじめに

近年，ユーザの行動情報やユーザの置かれた環境情報，位置情報から，その場の状況に適したサービスを提供する，コンテキストアウェアサービスが注目されている．コンテキストアウェアサービスでは，行動情報や環境情報，位置情報のようなコンテキスト情報を適切に取得する必要がある．コンテキスト情報のうち，行動情報はユーザの生活習慣や興味，意図を知るための重要な手がかりであり，マーケティングのための行動予測などに活用することができる．

一方，ソーシャルネットワーキングサービスの一種である Twitter が近年普及しており，膨大な投稿の中には行動を表すものも多く存在している．Twitter での投稿は tweet と呼ばれている．Twitter では，ユーザはパーソナルコンピュータやスマートフォンなどを用いて気軽に投稿を行うため，ユーザが考えていることや行ったことをリアルタイムで取得することができる．さらに，Twitter は Web ページや一般的なブログと比べて投稿頻度が高く，行動内容や行動の時間などをより細かく取得することが可能である．これらの点から，行動情報取得の対象として tweet データを用いた場合には，粒度が高い行動情報が抽出できると考えられる．しかし tweet データは，Web ページや一般的なブログと比べて，文法として不正確な場合や Twitter 特有の単語を用いる場合が多く，形態素解析といった従来の自然言語処理の手法では，適切に行動情報を抽出することが難しい場合がある．

この問題に対処するため本研究では，文字 n-gram 法を用いることにより，行動を表す tweet データを抽出する手法を提案する．文字 n-gram 法ではわかち書きを行わず，原則的に文章における n 文字以下のすべての素性を考慮することができる．従って，形態素解析における誤りを考慮する必要がなく，Twitter 特有の単語も扱うことができる．また辞書を用いた手法では扱うことのできない略語や新語についても，使用頻度が高いものであれば扱うことができるという利点がある．これらの点で，tweet データから行動情報を取得する場合に有効な手法であると考えられる．

連絡先: 矢野裕司，電気通信大学大学院情報システム学研究科，〒 182-8585　東京都調布市調布ヶ丘 1-5-1, yano@media.is.uec.ac.jp

2. 関連研究

本研究と同様に，Twitter から行動を抽出する研究が行われている [Banerjee 09][Nguyen 12]．Banerjee らは，英語の tweet データ中の動作とカテゴリ，時間を表す単語の共起頻度に基づいて行動を抽出している．また Nguyen らは，日本語の tweet データの文章構造から動作主，動作，対象，時間，位置といった動作の属性を抽出している．これらの研究では，形態素解析を利用しているため，誤った構文や Twitter 特有の単語を考慮することができない．著者らは，構文を考慮せずに Twitter 特有の単語を扱うために，行動を表す単語を集めた辞書を構築し，この辞書を用いて行動情報を抽出する手法を提案 [矢野 13] した．しかしこの手法では，人手で行動を表す単語を選択するため，手間がかかるという問題点がある．そこで本研究では，文字 n-gram 法により文章を特徴化し，行動を表す投稿を分類することで，少ない手間で行動情報を抽出する。

3. 提案手法

本研究では，行動を表す tweet データを抽出する手法を提案する．本手法の流れを図 1 に示す．まず投稿された tweet は，文字 n-gram 法により素性として分解される．次に，この素性

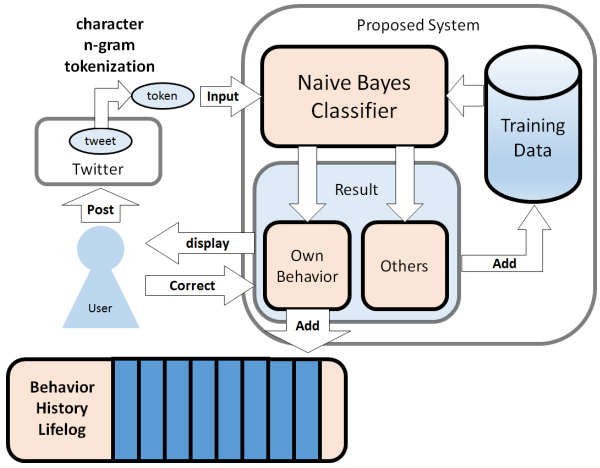

図 1: 提案手法の流れ

を入力として，ナイーブベイズ分類器により，行動を表すか否かに分類される．分類器により分類された tweet は，一日分のようなある一つのまとまりでユーザに提示される．ユーザは，クリックやタップ操作といった簡単な方法により，分類器による誤分類の結果のみを修正する．ここで行動を表すと分類された tweet は，個人の行動履歴として蓄積される．修正された分類結果および修正されなかった分類結果は学習データに加えられ，以降の tweet を分類する際に使用される．これを繰り返すことにより，学習データの規模は大きくなり誤分類が減少するため，ユーザの修正のコストが減少し，高精度で分類することができる．また，学習データは他人のユーザとも共有されており，最初の使用時であっても他のユーザの tweet を利用して分類することもできるようになっている．

4. 実験

4.1 実験方法

提案手法の有用性を示すために，実際の tweet を用いて分類実験を行った．今回の実験では，無作為に選んだ 10 ユーザの 2012 年 1 月 21 日から 1 月 31 日までの 11 日間における日本語の tweet を対象とする．5 人の実験協力者によって対象の tweet を行動を表すか否かにすべて分類し，その多数決を取り行動を表すとされた tweet を正解データとした．

評価は，正解データを利用して，式 (1) に示す precision および式 (2) に示す recall，そして precision および recall の調和平均をとった式 (3) に示す F-measure で行う．

$$precision = \frac{R}{N} \quad (1)$$

$$recall = \frac{R}{C} \quad (2)$$

$$F-measure = \frac{R}{\frac{1}{2}(N+C)} \quad (3)$$

ここで，R は抽出結果のうち正解データと適合した tweet 数，N は抽出結果の tweet 数，C は正解データの tweet 数である．

実験は交差検定に基づいて，tweet データを 1 件ずつテストデータとして選択し，テストデータ以外のすべてのデータにより学習した結果を用いて，テストデータを分類することで評価を行った．

4.2 実験結果と考察

表 1 に，提案手法および従来手法 [矢野 13] による行動抽出の F-measure の値を示す．表 1 の結果より，提案手法は従来手法に比べて良い結果を示した．また各ユーザで比較すると，user6 を除くすべてのユーザにおいて，F-measure が向上したことがわかる．

また，提案手法および従来手法における，それぞれのユーザについての評価結果を図 2 に示す．図 2 における横軸は Precision，縦軸は Recall であり，それぞれのデータに付けられている数字はユーザの番号を表している．図 2 より，提案手法はすべてのユーザにおいて Precision が向上していることがわかる．また，Recall についても，平均値でみると 0.01 ほど向上しており，ほとんど変わらない結果を示した．従って，両方の評価指標で従来手法よりも高い値を示している提案手法は有効であると考えられる．そして提案手法は，従来手法と異なり行動辞書を人手で作成するといった手間がかかる作業を必要とせず，学習データとして新規の正解データを追加することで分類器が更新される．

表 1: それぞれの分類手法における F-measure

	Proposed	Previous
user1	0.754	0.657
user2	0.655	0.650
user3	0.576	0.448
user4	0.513	0.496
user5	0.534	0.419
user6	0.508	0.512
user7	0.815	0.736
user8	0.638	0.478
user9	0.794	0.683
user10	0.731	0.706
average	0.652	0.579

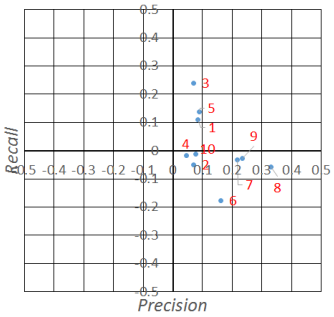

図 2: それぞれの手法による precision と recall の差

5. おわりに

本研究では，日本語の tweet から，コンテキストアウェアサービスのための行動情報を抽出する手法を提案した．そして実際のユーザが投稿した tweet から，行動を表す tweet を抽出し，評価を行った．その結果，従来手法よりも Precision，Recall の両方の評価指標で良い性能を示した．また提案手法は従来手法と比べ，辞書作成やその更新のための作業が必要なく，人間の作業時間を削減できるという利点を持つ．

今後は，抽出した行動を表す tweet に対して，具体的にどのようなことを行っているのかを表すラベルを付与する．これにより，抽出した行動をより様々なコンテキストアウェアサービスに活用することができると考えられる．

参考文献

[Banerjee 09] N. Banerjee et al.: “User Interests in Social Media Sites: An Exploration with Micro-blogs,” *Proc. CIKM*, pp. 1823-1826, 2009.

[Nguyen 12] T. Nguyen et al.: “Self-Supervised Capturing of Users’ Activities from Weblogs,” *IJIIDS*, vol. 6, No. 1, pp. 61-76, 2012.

[矢野 13] 矢野裕司, 横井健, 橋山智訓: “行動を表す単語に着目した Twitter からの行動抽出,” 第 *12* 回情報科学技術フォーラム講演論文集, vol. 4, pp. 157-164, 2013.