3ZC-08

ゲリラ豪雨パターン分類のための Twitter を用いたラベル付け自動化

藤田 拓也 † 大枝 真一 †† 中谷 剛 †††
木更津工業高等専門学校 制御・情報システム工学専攻 † 木更津工業高等専門学校 情報工学科 ††
独立行政法人 防災科学技術研究所 観測・予測研究領域 †††

1. はじめに

防災科学技術研究所 (NIED) ではゲリラ豪雨パターンを自動的に検知する試みを行っている．そのためには，ゲリラ豪雨であるというラベルが付いた XRAIN データが必要となる．XRAIN データにゲリラ豪雨かそうでないかのラベルを付ける手段として Twitter を用いている．現在，NIED では Twitter から雨に関するツイートを収集して，ゲリラ豪雨に関するツイートと XRAIN データを用いて，ゲリラ豪雨パターンを特定している．しかし，このラベル付けの作業を人間の手で行っているため時間がかかる．

そこで，本研究ではラベル付けの自動化システムを構築することを目的とする．Twitter から雨に関するツイートを収集して，それを解析することにより，ゲリラ豪雨に関するツイートとツイートされた地域を特定する．これらの情報とツイートされた時刻を用いて，その時刻の XRAIN データをゲリラ豪雨パターンとして利用する．

本研究ではナイーブベイズを用いてゲリラ豪雨に関するツイートとそうでないツイートに分類する．ナイーブベイズで文書を分類する際，文書は Bag-of-Words という単語の集合として扱われる．Bag-of-Words を作るには、文書を形態素 (言語で意味を持つ最小単位) に分割する必要がある．本研究では，日本語のツイートを形態素に分割するために MeCab を用いて文脈の解析や単語の分割を行う．

Auto labeling using Twitter for classification of localized heavy rain

†Takuya FUJITA, National Institute of Technology, Kisarazu College
††Shinichi OEDA, National Institute of Technology, Kisarazu College
†††Tsuyoshi NAKATANI, Department of Monitoring and Forecasting Research, National Research Institute for Earth Science and Disaster Resilience

2. ゲリラ豪雨

「ゲリラ」は予測困難性，局地性，激甚性などの意味を持っているが [1]，ゲリラ豪雨はまさに予測が非常に困難であり，突発的で局所的な豪雨である．気象学的に明確な定義はなく，気象庁は予報用語として「ゲリラ豪雨」を用いておらず，局地的大雨などと表現している．

ゲリラ豪雨は定義がない上に曖昧であるため，本研究ではその曖昧さを排除するために，人間がゲリラ豪雨であると考える雨をゲリラ豪雨と定義することにする．その情報を収集する手段として Twitter を用いる．

3. システムの概要

本研究で開発するシステムの概略図を図 1 に示す．

Twitter の情報からツイートの内容とツイートされた時刻を抽出して，ナイーブベイズの入力データとして与える．入力されたツイートは形態素に分割され，単語の出現確率を元にナイーブベイズにより分類される．本研究では，ゲリラ豪雨というトピックのみが必要であるため，ゲリラ豪雨とそれ以外の 2 値のトピック分類を行った．ゲリラ豪雨に関するツイートとツイートされた時刻を用いて，その時刻にゲリラ豪雨と考えられる XRAIN データの過程をゲリラ豪雨とラベリングする．

4. XRAIN

XRAIN は eXtended RAdar Information Network（高性能レーダ雨量計ネットワーク）の略である．また，国土交通省ではゲリラ豪雨などによる深刻な水害の早期警戒に役立てるために，平成 22 年より XRAIN によるレーダ雨量情報を提供している．従来の XRAIN は X バンド MP(マルチパラメータ) レーダ雨量計のみで構成されていたが，現在は X

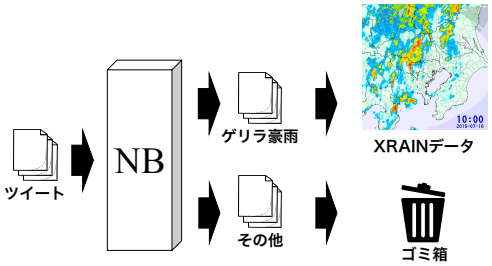

図 1　システムの概略図.

バンド MP レーダ雨量計と C バンド MP レーダ雨量計を組み合わけることにより，高精度・高分解能 (250m メッシュ)・高頻度 (配信間隔 1 分) のリアルタイムに近いレーダ雨量情報の配信を実現している [2].

5. トピック分類手法

未知文書にトピックを自動的に与える手法である．以下，ナイーブベイズによるトピック分類手法について述べる．

5.1 ナイーブベイズ

ナイーブベイズは確率に基づいた分類方法である．ナイーブベイズは以下の式 (1) のベイズの定理の右辺が最大となるクラス c を出力するという仕組みに基づいた手法である．

$$P(c|d) = \frac{P(c)P(d|c)}{P(d)} \tag{1}$$

ここで $P(c|d)$ は未知の文書 d が与えられたときにクラス c である確率を示す．また，$P(d|c)$ はクラス c が与えられたときに文書 d が生成される確率を示す．

本研究では実際の確率値は必要ではないため，式 (1) における分母 $P(d)$ はクラスに依存しないため不要である．つまり，分子 $P(c)P(d|c)$ を最大にする $C_{\max}$ を求めれば良い．$C_{\max}$ は以下の式 (2) で求められる．

$$\begin{aligned} C_{\max} &= \arg\max_c \frac{P(c)P(d|c)}{P(d)} \\ &= \arg\max_c P(c)P(d|c) \end{aligned} \tag{2}$$

5.2 Bag-of-Words

トピック分類では，単語の順序は必ずしも必要ではなく，文書中にどのような単語がどのような頻度で出現するかの情報で十分な場合が多い．文書 d を単語 w の集合として扱い，文書をベクトルで表現する手法として Bag-of-Words がある．単語の集合から特徴語を抽出して文書ベクトルを生成する．文書は Bag-of-Words で単語の集合として表し，$P(d|c)$ は以下の式 (3) で求めることができる．また，$P(w_i|c)$ は以下の式 (4) で求めることができる．

$$P(d|c) = P(w_1 \wedge \cdots \wedge w_k|c) = \prod_k P(w_i|c) \tag{3}$$

$$P(w_i|c) = \frac{T(c, w_i)}{\sum_{w' \in V} T(c, w')} \tag{4}$$

ここで $T(c, w_i)$ はクラス c に単語 w_i が出現した回数を意味している．また，V は訓練データ中の全単語集合を意味している．

6. まとめ

本研究では Twitter を用いたラベル付け自動化システムの提案を行った．また，ゲリラ豪雨の定義を行うことにより，ナイーブベイズによるクラス分類を実現した．

ゲリラ豪雨が増加傾向にある現在，ゲリラ豪雨の予測はこれからさらに重要になり，減災に役立つと言える．ビッグデータ解析分野の自然言語処理技術とセンサーデータの融合によって，減災に役立てる研究は全く新しいものである．また，ラベル付け自動化システムの実現によってゲリラ豪雨パターン検知の自動化の可能性が広がることが期待できる．

今後の課題として，有識者がツイートにラベルを付ける作業を行っているため，これからゲリラ豪雨に関するツイートを分類できるか検証を行う必要がある．

参考文献

[1]「ゲリラ豪雨」は気象学的に定義できるか？，日本気象学会, “http://dl.ndl.go.jp/view/download/digidepo_10595852_po_ART0009572832.pdf?contentNo=1&alternativeNo=”.
[2] 国土交通省, “http://www.mlit.go.jp/report/press/mizukokudo03_hh_000905.html”.