SNSコミュニティにおけるデマの状態推定モデルに関する研究

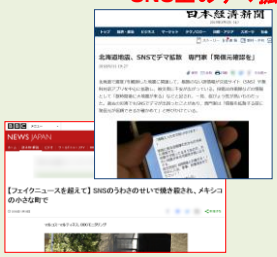
Study on models inferring diffusion status of hoaxes in SNS community

牛込 龍太郎・システム分科会・中央大学

Hoaxes like disinformation or misinformation spread in SNS (Social Networking Service) are known as a social problem because they often affect not only SNS users but also public especially when natural disasters or terrible accidents occur. The purpose of our research is to obtain the knowledge on reduce generating or spreading hoaxes through investigating when hoaxes are generated and how hoaxes become popular or unpopular. In this research, we focused on phrases in tweets and proposed a model which infers diffusion status of hoaxes combining phrases with posting timestamps.

研究背景・目的

SNS上のデマ拡散の社会問題化



[既存研究]

デマ・フェイクニュース関連
→投稿文の文字列、投稿時刻、アカウント情報から検知

SNS関連
→SNS内部の情報伝播をモデル化

[目的]

投稿文の文字列情報によってデマの状態をモデル化及びモデルの検証の実施
検証のための実データとしてTwitter使用

事前準備

[本研究における「デマ」の定義]

実際に発生した事象や存在する事物と矛盾する主張

[仮定：デマの状態と語句の関連性]

デマが「生きている」状態
||
デマが人を騙すことができている

人がデマに騙されていないとき、デマの価値は限りなく皆無。
SNSの投稿中にデマに対して疑念を抱いている表現が、人がデマに騙されている指標になり得る。

提案手法

データセット
100件程度
$D = \{(s_i, t_i)\}$
(s_i:ツイート文、t_i:投稿時刻)

$s_i \rightarrow x_i = (c_{i,1}, c_{i,2}, \ldots, c_{i,7}),\ c_{i,j} \in \{0,1\}$

ベイズ推定への適用

カテゴリ分類

カテゴリ別語句
c_i:カテゴリ
真偽 負
感情 負
行動 負
疑問
状態 負
状態 推定
人物 負

表：カテゴリと語句例

カテゴリ	例
真偽 負	架空、偽、騙され
感情 負	かわいそう、おかしい、怖い
行動 負	謝罪、悪趣味
疑問	かな、なぜ
状態 負	酷い、悪質
状態 推定	らしい、はず
人物 負	DQN

表：データセット詳細

データセット名	期間	件数	検索キーワード
F1	18.5.13-16	2149	-
N1	17.6.1-2	3776	ゲリラ豪雨
N2	18.2.7-8	4308	大阪 駅

[カテゴリ分類]

各データセットから100件程度のツイート文から、ユーザが話題に対して悪印象を持っていることを示す語句を、語句のもつ印象や性質の観点に基づき、手作業で抽出・分類

[ツイート文事前処理]

ツイートs_iがC_jの語句を含むとき、$c_{i,j} = 1$とし含まないとき$c_{i,j} = 0$とする。$c_{i,j}$の総和をx_iとし、ツイートのもつカテゴリ数とする。

$$x_i = \sum_{k=1}^{7} c_{i,j}$$

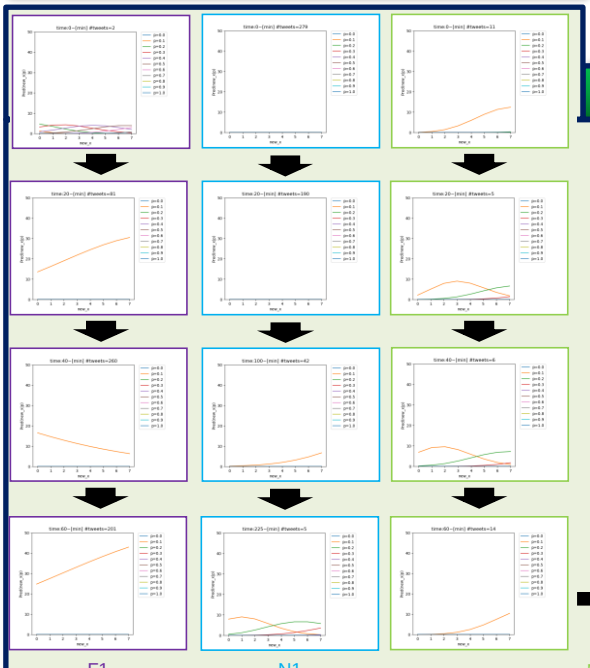
同時にツイートデータを等時間間隔に分割する。

[ベイズ推定への適用]

モデル式：$f(x_i|p) = {}_nC_{x_i} p^{x_i}(1-p)^{n-x_i}$ (p：ユーザが疑念を抱いている確率, n：カテゴリの種類数)

事前分布：ベータ分布

モデル式は二項分布をベースとしており、各ツイートの時間区分ごとにモデルの積を取った形も二項分布となる。そこに事前分布としてベータ分布を導入し予測分布が求まり、これをデマの拡散状態推定モデルとして利用する。



結果

- 区間内のツイート数が多いとき、通常の話題では大半のツイートのカテゴリ数が0になるので予測分布の確率密度の値は全てのpにおいて0となる一方、デマの話題ではカテゴリ数が0より大きいものが存在するため一部のpの値で確率密度の値に変化が出ており、最大値が通常の話題のそれよりも大きくなった。
- デマの話題における確率密度の値はデマであると判明した時点ではその最大値が右方向へ移動する(ユーザが疑念を抱く印象の語句を多く使用する)傾向にあると考えられる。なお、今回検証に使用したデマの話題は2つの偽情報を含むため、一方の嘘に気づいた後もう一方の嘘に気づく、といったユーザが存在したため、一度最大値が左方向へ移動しその後再度右方向へ移動している。
- 区間内のツイート数が少ないとき、各ツイートのカテゴリ数の変化が予測分布の確率密度の値に影響しやすくなるため、値の変化が激しい。デマの話題と通常の話題との間で差異は見出せなかった。

[今後の課題]

- 区間内のツイート数が多い場合の、デマの話題と通常の話題のデータセットから生成された予測分布の形状の差異を利用してデマの検知を行う
- 語句のカテゴリ分類を機械的に行うなどして客観性を高める
- ノイズを含むデータへの対応