A-006

ベイジアンネットワークによる SNS 上での流言の分類
Classification of hoaxes on Social Network Service using Bayesian Networks

牛込 龍太郎† 松田 健‡ 園田 道夫† 趙 晋輝†
Ryutaro Ushigome Takeshi Matsuda Michio Sonoda Jinhui Chao

1. はじめに

近年 SNS(Social Network Service) はユーザ間のコミュニケーションツールとして広く利用され，情報発信の手軽さから速報性に優れる．しかし，情報の信憑性を保証されていないことが多く，デマや誤りを含む情報が拡散される可能性がある．特に大規模な災害や事件・事故に関連する誤った情報は，救助活動や支援活動の妨げとなる懸念がある．したがって，投稿内容の信憑性の判断に対する機械的な支援の可能性を考えることは重要であり，そのためには大量のデータから信憑性の程度を示す特徴を見出す必要がある．その手段として本研究では機械学習の 1 つであるベイジアンネットワークを用いることで，投稿内の語句の関係性を見出し，投稿内容をデマか正常な文章かについての判定の可能性について検討する．本研究では SNS の 1 つである Twitter の投稿「ツイート」を利用した．

2. 既存研究

Twitter におけるデマの検知を機械的に支援する手法としては，ツイートに対するユーザの反応を n-gram を用いてテキストから分析し、ユーザが情報の真偽を判断する補助となるシステムが提案されている [4]．文献[4]では，ツイートをユーザの反応と根拠の有無の観点でそれぞれ SVM を用いて分類し，ユーザの反応の分類について精度の向上が実現されている．本研究での提案手法では同じ分野の複数の話題にも適用できることを念頭に置いている．

3. ベイジアンネットワーク

ベイジアンネットワークは確率分布を図式的に表現した確率的グラフィカルモデルの 1 種である．確率的グラフィカルモデルではグラフの各ノードが確率変数，リンクが変数間の確率的関係を表現する．そして，モデルが含む全ての確率変数上の同時分布が，一部の変数のみの積としてどのように分解可能か，ということをグラフから読み取ることができる．グラフィカルモデルの構造を調べることで条件付き独立性などのモデルに関わる知見が得られる．ベイジアンネットワークの特徴として，リンクが特定の方向性をもつため，確率変数間の因果関係を表現するのに適していることが挙げられる．

例として，3 つの確率変数 A，B，C 上の任意の同時分布 $\mathrm{p}(A, B, C)$ を考える．確率の乗法定理から

$$\begin{aligned}\mathrm{p}(A,B,C) &= \mathrm{p}(C|A,B)\mathrm{p}(A,B)\\ &= \mathrm{p}(C|A,B)\mathrm{p}(B|A)p(A)\end{aligned}$$

と書ける．これをグラフ化したものが図 1 である．

確率変数の集合を $X = \{x_1, \ldots, x_n\}(n \geq 1)$，$\mathrm{pa}(x)$を x の親ノードの全ての確率変数とすると，ベイジアンネットワークが表す確率分布は一般に次のように書ける．

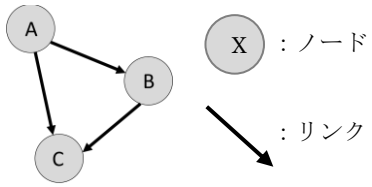

図 1 ベイジアンネットワークの例

$$\mathrm{P}(X) = \prod_{x \in X} \mathrm{p}(x|pa(x))$$

本研究では，文中での単語・語句の出現順序に依存せずに語句の共起度をみるため，語句の条件付き確率を算出しベイジアンネットワークを作成した．

4. ベイジアンネットワークを使用した実験

4.1 データ

表 1 データ詳細

データ名	件数	内容
Hoax1	118	市街地にて大雪が降っていることを画像付きで伝える内容．実際には降雪はなく，添付された画像は前年のものであった．
Hoax2	196	日本国内にある山が噴火していることを画像付きで伝える内容．実際に噴火はなく，添付画像は国外の山のものであった．
Normal	299	初夏の時期に雹が降っていることを伝える内容．

本研究で使用したデータの詳細を表 1 に示す．3 種のデータのうち Hoax1 と Hoax2 はデマ，Normal はデマではない通常の内容である．データの収集には Twitter の Web ページでの検索機能とキュレーションサイト Togetter，Python のライブラリ tweepy を使用した．各データはデマもしくは通常ツイートをリツイートしたものである．

また本実験では日本語評価極性辞書[2][3]が含む表現の一部と，前述のツイートデータの頻度解析から辞書データを作成した．日本語評価極性辞書は東山らによって作成されたもので，名詞約 8500 語，慣用句を含む用言約 5000 語に対して，その語句が持つ印象を positive，negative，それらのどちらでもない（even）の 3 種に分けたものである．本実験で使用する辞書データの具体的な作成手順は，まず極性辞書からデマの発信元への罵倒や，デマの投稿に対して非難をするような表現を選出し，またツイートデータの頻度解析から，極性辞書に含まれない表現やデマツイートを略した「デマツイ」のような略語表現を追加する．これに加え，「大雪」や「大雨」といった自然災害に関する表現，「火災」や「殺人」といった事件・事故の話題に用いられ得る表現を個別に追加しそれぞれ辞書”disaster”，辞

† 中央大学
‡ 長崎県立大学

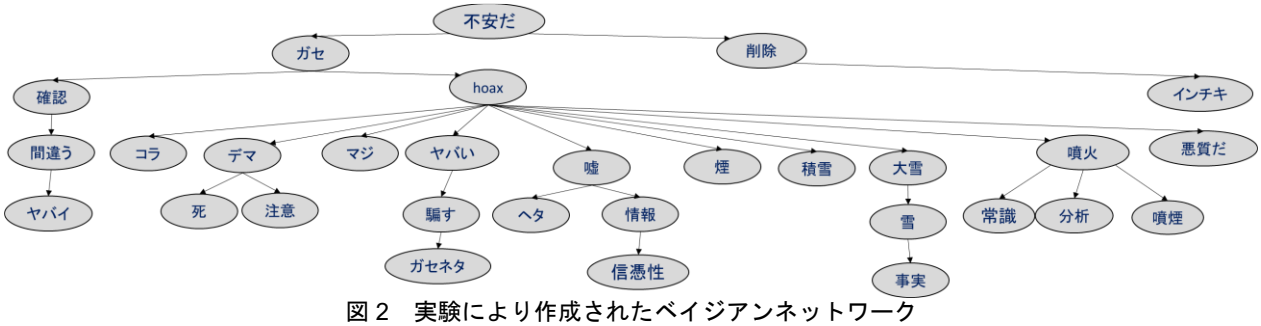

図 2　実験により作成されたベイジアンネットワーク

書”accident”とする．その結果各辞書に含む語句は 262 語になった．本研究ではデータの話題から”disaster”のみを使用した．

4.2 実験方法

3 章の表 1 で示したツイート合計 613 件をそれぞれ形態素解析し，各形態素が辞書”disaster”に含まれるかをツイート毎にまとめる．また，ツイート内容がデマであるか通常のツイートであるかを 1/0 値で表し属性として追加する．

機械学習ソフトウェア Weka を使用してベイジアンネットワークモデルを構築し，交差検証を用いてモデルの精度を算出させる．交差検証の分割数は 5 とした．

5. 結果・考察

表 2　実験結果

hoax	TP	FP	Precision	Recall	F-measure
1	0.997	0.003	0.997	0.997	0.997
0	0.997	0.003	0.997	0.997	0.997

実験の結果，分類の精度については表 2 のようになり，図 2 のような形状のベイジアンネットワークのグラフが得られた．なお，リンクと接続していないノードは省略している．表 3 にはデマである場合(hoax=1)とデマでない場合(hoax=0)においての，語句の出現確率とその倍率を示す．表 2 から比較的高い割合でツイートが分類されていることが確認できる一方，図 2 のグラフからは，「不安だ」ノードがグラフの根になるなど 7 つの語句が「hoax」ノードの子以下に収まっていないことが確認できる．ベイジアンネットワークの定義に基づくと，これら 7 つの語句の出現確率は「hoax」の確率，すなわち話題がデマであるかどうかが影響しないことになる．しかし表 3 をみると，「hoax」ノードの子以下に位置していない語句の倍率が，「hoax」の確率の操作に伴い変化していることが確認できる．このことから，「hoax」ノードを根にもち，デマと通常について分類できるグラフモデルを構築できる可能性は低くないものの，辞書が持つ語句については再考の余地があると考えられる．

6. 結び

本研究では同じ分野の複数の話題に関するデマとデマでない通常のツイートを入力としてベイジアンネットワークを構成し，その考察を行った．今後の課題としては，図 2 のグラフにおいて意味の観点で関連性のある語句が集中している箇所が複数見られたことから，辞書語句のグルーピングの可能性についての検討を行うほか，極性辞書”accident”を使用する実験の実施，ツイートデータを話題の観点からグルーピングする操作方法の検討，話題に応じて辞書に含ませる単語の選出方法の検討などが考えられる．また，得られたネットワークに新たなテストデータを入力することでの分類精度の変化についても検証の必要がある．

参考文献

[1] C. M. Bishop, “パターン認識と機械学習 下”, p.p. 71-84 Vol.2
[2] “日本語評価極性辞書（名詞編）”, 東山昌彦, 乾健太郎, 松本裕治, “述語の選択選好性に着目した獲得”, 言語処理学会, 2008.
[3] “日本語評価極性辞書（用言編）”, 小林のぞみ, 乾健太郎, 松本裕治, 立石健二, 福島俊一, “意見抽出のための評価表現の収集”, 自然言語処理, Vol.12, No.3, pp.203-222, 2005.
[4] 藤川 智英, 鍛治 伸裕, 吉永 直樹, 喜連川 優, “マイクロブログ上の話題抽出とユーザの態度の分類に基づく流言検出支援システム”, DEIM Forum, 2012

表 3　一部語句の場合別出現確率

Word	hoax=1	hoax=0	倍率
ヤバい	0.6206	0.0016	387.875
煙	0.6206	0.0016	387.875
大雪	0.373	0.0016	233.125
嘘	0.0619	0.0016	38.6875
マジ	0.0492	0.0016	30.75
デマ	0.0365	0.0016	22.8125
噴火	0.0269	0.0016	16.8125
ネタ	0.0174	0.0016	10.875
悪質だ	0.0174	0.0016	10.875
騙す	0.0115	0.0012	9.5833
ガセ	0.0106	0.0015	7.0667
死	0.0053	0.001	5.3
注意	0.0053	0.001	5.3
雪	0.0312	0.0071	4.3944
確認	0.0047	0.0014	3.3571
事実	0.0047	0.0017	2.7647
ガセネタ	0.0083	0.0045	1.8444
不安だ	0.0071	0.0041	1.7317
間違う	0.0076	0.0051	1.4902
インチキ	0.0033	0.0028	1.1786
ヤバイ	0.0069	0.0059	1.1695
情報	0.007	0.026	0.2692
信憑性	0.0043	0.021	0.2048
積雪	0.0015	0.6283	0.0024