Twitter上の特定トピックにおける話題ネットワークの可視化

飯尾 直樹[1,a)]　寺田 実[1,b)]

概要：SNS 等の双方向メディアでは意見の主張は活発に行われているが, 意見を網羅的に収集することは容易ではなく, 一方的な主張や主観的な意見によって事実を正確に認識出来ない可能性がある. この社会的問題を解決するために, 本研究では日本での利用率が高く Twitter での特定の話題に対するツイートを広範囲に対象として, トピックモデルを用いて主張ごとに分類し, 直感的に理解しやすい形での可視化を行うシステムを提案する.

1. はじめに

近年のスマートフォンなどの電子デバイスの普及によって, SNS 等のソーシャルメディアの利用率は非常に高くなってきている. 実際に日本人の SNS 利用率は 60%であり[*1], 半分以上の日本人が SNS を利用している.

SNS 上ではユーザが容易に発言できることから, 様々な社会の話題に対しての大量の意見を見ることができる. しかし,SNS 上での話題の移り変わりは激しく, 話題自体や背景となる情報の収集や理解が難しい場合がある. またユーザによっては真実とは異なる内容の投稿を行ったり, プロパガンダ的に偏った内容で投稿されている場合もある. これらの課題を踏まえて SNS 上の話題や意見を正しく把握するために, 話題に対して網羅的に情報や意見を収集し, 複数の視点で意見を理解する必要があると考えられる. 総務省により定義されているインターネット・リテラシーの観点においても,“情報を収集する能力” と “ICT メディアにおける送り手の意図を批判的に読み解く能力” が定義されている[*2]ことから, SNS を適切に扱うための必要なリテラシーであると言える.

2. 関連研究

SNS 上の話題の抽出や可視化を行う研究は近年, 活発に行われている. 星らの研究 [1] では Twitter のタイムラインから特定の話題を抽出し, 同一の話題のツイートを連結することでタイムライン上の多様な情報の理解を支援する方法を提案した. 小川らの研究 [2] では話題の理解を示すために, ツイートに対するリプライやリツイートなどの関連性を持つツイートを取得し, 円状木構造グラフで示すことでツイート同士の関連性を可視化する手法を提案した. 榊らの研究 [3] では Twitter 上の大規模な情報拡散を直感的に理解するための手法として, 階層構造とユーザ同士の社会ネットワーク構造を利用することで, Twitter の情報拡散の関係性を示している

3. 提案システム

本研究では, SNS として日本での利用率が高い Twitter を対象として, 話題に対するユーザの主張を内容ごとに分類することで, 話題に対する客観的理解を支援する手法を提案する.

3.1 概要

システムのフローは以下の様に実行される.

(1) Twitter 上の特定の話題に対する意見・反応を収集
(2) 収集したツイートをトピックモデルを利用して分類
(3) 分類されたツイート集合から特徴語を抽出
(4) 特徴語でネットワークを構成し, 可視化を行う

3.2 話題

Twitter 上で特定の話題に対して反応しているユーザが一定数存在する状況を想定する. 本研究における話題とは話題自体を表す単語集合であり, 話題に対する反応はその単語集合に紐づく単語集合, つまりは関連性が高い単語集

[1] 電気通信大学大学院情報理工学研究科
[a)] i18310101@edu.cc.uec.ac.jp
[b)] terada.minoru@uec.ac.jp
[*1] 総務省 平成 30 年版 情報通信白書 ソーシャルメディアの利用状況
[*2] 総務省 情報通信政策研究所 平成 24 年 青少年のインターネット・リテラシー指標

合であるとする.

単語検索を行う単語自体を話題の中心語とする.

3.3 コーパスの収集

本研究では呟いているユーザを限定せず網羅的にツイートを収集する必要があるため, Twitter Developers のツイート検索 API を利用してツイートの収集を行う. また, Twitter には情報拡散の手段としてリツイート機能が存在する. リツイートはツイートの共有機能であり, 一般的にツイートに賛同する場合に用いられるため, リツイートを行ったユーザはリツイート内容と同じ内容を投稿したとしてコーパスを定めた.

3.4 分類

収集したツイート内容に対して, 意見の概要ごとに分類を行う. 分類には Python のライブラリである gensim を用いて, LDA によるトピックモデルで分類を行う. ツイートの文章は句読点や改行などが不規則であり, LDA で分類を行う単位を一文章とすると極端に短い文章となってしまう場合があった. そのため本研究では, 一つのツイートの文章量は最大で 140 文字であることから, 一つのツイートを一文として LDA の分類を行った.

3.5 潜在的ディレクレ配分法

潜在的ディレクレ配分法 (Latent Dirichet Allocation) [4] はトピックモデルの手法の一つとして, 文書に対しては潜在的な複数のトピックが存在しているとし, その上で文書生成を行う過程のモデルである.

3.6 特徴語抽出

LDA を用いてツイートを分類したのち, その分類結果から特徴語を抽出するには, word2vec で得られる単語同士の類似度を用いる. word2vec [5] はニューラルネットワークを用いた分散表現を取得するための一手法である. word2vec では skip-gram モデルを利用して抽出を行った. skip-gram モデルはある単語を入力層とし, その単語の周辺に出現する単語を出力層として周辺単語の予測を学習するモデルである. 類似度の高い単語は似た文脈に出現しやすいという分布を元に, その際に計算される隠れ層は単語のベクトルを表現するとする手法である. この単語のベクトルを用いることで限定的な話題における単語の意味を考慮することが可能であり, 単語同士の意味的な類似度を計算することが可能となる. 隠れ層の計算は Python の gensim ライブラリを用いて学習を行った.

3.7 既存研究

LDA と word2vec を併用して用いる既存研究は存在しており, Zhibo Wang らの研究 [6] では LDA と word2vec を利用した特徴語抽出方法を提案している. Zhibo Wang らの研究では意味的に距離があるドキュメントから特徴を抽出できることを示している. 本研究では LDA と word2vec を併用する有効性から, 可視化を行う際にトピックモデルで分類されたトピック同士を意味的類似度で再度結合することで過剰なトピック分類を適切な範囲に抑えることができると考えた.

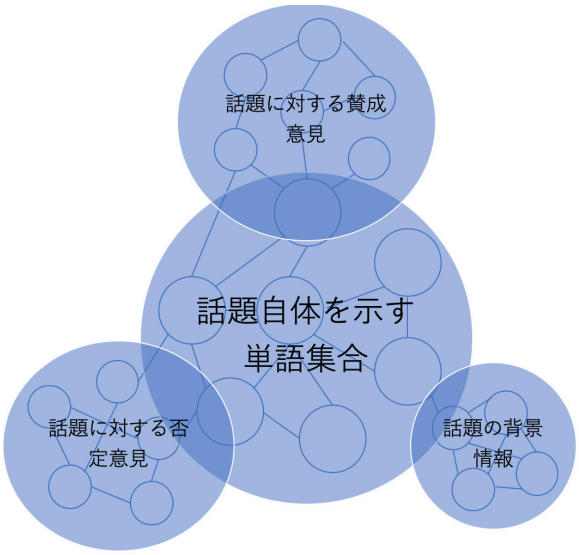

図 1 可視化のイメージ

3.8 可視化

抽出された類似度を元に可視化を行う. 話題は話題自体を表す単語集合であり, これらの単語集合は話題の中心語と単語としての類似度は高くなり, かつこれらの単語同士の類似度も高くなる. また話題に対する反応においても単語集合に対して紐づけられる単語集合であることから, 話題自体を表している単語集合と類似度が高い単語は話題に対する一部の反応と考えられる. つまりは話題の中心語と距離が近いノードは客観的な話題自体を表しており, 距離が遠いノードは話題に対する反応を表していると考えられる. 図 1 にこれらを踏まえた可視化のイメージを示す. 可視化のアルゴリズムは図 2 で示す. 話題の中心語を中心に中心語と類似度が高い特徴語をノードとして配置する. 配置されたノードから再帰的に類似度が高い特徴語をノードとして配置する. 再帰は一定の閾値以上の類似度が高い単語が見つからなくなるまで実行され, 単語同士の類似度によるネットワークが構成される. 可視化の実装には Python の networkx と PyGraphviz のライブラリを用いた.

4. 分析

実際に Twitter 上で多くの反応が寄せられた話題に対して, 提案システムを用いて可視化を行った.

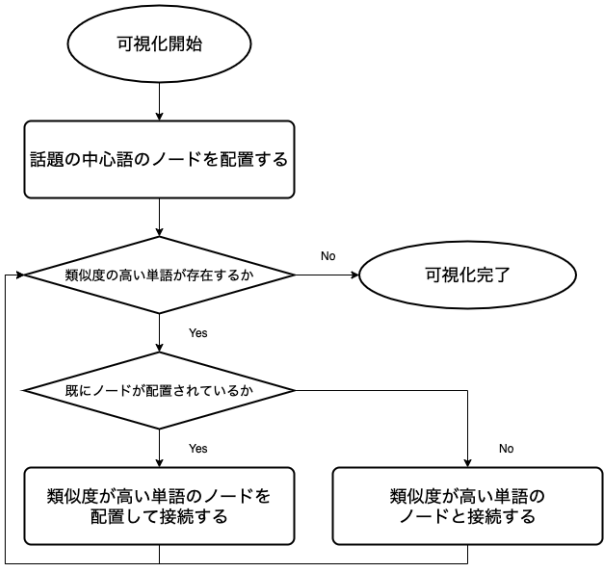

図 2 可視化のアルゴリズム

4.1 事例

検証事例として, 今回は 2019/12 時点で Twitter 上やニュースメディア等で話題となっていた “桜を見る会” を話題の対象とした. 収集を行った時期とツイート数は以下である.

- 収集時期: 2019/11/26〜2019/12/3
- 検索単語: 桜を見る会
- ツイート数: 92992 件

LDA によって分類するトピック数は 5, 10, 15 で可視化を行った. 図 3, 4, 5 が今回の事例に対するそれぞれの可視化の結果である.

4.2 考察

結果の有効性は以下の項目で判断を行う.

(1) 話題自体を説明する特徴語が中心語に近い位置で可視化されている
(2) 事例における意味が近い特徴語が接続されている
(3) 主張が分類され, 特徴語として可視化されている

今回の “桜を見る会” の事例から話題自体を説明する特徴語を考えるために, ニュースメディア等の報道から事例の概要を要約すると “安倍首相が特定の人物に招待を送っていた問題”, “参加者の名簿を破棄していた問題”,“反社会勢力との関わり” などが問題として取り扱われていた. 本研究の結果では図 4 の結果が上記の要約を適切に表していると見ることができる.

“桜を見る会” から直接接続されているノードを見ると “反社会勢力”, “安倍”, “招待”, “参加” などが存在しており, これらは上記の要約の特徴語であると認識できる. 特徴語同士の関係においても “野党”, “追求”, “議員”, “国会” の組み合わせなどは意味的に近い関係にあると考えられる. 他にも “官房長官”, “反社会勢力”, “定義” の組み合わせも

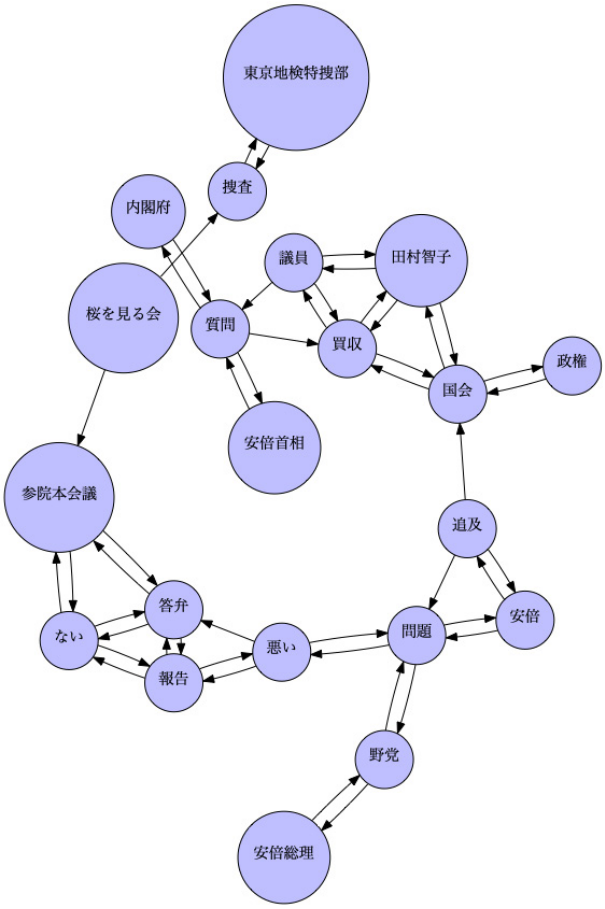

図 3 可視化結果:モデル数=5

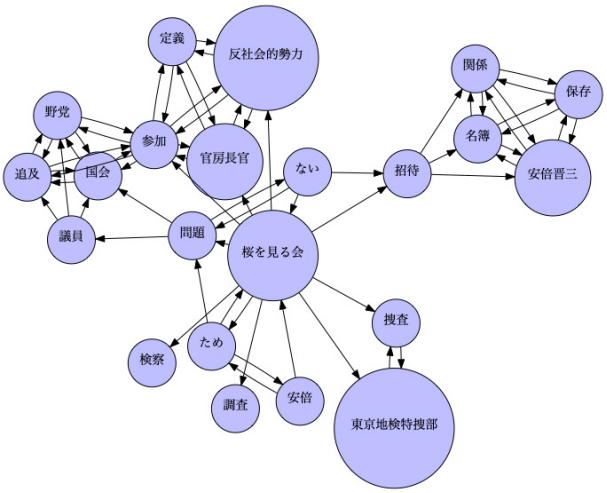

図 4 可視化結果:モデル数=10

官房長官の発言として意味的に近いと考えられる.

(3) については図 3, 4, 5 ともにユーザによる主張は特徴語として十分な可視化は出来ていない.

5. おわりに

本研究では Twitter を対象としてツイートを意味的に分類し, 意味的な距離を含む可視化を行うことで, 話題に対する客観的理解を支援を目標とする手法を提案した. 結果と

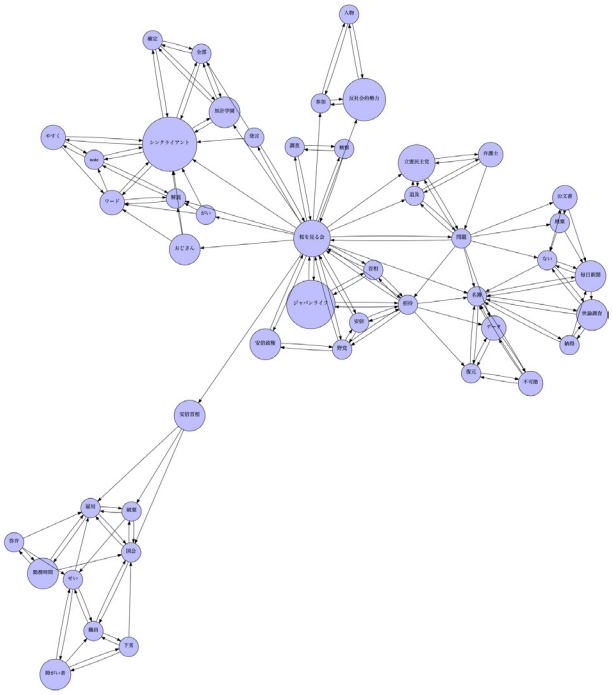

図 5　可視化結果:モデル数=15

しては話題自体を意味する一部特徴語の抽出, 意味的に近い特徴語を繋げることによる可視化を行うことには成功したと考えられる. しかしユーザによる主張は現段階では適切に抽出できておらず, 抽出しているトピックも過不足なく抽出できているわけではなく, これらを解決するために閾値の調整や品詞の適切化, モデリングの調整を行う必要がある. これらの調整を行うとノード数などが過剰になり, 図 5 の様に可視化結果の可読性を損なう可能性がある. 今後の課題としては可読性を損なわない範囲での情報量と関係性を増加させることがあげられる.

参考文献

[1] 星　皓介，山田鋼一，絹川博之：Twitter タイムラインからの話題の抽出とその評価，情報処理学会第 76 回全国大会 (2014).
[2] 小川貢平，芝田圭佑，藤井友紀子，濱川　礼：Twitter におけるツイートの関連性可視化システム (2014).
[3] 榊　剛史，鳥海不二夫，大知正直：ソーシャルメディア上の大規模情報拡散に関する俯瞰的可視化手法の提案，人工知能学会全国大会 (2019).
[4] Blei, D. M., Ng, A. Y. and Jordan, M. I.: *Latent Dirichlet Allocation*, Journal of Machine Learning Research (2003).
[5] Mikolov, T., Chen, K., Corrado, G. and Dean, J.: *Efficient estimation of word representations in vector space*, In Proceedings of Workshop at ICLR (2013).
[6] Wang, Z., Ma, L. and Zhang, Y.: *A Hybrid Document Feature Extraction Method Using Latent Dirichlet Allocation and Word2Vec* (2016).