

固有表現および単語の出現率に基づく関連ツイートの検索

向井 勇希[†] 杉本 徹[‡]芝浦工業大学大学院 理工学研究科[†] 芝浦工業大学 工学部[‡]

1. 研究背景と目的

現代の SNS はユーザ同士の会話の場となるだけでなく、最新の情報を収集するツールとしても社会に欠かせない存在となっている。しかし、SNS 上の情報はユーザが容易に発信できるために信憑性に欠けるものも多く、誤った情報が拡散されることで社会に混乱をもたらす危険性も大きい[1]。

この問題に対処するために、Twitter 上のツイートを入力として受け取り、関連ツイートを肯定や否定などのカテゴリに分けて表示することで、入力の内容に対して他のユーザが持つ印象を可視化するシステムの構築を目指している。本研究では入力ツイートから固有表現や各単語の出現率の時間的変化などの情報を用いてキーワードを複数抽出し、それらを用いて関連ツイートを収集する手法を提案する。

2. 研究の概要

先に述べたように、最終的にはある入力ツイートに関連するツイートを内容のカテゴリに分けて表示し、関連性を可視化することで内容の考察を補助するシステムの構築を目指している。図 1 にシステムの出力のイメージを示す。

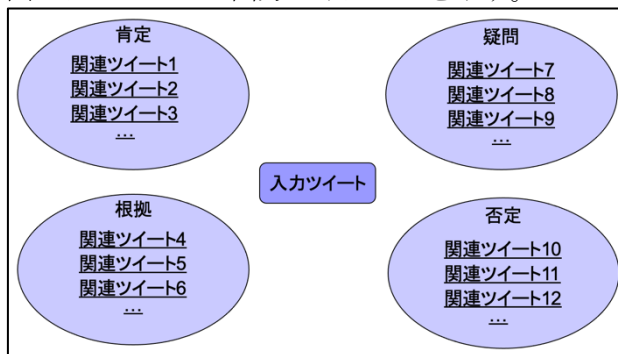


図1 システムの出力のイメージ

本研究では入力ツイートに含まれる固有表現や各単語の出現率の変化などの情報を用いて、ユーザが入力ツイートの内容を考察する上で有用な関連ツイートを効率的に検索することを目指す。

Related Tweet Search Based on Named Entity and Word Occurrence

[†]Yuki Mukai [‡]Toru Sugimoto

[†]Graduate School of Engineering and Science, Shibaura Institute of Technology

[‡]Faculty of Engineering, Shibaura Institute of Technology

3. キーワードの抽出

本研究では入力ツイートに含まれる単語からキーワードを抽出し、それらを用いて AND 検索を行うことで関連ツイートを収集する。入力ツイートを形態素解析し、その中に含まれる名詞に対して以下の条件に応じたスコアを付与する。

①ハッシュタグ

ユーザ自身がツイートに付与するハッシュタグはツイートの話題を直接表現するものであることが多く、関連ツイートを検索するキーワードとして有用であると考えられる。各名詞に対して以下のスコア S_{hash} を付与する。

$$S_{hash} = \begin{cases} 40 & (\text{ハッシュタグ内の名詞}) \\ 0 & (\text{それ以外の名詞}) \end{cases}$$

②固有表現

予備調査として、無作為に収集したツイート 100 件について人手でキーワードを抽出した。その結果、含まれている固有表現の 82% がキーワードとして抽出されていた。また、それらの固有表現は全て固有名詞であった。以上より、各名詞に対して以下のスコア S_{ner} を付与する。

$$S_{ner} = \begin{cases} 40 & (\text{固有名詞}) \\ 0 & (\text{それ以外の名詞}) \end{cases}$$

③単語の出現率変化

Twitter の持つ特徴の一つとして、情報拡散のリアルタイム性が挙げられる。それゆえに、短期間に出現頻度が上昇している単語は盛んに取り上げられている話題を表していると言える[2]。こうした傾向が見られる単語をキーワードとして抽出する。

出現率の変化は以下のように測定する。まず、対象単語を含むツイートを抽出元のツイートが投稿された時刻から遡って 500 件検索する。もっとも古いツイートが投稿されてから抽出元のツイートが投稿されるまでの時間差を $time_f$ として記録する。次に、遡って検索する起点を抽出元ツイート投稿の 1 ヶ月前に変更し、同様に投稿時刻の差を記録する。これを 1 年前のツイートに到達するまで繰り返し、最も大きな投稿時刻の差、すなわち出現率が最も低かった時の時間差を $time_r$ として記録する。

それぞれの単語に対して、こうして求めた 2 つの時間差の比をスコア S_{time} として以下のように付与する。

$$S_{time} = \frac{time_r}{time_f}$$

④スコアの付与とキーワードの決定

各単語をキーワードとして抽出するか決定するスコアを以下の式で定義する。

$$score = S_{hash} + S_{ner} + S_{time}$$

$score$ が 50 以上の単語をその値が大きい順に 3 つまでキーワードとして抽出する。 $score$ が 50 を超える単語が存在しない場合は最も $score$ が大きい単語を 1 つキーワードとして抽出を行う。こうして抽出した単語で AND 検索を行うことで関連ツイートの検索を行う。

4. 評価実験

4.1 実験方法

提案手法で入力ツイートの関連ツイートの検索を行い、その結果を評価した。検索で得られたツイートは入力ツイートとの関連の種類別に分類して表示することを想定している。よって、検索で集まったツイートが元のツイートの内容に十分に関連した投稿であり、かつその関連性に何かしらのカテゴリを割り当てられるものであることが望ましい。

まず元ツイートとそこから抽出したキーワードで検索した関連ツイート 10 件の対を 10 個用意した。実験協力者 10 名に依頼し、それらの内容の関連度を 5 段階で評価し、5 段階評価の 3 以上のツイートに関しては事前に用意した選択肢の中から「肯定」、「否定」、「補足」、「疑問」、「発展」、「言及」の 6 種類の関連カテゴリを複数選択が可能な形で付与してもらった。なお、元ツイートには内容の考察の必要性が高いと想定される時事問題や思想論に関するツイートを中心に使用し、実験協力者が選択肢の中に適切な関連カテゴリが存在しないと判断した場合は自由記述で関連の種類を書いてもらった。

4.2 実験の結果

図 2 に関連度の 5 段階評価の結果を示す。

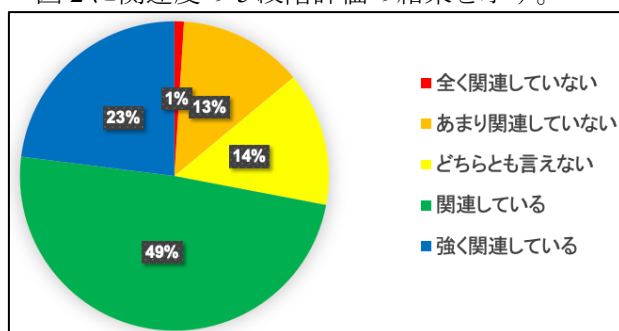


図 2 関連度評価の結果

収集した関連ツイートの 72%が元のツイートとの関連度が 4 以上であり、おおむね元のツイートと関連があるものを集められている。また、表 1 にそれぞれのカテゴリが付与された割合を示す。

表 1 各カテゴリが付与された割合

肯定	否定	補足	疑問	発展	言及
14.8	24.4	21.2	4.5	16.3	57.1

表 2 に具体的な関連ツイートと付与されたカテゴリの例を示す。ツイートの文章は一部省略している。

表 2 関連ツイートと付与されたタグの例

元	ヒグマに詳しい人「むしろ道南の方が知床とかより危険」	
関連	道南のヒグマはガチ勢だからホントに怖い。知床とは訳が違う	肯定
関連	「道南のヒグマは危険」っていうか、「知床のヒグマは割とおとなしい方」と聞いた。	疑問

5. 考察

関連度が低いと判定されたツイートの多くはキーワードに固有名詞を含まないものであった。一般名詞のキーワードに関しては元のツイートで主語になっていたものが関連ツイートで目的語として用いられるなど文中での位置付けが大きく異なる場合が多く、格関係の一致を考慮するなどスコアの付与方法の改良が必要である。

6. まとめ

ツイートから固有表現や単語の出現率といった情報を元に抽出したキーワードを用いて関連ツイートの検索を行い、実験の結果収集したツイートの 7 割以上が元のツイートと関連がある内容であることを確認した。一方で直接話題を表す単語が含まれない関連ツイートは収集できていないため、他の手法を用いてこれらのツイートに対応することが今後の課題である。また、最終的には元のツイートとの関連性を可視化することを目指しているため、収集したツイートを分類、あるいはクラスタリングする手法も今後検討していく予定である。

参考文献

- [1] 鳥海 不二夫, 篠田 孝祐, 兼山 元太, “ソーシャルメディアを用いたデマ判定システムの判定精度評価”, 情報処理学会デジタルプラクティス, Vol.3, No.3, pp.201-208 (2012).
- [2] 向井 勇希, 杉本 徹, “出現頻度と周辺文脈に基づくツイート中の新語の検出”, FIT2021 講演論文集 (2021).