2L-02

話題語に関連する情報のマイクロブログからの抽出

鈴木　諒雅†　　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

近年，ソーシャルメディアのユーザが増加し続けており[1]，その膨大なユーザにより個人の発言が多く投稿され，マイクロブログには情報源としての役割が期待されている．マイクロブログのひとつである Twitter は，特に話題のトレンドに関しての投稿が多く，リアルタイム性が高いという特徴を持つ．しかし，単純に話題の語(以下話題語と呼ぶ)で検索しただけでは，スパムや bot などの話題語は含むが内容を含まない投稿や，話題語と内容は含むがユーザにとって役に立たない投稿が大量に出力される．そのため，意見・知識・経験といったユーザにとって役に立つ関連情報を適切かつ効率的に収集することは容易でない．また，話題語に関する内容が多岐にわたるためにユーザ自身がどう検索してよいかわからないという問題も存在する．マイクロブログにおける情報取得を効率化することを目的とした研究として，ブロガー同士のリンク構造に着目した岩木らの研究[2]などが挙げられる．

本研究では，話題語は含むが内容を含まない投稿に対してはフィルタリングによって除去する手法を提案する．また，話題語と内容は含むがユーザにとって役に立たない投稿に対しては，話題語に対する関連語を提示して選択させるユーザとシステムのインタラクションによって，ユーザが求める情報の方向性を決定し，ユーザにとって役立つ情報を抽出，提示する手法を提案する．

２．提案手法の概要

提案手法の流れを図 1 に示す．はじめに Twitter 検索 API を用いてユーザが入力した話題語を含むツイートを 1000 件収集する．この時点で，検索オプションを用いて URL を含むツイートを除外する．URL を含むツイートでは投稿者本人による意見や経験談などが含まれないツイートが多いためである．また，リツイートとリプライも除去する．そして MeCab[3]を用いて形態素解析を行い，ツイートのベクトル化を行う．ベクトル化には名詞，動詞，形容詞のみを用いる．

その後のフィルタリングで不要なツイートを除去し，その後話題語に対する関連語を抽出する．関連語抽出の前にフィルタを適用するのは，スパムツイートなどが関連語抽出の結果に悪影響を及ぼすためである．そして抽出された関連語リストをユーザに提示し，ユーザは興味のある関連語を選択する．関連語リストを見ることで，ユーザは話題語に関するどのような関連情報が存在するのかを大まかに把握することができる．その後，検索クエリを「話題語　関連語」として検索した結果に，再びフィルタを適用した結果得られたツイートリストをユーザに提示する．

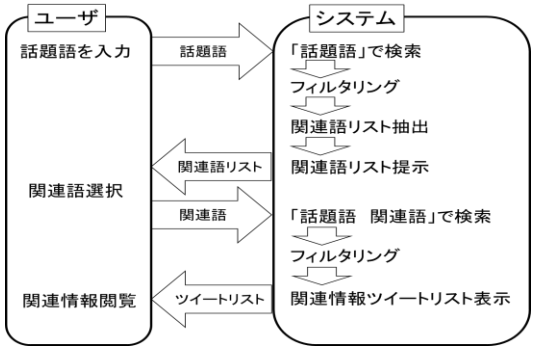

図 1 処理の流れ

３．フィルタリング

フィルタリング部では，主にスパムツイートを除去する．話題語によっては検索結果の半分以上がスパムツイートになることもあり，スパムツイートの除去は不可欠である．スパムツイートにはいくつか種類があり，現在のトレンドワードを羅列するものや，関係の無い宣伝をするものなどがある．

３．１ 類似度フィルタ

類似度が高いツイートの組を除去するフィル

Extraction of information related to a topic word from microblogs
†Ryoga Suzuki　Graduate School of Engineering and Science, Shibaura Institute of Technology
‡Toru Sugimoto　College of Engineering, Shibaura Institute of Technology

タである．ほとんどのスパムツイートは，ほぼ同一の内容が複数回に渡って投稿されるため，類似度が高いツイートが複数出現するという特徴を持つ．そこで，各ツイートに対して，他の全ツイートとの類似度をそれぞれ算出し，しきい値を超えたツイートが一定数以上存在する場合に除去する．

類似度の尺度としてユークリッド距離の逆数，コサイン類似度，一致単語数/全単語数の 3 つの中から，最もスパムを除去することができた一致単語数/全単語数を用いることにする．

３．２ 動詞比率フィルタ

ツイート中に出現する動詞の割合が少ないツイートを除去するフィルタである．これは，文章ではなく単語の羅列であるスパムでは動詞や助詞が少なくなる傾向があることを利用している．

３．３ 単語数フィルタ

単語数が 2 以下のツイートは内容が極端に少ないと判断して除去する．

３．４ 除去しきれないスパム

単語の羅列ではなく，出現数が少ないというスパムツイートは上記のフィルタでは除去することができないが，出現数が少なければこの後の関連語抽出に悪影響を及ぼさないため，ここでは考慮しない．

４．関連語の抽出

関連語抽出部では，話題語に対する適切な関連語リストを検索結果から抽出する．また，ここでは関連語候補として名詞のみを考える．

４．１ 関連語抽出の概要

話題語でツイートを検索した結果に出現したすべての名詞に対して関連語としての適切さを数値化し(以下スコアと呼ぶ)，それに基づいてソーティングした上で上位の単語を関連語として抽出する．そのため，スコアの算出方法として何を選択するかによって，抽出結果が変わる．

４．２ tf による抽出

話題語でツイートを検索した結果における単語の出現数をスコアとする方法である．話題語と共起する頻度であるため，基本的には関連語として適切な単語のスコアが高くなる．

４．３ tf に重みを掛けた抽出

tf による抽出では，例えば「人」などといった一般的な単語の値が大きくなるという問題点が存在する．その問題に対して，大規模日本語 N-gram データ[4]の 1-gram を使用することで解決を図る．1-gram データでの数値が高ければ高いほど一般性が高いとして，この数値の逆数を重みとして tf に掛ける．これにより，tf のみを用いるよりも関連語として適切な単語をユーザへ提示することができる．

５．評価実験

話題語「マイナンバー」を用いて，提案手法の妥当性を確認する評価実験を行った．スコアの算出には，tf に 1-gram データの逆数を掛ける４．３節の方法を用いた．表 1 は実験の際に抽出された関連語リストの一部である．

表 1 話題語「マイナンバー」における関連語リスト

局員	流出	役所	年賀状	送金
配達	提出	書類	郵便	制度

５．１ 被験者

本学の学生 6 名を対象とした．

５．２ 方法

(1) 検索クエリ「マイナンバー」での検索結果を 100 件提示する．
(2) 被験者が必要な関連情報であると感じたツイート数を数えてもらう(ベースライン)．
(3) 関連語リストを提示する．
(4) 関連語をひとつ選択してもらう．
(5) 検索クエリ「マイナンバー　関連語」での検索結果を 100 件提示する．
(6) 被験者が必要な関連情報であると感じたツイート数を数えてもらう(提案手法)．

５．３ 結果

評価実験の結果を表 2 に示す．

表 2 ベースラインと提案手法の必要な関連情報数

	被験者A	被験者B	被験者C	被験者D	被験者F	被験者G
ベースライン	11	24	7	7	6	10
提案手法	24	41	14	16	13	11

全ての被験者で，提示した関連語を検索クエリに含めた方が必要な情報を含むツイートが多く出力されるという結果となった．

６．結論

マイクロブログから話題語の関連情報を得る手法を提案した．抽出した関連語を用いて検索することで，ユーザが求める情報を効率的に取得することができるという結果が得られた．

参考文献

[1] 総務省：平成 27 年版情報通信白書， 第 2 部 第 4 章 第 2 節， 2015.
[2] 岩木 祐輔， アダムヤトフト， 田中 克己：マイクロブログにおける有用な記事の発見支援， DEIM Forum 2009， 2009.
[3] MeCab：http://mecab.sourceforge.net/
[4] 工藤拓， 賀沢秀人：Web 日本語 N グラム第 1 版， 言語資源協会， 2007.