3P-3

マイクロブログからの地域の話題抽出に関する研究

北野光一† 寺口敏生† 田中成典‡ 西江将男† 中本聖也‡
関西大学大学院総合情報学研究科† 関西大学総合情報学部‡

１．はじめに

近年，携帯端末やカーナビゲーションの普及に伴い，現在位置を基に地域情報を参照するサービスが増加[1]している．しかし，これらの位置情報サービスに用いられる地域情報はリアルタイム性に乏しいという問題がある．そこで，著者らは，リアルタイム性の高いソーシャルメディアであるマイクロブログから最新の地域の話題を抽出する手法について研究を行っている．話題抽出に関する既存研究には，文書に出現する単語の特徴を用いて話題を抽出する手法[2]-[4]と学習データを用いて文書集合から話題を抽出する手法[5]がある．しかし，前者の手法では，造語や省略語などの自然言語の曖昧性の問題から単語間の類似度を正しく評価できない問題がある．一方，後者の手法では，新しい情報に含まれやすい未知語を学習データとしてリアルタイムに用意できない問題がある．これらの問題から既存手法はリアルタイムな話題の抽出に活用することが難しい．そこで，本研究では，マイクロブログの文書中に含まれる単語を検索エンジンに入力し，検索結果の Web ページに含まれる単語群を用いて単語間の類似性を評価することで，造語，省略語や未知語を考慮したリアルタイム性の高い地域の話題を抽出する手法を提案する．

２．研究の概要

本研究では，マイクロブログからリアルタイム性の高い地域の話題を抽出する手法を提案する．本システムの概要を図 1 に示す．本システムは，1）マイクロブログ収集機能と 2）話題抽出機能により構成される．入力データは，地名や建物名などの場所に関する名称とし，出力データは，リアルタイム性の高い地域の話題とする．

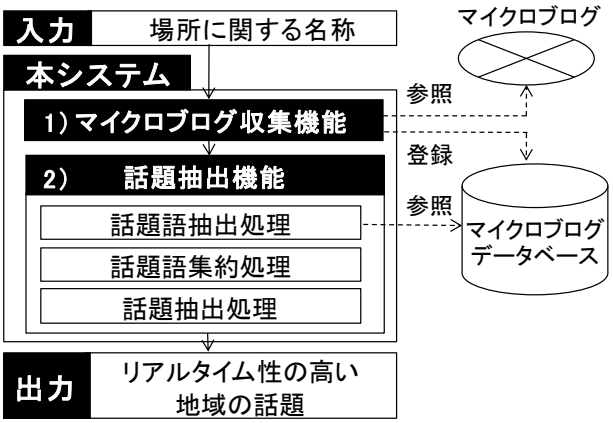

図１ 本システムの概要

２．１ マイクロブログ収集機能

本機能では，マイクロブログから地域に関する文書を収集する．まず，入力した場所に関する名称を検索語としてマイクロブログから文書を収集する．次に，収集した文書には，地域と関連のない情報が多く含まれるため，自動的に投稿を行う BOT の文書，他者への返信や引用といった投稿者自身の体験ではない文書やリアルタイム性の低い過去と未来の話題を扱う文書をノイズとして除去する．最後に，これらの文書を地域に関する文書としてマイクロブログデータベースに格納する．

２．２ 話題抽出機能

本機能では，地域に関する文書からリアルタイム性の高い話題を抽出する．本機能は，話題語抽出処理，話題語集約処理と話題抽出処理により構成される．話題語抽出処理では，マイクロブログデータベースに格納された文書に対して形態素解析を行い，連接する形態素の品詞の組み合わせから複合語を取得することで，話題となる話題語を抽出する．話題語集約処理では，N-gram を用いて文字単位で類似する話題語を集約する．そして，話題語を検索エンジンに入力し，検索結果の Web ページに含まれる単語を用いて話題語の特徴ベクトルを算出する．これらの特徴ベクトルから話題語間の類似度を算出す

Research for Extracting Area Topics from Microblog
† Koichi Kitano, Toshio Teraguchi, Masao Nishie,
Graduate School of Informatics, Kansai University, 2-1-1 Ryouzenji-cho, Takatsuki-shi, Osaka 569-1095, Japan
‡ Shigenori Tanaka, Seiya Nakamoto
Faculty of Informatics, Kansai University, 2-1-1 Ryouzenji-cho, Takatsuki-shi, Osaka 569-1095, Japan

ることで，類似度の高い話題語同士を共通の話題語として集約する．話題抽出処理では，集約した話題語群の時間帯毎の出現頻度を考慮し，現時点から一定期間中に特徴的な話題語群をリアルタイム性の高い話題として抽出する．

３．実証実験と考察

本提案手法の有用性を実証するために，マイクロブログの Twitter を解析対象とし，話題抽出の精度について本提案手法と N-gram を用いた既存手法[3]で比較実験を行う．

３．１　実証実験

本実験では，Twitter 上にて多くの人が投稿した地域の話題 25 件を実験対象とし，Twitter 上から関連する文書約 4,500 件を収集する．収集条件として，話題となった時間帯に投稿された文書の中で，関連する地名を含む文書を選択する．これらの実験対象から関連する地名と時間帯に基づき正しく話題が抽出できたかを評価する．評価指標としては，抽出した話題中における正解データの割合を話題抽出率として算出する．

３．２　結果と考察

実証実験における話題抽出精度の実験結果を表 1 に示す．既存手法では，同時期に複数の話題がマイクロブログ上に投稿される場合や，話題が様々な表現で記述される場合に，別々の話題として抽出を行った．一方，本提案手法では，意味の類似した話題語を集約して一つの話題として抽出したため，精度が向上したと考えられる．したがって，本提案手法は，既存手法と比較して有用であることを実証した．本提案手法を用いて抽出した 2010 年 12 月 31 日 12:00～13:00「金閣寺」の話題の抽出結果を表 2 に，2011 年 1 月 10 日 5:00～8:00「西宮神社」の話題の抽出結果を表 3 に示す．表 2 と表 3 の結果の通り，話題に関する意味の類似した造語，省略語や未知語を集約できていることを確認した．一方，意味が類似しない話題語間でも，検索エンジンから検索結果として得られた Web ページの構造が類似する場合には，共通する単語が増加するため，類似した話題語として集約され正しく話題語を集約できない問題が生じた．この問題に対しては，Web ページの HTML タグ構造に着目し，本文とは関係のない不要な部分を除去することで，解決できると考えられる．

表 1　話題抽出精度の実験結果

	本提案手法	既存手法
話題抽出率	88%	80%

表 2　2010 年 12 月 31 日 12:00～13:00
「金閣寺」の話題の抽出結果

話題	話題語
雪化粧	雪化粧，雪景色，真っ白， 積雪，銀世界

表 3　2011 年 1 月 10 日 5:00～8:00
「西宮神社」の話題の抽出結果

話題	話題語
福男	福男，野球ユニフォーム姿， 十日戎，一番福，一番福男選び

４．おわりに

本研究では，マイクロブログからリアルタイム性の高い地域の話題を抽出する手法を提案した．実証実験の結果，同じ意味を持つ造語，省略語や未知語を集約し，本システムの有用性を証明できた．今後の課題として，本提案手法では話題抽出機能において時間の変化のみに着目したが，地域毎に特徴的な話題があると考えられるため，地域毎の話題の発生傾向を学習させることで地域性の高い話題の抽出手法の実現に取り組む．また，今後の発展として，本研究では地域における話題の抽出を目的としたが，地域に発生する犯罪情報や災害情報も抽出対象としてリアルタイムに抽出することで，地域の安心安全の分野に適応することを考えている．

参考文献

[1] 総務省：平成 21 年度版情報通信白書，ぎょうせい，2009.7.

[2] 中村健二，吉村智史，北野光一，田中成典，古田均：GA を用いた Web ニュースの時系列情報を考慮したトピック抽出に関する研究，情報処理学会論文誌，情報処理学会，Vol.49，No.7，pp.2480-2492，2008.7.

[3] 藤木稔明，南野朋之，鈴木泰裕，奥村学：document stream における burst の発見，自然言語処理研究会報告，情報処理学会，Vol.2004，No.23，pp.85-92，2004.3.

[4] Hatzivassiloglou, V., Gravano, L. and Maganti, A.: An Investigation of Linguistic Features and Clustering Algorithms for Topical Document Clustering, Proceedings of 23th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, pp.223-231, 2000.7.

[5] Takeshi, S., Makoto, O. and Yutaka, M.: Earthquake Shakes Twitter Users: Real-Time Event Detection by Social Sensors, Proceedings of the 19th International Conference on World Wide Web, ACM, pp.851-860, 2010.4.