ISS-P-148

共起とハッシュタグを用いた ツイートのカテゴリ推定の検討

山本　宗典†　　須鎗　弘樹†
† 千葉大学大学院融合科学研究科

1. はじめに

Twitter では，タイムラインを閲覧する際に情報量があまりに多くなってしまうと，視覚的にユーザの負担が増えてしまう場合がある．そのような問題点への対策として知られるのがキュレーションであり，その代表である Togetter では話題によってツイートを他のユーザがまとめたものを閲覧することができる．このように，カテゴリや話題ごとにツイートをまとめて管理したり，他にも，各ツイートのカテゴリのみを見出しに表示しておいたりすることにより，ユーザの負担を軽減できることが期待される．しかし，既存のキュレーションではツイートのまとめを人手で行っており，個別のツイートのカテゴリを推定し，自動でまとめを行うようなものはまだ知られていない．

そこで本研究では，入力ツイートに対し，そのツイートが何のカテゴリに関するものなのかを推定する手法について検討し，現時点での推定結果を求める．

2. 関連研究

文書のカテゴリ推定に関する研究は以前より広く行われており，テキストカテゴライゼーションとクラスタリングという二つの手法に大きく分けることができる．

テキストカテゴライゼーションはあらかじめ設定されたカテゴリに自動的に分類する手法であり，近年では機械学習による研究が主流である[1]．この場合，未知のカテゴリに推定することは不可能であり，拡張性に欠ける．一方，クラスタリングは，文書集合の中で類似する内容の文書をグループ化し，そのグループ内で頻出する単語などを利用し，カテゴリを推定する手法であり[2]，例えば，一つだけの文書を入力する際などには使用できないことが欠点である．

それに対し本研究では，これら二つの手法の欠点を補うハッシュタグを用いた新たな手法を検討する．

4. 推定手法

推定手法としてはまず，推定したい入力文に対して，その文のカテゴリとの関連性が低いと思われる単語を取り除くためにキーフレーズ抽出を行う．抽出には Yahoo!のキーフレーズ抽出 API を使用した．キーフレーズを抽出した後は，それらのフレーズをクエリとして，Twitter Search API により，各クエリに対してハッシュタグを含むツイートのみを一定数ずつ取得する．最後に，取得したツイート群に発生したハッシュタグをカウントし，カウント数の最も多いものを推定結果とした．

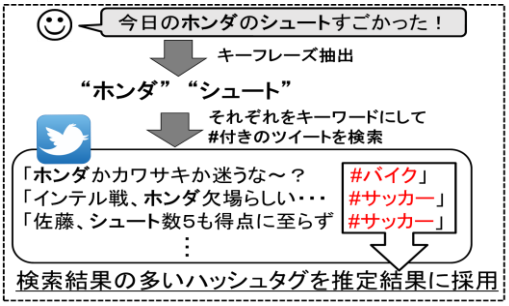

図 1 推定手法の流れ

表 1 正答推定ツイート数

	サッカー	野球	競馬	映画	アニメ	音楽	AKB48	ファッション	グルメ	政治
提案手法	24	18	32	7	5	28	42	34	13	12
SVM	40	21	44	25	31	38	34	42	48	39

5. 実験

実験では，正解とするカテゴリがハッシュタグとして付与されたツイートからハッシュタグの部分を除去したものに対して推定を行い，サッカーや政治など 10 カテゴリ各 50 計 500 ツイートを対象とした．また，比較対象として各カテゴリ 500 計 5000 ツイートを用いて SVM で学習し，推定した結果も求めた．SVM のカーネルには線形カーネルを使用し，素性には 936 次元の bag of words を用いた．

各カテゴリ 50 ツイート中，正しく推定できたツイート数を表 1 に示す．全体の正答率は，提案手法では 43%，SVM では 72.4%となった．ただし提案手法では，推定結果がそのカテゴリを表しているといえるものでも，元々の正解カテゴリに完全に合致しておらず不正解となったものが多いので，その点を考慮した評価を行えば正答率は上昇するであろう．また，SVM による推定はカテゴリ数が決まった中での分類なので一概には適切な比較対象とはいえない．

6. まとめと今後の予定

本研究では，ツイートのカテゴリを推定する手法を検討し，現時点での実験結果を示した．今後は正答率が上がるよう手法改良を行う予定である．具体的には，それぞれの単語が持つ意味を単語間で計算することを検討中である．

参考文献

[1] 松本一則ほか，“web コンテンツのジャンル推定に向いた実用的な 2 段階 SVM の構築”，情報処理学会研究報告情報学基礎研究会報告，2013-IFAT-111(22),1-3, 2013-07

[2] 佐藤進也，高橋公海，松尾真人，“特徴抽出を目的とした文書クラスタからの一貫性阻害要素除去”，情報処理学会論文誌，データベース 6(3), 1-12, 2013-06

Copyright © 2015 IEICE