Twitter のハッシュタグを用いた多様なメディアにおけるトピック同定

佐藤　和人

近年，Twitter や Facebook などのソーシャルネットワーキングサービス(SNS)をはじめとする Web 上のサービスへの社会的関心が高まっている．Web では多くの情報をリアルタイムに収集できるため，この利用価値は高いが，Web に投稿されるテキストは整理されていない場合が多く，求める情報を選択的に収集することは困難になってきている．このため，機械的な処理によって Web 上のコンテンツを分類するという研究課題は重要さを増している．

SNS にはコンテンツを分類する手法として「タグ」が存在し，Twitter においては「ハッシュタグ」が導入されている．タグは，付与することで特定のトピックに関連する投稿を検索することが容易になるという効果が期待できる．しかし多くの場合，タグによって検索できるのはタグの利用されているメディア上のコンテンツのみであり，そのトピックに関係する他のメディア上のコンテンツとの関連はない．

本研究では，Twitter 以外のメディアのテキストに対し，Twitter で利用されているハッシュタグを推薦することで対象のテキストのトピックを同定する手法を提案する．具体的には特定のハッシュタグの現れるツイート全てを繋ぎ合わせたハッシュタグ文書を仮定し，ここに現れる語彙について TF-IDF ベクトル化を行う．これを *k*-means 法でクラスタリングすることで特定のトピックを表すハッシュタグが同一のクラスタにまとめられる．得られたクラスタを用いて別のメディアのテキストにハッシュタグクラスタを推薦する．これにより，ユーザの設定したタグを利用したトピック同定は可能か，あるいは異なるメディアにおける統一的なタグの利用は可能かを明らかにする．対象は収集するハッシュタグと同時期の新聞記事とする． 実験では，単にハッシュタグを推薦する場合とハッシュタグクラスタを推薦する場合との比較を行う．クラスタの推薦には k 個の近傍なハッシュタグの属するクラスタのうちより数の多いクラスタを推薦する k 近傍法と近傍な重心を持つクラスタを推薦する重心法の 2 種類を用いる．

この結果として，クラスタを推薦する場合ではハッシュタグ単体を推薦する場合と比較して適合するハッシュタグの絶対数は，*k*近傍法では 2.20 倍，重心法では 1.80 倍となった．また，重心法によるクラスタの推薦では，ハッシュタグ単体の推薦に比べて適合率のマイクロ平均がやや上がった．より多くのハッシュタグを推薦するクラスタ推薦において，ハッシュタグ単体を推薦する場合と比較してもそれほど適合率のマクロ平均が下がらなかった．以上の結果から，他のメディアのテキストに対して推薦を行う場合，クラスタを推薦する手法はハッシュタグ単体を推薦する手法と比較して，適合率のマイクロ平均の面でも，推薦を正しく行えたタグの絶対数という面でも，優れていることがわかった．

（指導教員　若林　啓）