表記の多様性を考慮したハッシュタグ推薦

井上　優作

Twitter や Facebook などのソーシャル・ネットワーキング・サービス (SNS) には日々大量にテキストが投稿されており，それらを系統的に整理して検索可能にすることは重要な課題である．特に，SNS に投稿される情報はリアルタイム性が強く，またユーザにしか知り得ない感想や情報が含まれることから，ライブや展示会，交通障害，災害，テレビ番組などといったイベントに関する情報源としての利用に注目が集まっている．SNS から特定のイベントに関連した投稿を収集することができれば，当該イベントの主催者や当該イベントに関心のあるユーザにとって有益である．また，多くの SNS では，投稿の内容のテーマを表す文字列を「ハッシュタグ」として付与することで，同様の内容の投稿を検索しやすくする機能がある．本研究では，このハッシュタグから得られる情報を用いて特定のイベントについての特徴を学習し，ハッシュタグの推薦を通して SNS 上のユーザが特定のイベントに参加していたかどうかを推定する手法を提案する．

具体的な手法としては，SNS に投稿されたハッシュタグを伴うテキストからハッシュタグごとの TF-IDF ベクトルを作成し，そのベクトルを適切なクラスタ数でクラスタリングすることでクラスタが現実のイベントに対応するようにする．クラスタリングが必要なのは，現実におけるイベントに対応するハッシュタグが SNS 上に 1 つしか存在しないとは限らないからである．本研究では k-means 法でクラスタリングを行う．そして，対象のイベント時間中にあるユーザが投稿したテキストを結合したものを 1 文書と見なして，その文書に推薦されたハッシュタグクラスタ内に対象イベントに関するハッシュタグが含まれるかどうかでそのユーザが対象イベントに参加していたかどうかを判定する．

実験では，ハッシュタグをクラスタリングするのに適切なクラスタ数と，そのクラスタ数における精度と再現率を，ハッシュタグクラスタの重心を用いる方法（「重心法」とする）と，単独のハッシュタグを基準とした k-近傍法（「近傍法」とする）の 2 つの推薦方法で比較する．結果として，クラスタリングする際のクラスタ数は，重心法ではクラスタ数が大きい方が，近傍法では全体の異なりハッシュタグ数に対して 0.6-0.7 倍の場合が正解率が高く，全体を通して近傍法の方が正解率が高かった．精度と再現率の実験では，精度は重心法と近傍法の両方でほぼ 1 だったが，再現率は重心法よりも近傍法の方が値が高く，結果的に F 値も近傍法の方が高くなった．以上をまとめると，全体の異なりハッシュタグ数に対して 0.6-0.7 倍程度の数でハッシュタグのクラスタリングを行い，近傍法によってユーザの投稿テキストに対してハッシュタグクラスタを推薦すればよいというのが本研究の結論である．

今後は，本研究のようにハッシュタグを用いずに，SNS に突発的に大量に現れた語彙を特徴と見なして自動でイベント発生の検出を行うシステムの研究が求められる．

（指導教員　若林　啓）