Twitter ユーザに対するゼロショットタグ付け

新田　洸平

ソーシャルネットワーキングサービスにおいてユーザの性質を明らかにすることは重要である．ユーザの性質が分かることで有益な判断が容易になる．例えば，ユーザの性質からそのユーザが発信する情報の傾向がわかることで，情報源として有効かどうかの判断が容易になる．また，何らかのイベントがあったとき，イベントに参加するユーザの性質からイベントの性質がわかることで，興味と合致しているかどうかの判断が容易になリ，イベント推薦などに活用できる．ユーザの性質を明らかにするような研究として，属性に基づくユーザ分類とユーザに対するタグ付けに分かれる．属性に基づくユーザ分類は，クラスがあらかじめ決められているような問題設定である．例えば，機械学習手法を用いてユーザを年齢や性別，政治的思考などで分けるような研究が行われている．ユーザに対するタグ付けは，クラスが明示的に決められていないような問題設定である．例えば，ユーザのプロフィールや投稿などの関連情報に含まれる名詞の単語をタグ付けするような研究が行われている．既存のユーザ分析における課題として，具体的な性質の表現が難しくタグの表現が単語に限定されていること，教師あり機械学習手法を用いる場合，学習データ中に含まれていないパターンのユーザを正しく予測できないこと，学習データの作成に大きなコストがかかることが挙げられる．本研究では，ゼロショット学習手法を用いてユーザとタグとの対応関係を学習することで学習データがない場合でもユーザに対してタグ付けする手法を提案する．具体的にはゼロショット学習と呼ばれる機械学習手法の応用手法を用いる．ゼロショット学習では，学習データにおいてラベルに対する事例が無いもしくは予測するために十分ではない場合において，新たな未知の入力に対しても何らかの入手可能な既知の情報をを用いることで学習データを用いたモデルの訓練時に一度も出現しないようなパターンの事例をも予測できる手法である．このゼロショット学習手法を応用して，Twitter におけるユーザのツイートとユーザが含まれるリストの名前をタグとして用いて，ユーザの表現とタグの表現を単語埋め込み空間に写像し，単語埋め込み空間上での対応関係を獲得することで，学習時に出現しないタグをもユーザに対して付与する．タグの表現をタグに含まれる各単語の埋め込みを平均したベクトルを使用することで，既存手法のタグの表現が単語に限定されるという課題にも対処している．実験では，提案手法の学習においては文書分類タスク，予測においては文書検索タスクに取り組み，ユーザとタグの適合度によって順位付けを行った．実験に用いたデータセットは Twitter から収集したリストとユーザ単位のツイートから作成し，文書検索タスクのベースライン手法と提案手法の有効性について，Hit@k, nDCG@k などを用いて上位 k 件に正解タグが含まれるかどうかで評価を行い，評価結果を比較した．

（指導教員　　加藤　誠）