マイクロブログにおける実生活に関する記事の抽出法の研究

山本　修平

マイクロブログの代表として Twitter は，多くのユーザがリアルタイムに，自分がしていることや考えていることなど，より身近な「今」を投稿している．投稿する記事の中には，ユーザの経験や知識，地域の生活に関することなど，地域性が高く新鮮な情報がある．このような情報は，他のユーザの実生活における活動の支援に役立つと考えられる．しかし，Twitter 上では誰かの投稿に対する相槌や共感といった記事も多く，実生活に言及している記事が埋没してしまっている．また，Twitter に投稿される文章は主語や目的語が抜けているものが多い．このため，膨大な記事の中から実生活に言及している記事を選択的に抽出することは，単純なキーワード検索では困難であった．

本研究では，ユーザの実生活における活動に有用な情報を提供することを目的に，Twitter 上から実生活に言及している記事を抽出する技術を確立する．抽出した記事に対して,「食事」や「交通」等の生活の局面などの属性として付与し，実生活データベースを構築する．このような属性を付与することで，生活の局面に応じた実生活情報をリアルタイムにユーザに提供できると考えている．また，本研究では「茨城県つくば市」に焦点を当てて実践的に研究を進めるため,「つくば市」を生活圏とするユーザの記事を対象とする．

提案手法では,「実生活」を 14 の局面からなる生活情報と定義する．局面毎に実生活に特徴的な単語を集めた「実生活辞書」を生成し，実生活辞書に基づいた実生活に言及している記事を抽出をする．辞書に登録された単語は，局面毎に「実生活らしさ」という値を持つ．抽出した記事には，14 のいずれかの局面が付与される．

予備調査で収集したつくば市を生活圏にする投稿者の記事 4,000 件と，public timeline からランダムに収集した記事 4,000 件に対して，実生活に言及しているか否かを人手により分類した．実生活と分類した記事には，更に 14 の局面のいずれかの情報を付与した．3,000 件の記事を基に実生活辞書を生成し，残りの 1,000 件をテストデータとして実生活情報の抽出実験を行ったところ，F 値で 70%を超える結果を得られた．また，記事に付与された局面の正解率を調べる実験も行った．その結果，5 つの局面で 50%以上の正解率を得られた．提案手法と SVM を用いた手法を比較した結果，提案手法が高い評価値を示し，有効性を確認できた．

今後の課題として，単語数と局面に考慮して動的に閾値を変動させる手法を実装し，抽出精度の変化を分析すること，実生活情報へ複数の局面を付与することが挙げられる．

（指導教員　佐藤 哲司）