プロ野球の実況ツイートを対象とした マルチラベル分類

山田　慎也

要旨

近年，ユーザが個人の視点から情報発信が可能なソーシャルネットワーキングサービスや，マイクログログから爆発的に普及している．代表的なものとしてTwitterがあげられる．Twitterの特徴として，簡易さから「今，何をしているのか」などのユーザの現在の状況を発信することが多く，リアルタイム性が高い情報交換ツールとして多くの人に活用されている．ツイートは，随時更新されるタイムライン（timeline）上に時系列順に表示され，ほぼリアルタイムで情報を獲得することが可能である．ここで注目すべき点は，Twitterの簡便性とリアルタイム性である，これらの性質より，インターネット上の一般的なブログと比べ，Twitterでは，自分が必要とする情報，データが入手しやすいと考えられる．Twitterでは，テレビ番組やスポーツ中継の際に実況書き込みが行われ，ユーザーの考えや意見，感想がわかるようになった．これらのことを背景に，本研究では，プロ野球の実況書き込みを対象に，野球の試合の実況ツイートを収集する．具体的には，2013年5月14日から6月20日の間に行われた,「2013年度日本生 命セ・パ交流戦」の読売ジャイアンツ全24試合を対象とした．野球の試合は，9イニングという決められた回数の中で，攻撃と守備がはっきり分かれているため，実況書き込みをしやすいと考えた．ツイート内容から判断して決めた解説，状況，感想，応援，野次，その他のラベルに分類する．実際に，収集した全47000のツイートをラベルごとに手で振り分けた．野球の生のツイートを集め，これだけのラベルを付与したデータは他にはない．しかしツイートの量が多く，手でふるのは困難である．そのため，機械学習手法を用いて，ラベルづけを行うモデルをデータから自動的に生成する．また，Tweetの自動分類を実現することは，ユーザの考えや感想，意見を集約したまとめサイトなどを構築する基礎となると考えている．一般的な分類は単一クラス問題となるが，ツイートの場合，140文字とはいえ，複数の内容を含み，結果として複数のラベルを付与することが必要となる場合も少なくない．従って，ツイートの分類をマルチラベル分類の枠組みでとらえることとした．分類の評価実験に関しては,「どの手法を利用するのが良いのか？」「どのく らい学習データが必要・適切なのか？」ということを明らかにする目的で，利用するアルゴリズムやデータの準備の仕方を変え，種々の評価値を用いた多角的に比較を行う．結果として，手法ごとの差はあまりなかったが，ラベル別では，よく当たるラベルと当たらないラベルに差が出た．