1V-5

Twitterにおける専門家判別手法の性能評価

鎌田 健史† 長谷川 大 佐久田 博司
青山学院大学 理工学部 情報テクノロジー学科‡

1 序論

ソーシャルネットワーキングサービス Twitter [1]の普及によって個人の情報発信がより活発になった．その結果，情報を発信することが人々にとってより身近なものとなり，より多くの情報が得られるようになった．これにより，我々は情報を取捨選択し，自分にとって有益だと思われる情報のみを選別する必要がある．Twitter 上では，ユーザは別のユーザをフォローすることで彼らが発信する情報を得ることができる．しかし，これらの情報を発信するユーザには，専門的な知識を持ち有益な情報を発信するユーザや，専門性の低い情報を発信するユーザが混在している．これらはプロフィールを一読しただけでは判断がつかない．したがってプロフィール以外からそのユーザの専門性を判断する必要がある．

本研究は，機械学習を用いることで，ユーザが対象のドメインに対して専門的な知識を持つ専門家かどうかを判別することを目的とする．本研究における機械学習のための素性として，Twitter のツイート内容に含まれる専門用語の有無，実世界の出来事との共起情報，抽象的な単語の使用頻度を用いる．

今回は，興味や関心を持ち情報を発信するユーザが多くいること，専門性のある内容からそうではない内容まで幅広い内容がツイートされていることから，サッカーを対象に行う．

2 関連研究

池田ら [1] は Twitter において，ユーザのツイートからそのユーザの年齢，性別，在住地域を判別した．彼らは，あらかじめ年代，性別，在住地域などをプロフィールに記載しているユーザを取得した．そして，それらのユーザのツイートから抽出した特徴語を，年代を判別する辞書，性別を判別する辞書，在住地域を判別する辞書とし，これらを素性として用いて，ユーザの分類を行った．しかし，彼らの研究ではユーザの専門性の判別は行っていない．

小紫ら [2] は，web ページ集合内で特定のクエリトピックに関連する単語の一般性を定量化する手法を提案した．この手法では，web ページ内でのその単語の出現回数や，web ページ内での出現位置，ページタイトルやリンクテキストなどを利用して数値化している．しかし，彼らは web ページのみを対象にしているため，Twitter における単語の一般性評価や，評価の結果をもとにしたユーザの専門性の判別は行っていない．

3 素性の選択

3.1 概要

本研究では，Twitter において，ユーザのツイートを取得し，その中に含まれる専門家推定に役立つ素性を検出することで専門性を推定する．

本研究では，サッカーの専門家を「サッカーについて，Wikipedia などインターネットでの検索によって簡単に得られることのできない知識を有している人物」とした．著者 1 名によるサッカーに関連するツイートを分析した予備調査において，上記の定義に合ったサッカーに関する専門的知識を持つユーザは以下の 3 つの特徴を有している傾向があることを確認した．

1. サッカーに関連する用語が多様である
2. 現実のイベントと共起したツイートが見られる
3. サッカーの理論的な説明の際に抽象的な単語を用いる

以下，各特長について，説明する．

3.2 サッカーに関連する用語の収集

サッカーについて常に，より幅広い話題について言及していれば，そのユーザは常にサッカーのことを考えていることが想像できる．こういったユーザは，サッカーという特定の項目に対して常に考えていることが予想されるので，より専門性が高いユーザであると推察される．

そこで，サッカーに関する用語を収集する．サッカー用語の辞書として，日本語版 Wikipedia に登録されているページから，カテゴリ名に「サッカー」を含む語がある約 83,600 ページのタイトルを収集する．表 1 に収集したタイトルの例を示す．

Performance Evaluation of expert determination in Twitter
† Kenshi KAMATA (a5809029@aoyama.jp)
‡ Department of Integrated Information and Technology, College of Science and Engineering, Aoyama Gakuin University
[1] http://twitter.com

表 1　Wikipedia にあるサッカー用語のタイトル例

カテゴリ	タイトル名
選手名	リオネル・メッシ
	ディエゴ・マラドーナ
	ラモス瑠偉
チーム名	トッテナム・ホットスパー FC
	AC ミラン
	レアル・マドリード
技術名	インサイドキック
	マルセイユ・ルーレット
	トータルフットボール
その他	サポーター
	国立競技場
	プレミアリーグ
	etc.

3.3　試合時間データベースの作成

予備調査において，専門性の高いユーザーは深夜にもツイートをすることがあるという特徴が見られた．これは専門性の高いユーザは国内外を問わず，常に試合を観戦し，それらについて Twitter 上で即座に自分の意見を述べたり，試合状況についてツイートするからであると考えられる．そこで，試合時間内のツイートを取得するため，試合時間辞書を以下の手順で作成する．

1. サッカー・カレンダー[2]に掲載されている 2013 年 1 月 6 日から過去 1 年分の試合開始時刻を集める．
2. これらの時刻から重複しているものは取り除き，2715 試合の試合開始時刻を取得する．

3.4　抽象語の収集

予備調査の結果から，専門性の高いユーザーは「統計的」,「勇気がある」などの抽象語を多用するという特徴が見られた．これは専門家は，自分の意見をフォロワーに説明するため，抽象的な単語を多用していると考えられる．

そこで，ツイート内に存在する抽象語を収集する．抽象語の辞書として，九州大学大学院言語文化研究院が web 上に公開している「A Passage to English 大学生のための基礎的英語学習情報」資料[3]に記されている英語の抽象名詞，抽象動詞，抽象形容詞を和訳した単語 2045 単語を辞書として用いる．抽象的な単語の例を表 2 に示す．

表 2　抽象用語の例

品詞	単語			
名詞	不安	節約	努力	恐怖
形容詞	正確な	不器用な	気まぐれな	攻撃的な
動詞	大切にする	利用する	抵抗する	戦う

[2] http://soccer-calendar.heroku.com/game

[3] http://www.flc.kyushu-u.ac.jp/ passage/passage11.html

4　実験データ収集

実験データ作成の概要を図 1 に示す．Twitter API[4]を利用して，ユーザのツイートを取得する．1 回の API リクエストで 200 件のツイートとその情報を取得できる．この情報から 200 件のツイート内容とツイート時間を取得する．そこからサッカーに関する用語や，抽象語を判別するために，ツイート内容から形態素解析を用いて名詞，動詞，形容詞を抽出する．形態素解析には MeCab[5]を用いる．動詞や形容詞は，ツイートから抽出された単語をそのまま利用するのではなく，それらの基本形を用いることとする．実験データのラベル付は 3 人の評価者によって行った．3 人がそれぞれユーザに対してラベル付を行い，3 人の内 2 人の評価が一致したものをラベルとして用いる．これらのデータを組み合わせることで実験データとする．

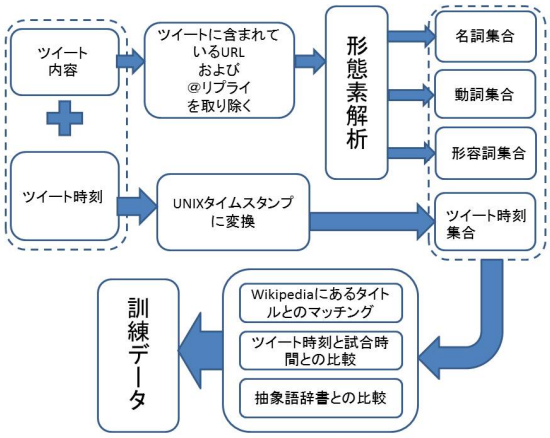

図 1　実験データ作成概要図

5　結論

本稿では，Twitter の情報から専門家とそうでないユーザを機械的に判別する手法を提案した．今後の予定として，機械学習の分類器として SVM を用いた実験により，精度評価を行う．

また今後は，本研究で用いた素性が他のスポーツや，科学技術など他のドメインにおいても有効かどうか，そうでないならどのような素性が有効か検証する必要がある．

参考文献

[1] 池田和史, 服部元, 松本一則, 小野智弘, 東野輝夫. マーケット分析のための twitter 投稿者プロフィール推定手法. マルチメディア、分散協調とモバイルシンポジウム 2011 論文集, 第 2011 巻, pp. 1308–1315, 2011.

[2] 小紫弘貴, 田島敬史. クエリトピックに対する一般的知識範囲の推定. *DEIM Forum 2012*, 2012.

[4] https://dev.twitter.com/

[5] http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html