単語の特徴に着目したTwitterユーザのリスト分類

松本 悠　　斎藤 博昭
慶應義塾大学理工学部 情報工学科
{matumoto, hxs}@nak.ics.keio.ac.jp

1 はじめに

近年ではTwitter[1]が重要な役割を担っている．これはTwitterを通じて友人や興味のあるユーザとの密接な繋がりを構築できるといった点が挙げられる．また個人のみの利用だけでなく，企業や政府機関などでもTwitterの活用がみられる．これは最新の情報の発信などで情報の格差の解消や，災害時におけるデマなどの誤った情報を排除する働きがあるため重要視されている．このようにTwitterは他ユーザとのコミュニケーションを目的とする側面と，自身にとって有用な情報を素早く取得することを目的とした側面があると考えられる．

Twitterにおいてはユーザをフォローすることにより，多数のユーザのツイートと呼ばれる140文字以内からなる短文を時系列に並べてタイムラインと呼ばれる画面に表示することができる．しかし多くのユーザをフォローしているユーザにとっては一度にすべてのツイートを見ることは困難であり，また有益な情報を見逃す可能性も低くはない．Twitterにはリストというユーザをグループに分け，そのグループ毎にタイムラインを表示させる機能があるため，上記のような問題をさけることは可能である．しかしながら多くのユーザにとってリストを作成するのは容易ではない．理由として次の2点が挙げられる．

1. リスト数はTwitter使用者によって異なる
2. それぞれのユーザがどのリストに含まれるかを考えなければならない

本稿では同じ内容を多くツイートするユーザは同じリストに分類される可能性が高い，Twitter使用者ごとに最適なリスト数は異なるという仮定のもと，ユーザを自動でリストに分類する手法を提案する．

[1]https://twitter.com

2 関連研究

Pennacchiottiらは Twitterユーザの分類を4つのタスクで行った [1]．分類に用いた特徴量としてはプロフィール情報，ツイート頻度，ツイート内容，ユーザとの関わりなどを使用した．特にツイートの内容を特徴として使用するにあたり，Latent Dirichlet Allocation(LDA) という文書分類においてよく使用されるモデルを使用している．LDAでは与えた文書からトピックを抽出することができ，抽出したトピックを特徴として使用することができる．これらの特徴を使用し，ユーザの分類を行った．

Yamashitaらはフォロー関係を用いて Twitterのユーザの分類に関する研究を行った [2]．フォロー関係を使用することで，好みなどが近いユーザどうしが同じクラスに分類される．山下らの手法ではユーザ間のフォロー関係をグラフで表現し，グラフから隣接行列を作る．隣接行列をもとにした類似行列というユーザ間の近さを表す行列を作り，ユーザの分類を行った．

Bergsmaらはツイートに付与される言語データおよび地域データを用いてユーザの素性 (国，言語，宗教など) に関する分類を行った [3]．言語データおよび地域データから各ユーザの属性を決定し，K-meansアルゴリズムを利用してユーザの分類を行った．分類の結果から得られた各クラスタを素性として用い，サポートベクターマシンを用いた多クラス分類器によるユーザの分類を行った．

3 提案手法

本提案手法の概要を図1に示す．提案手法ではTwitter使用者のスクリーンネーム[2]を入力とし，Twitter使用者のアカウントがフォローしているユーザのツ

[2]Twitterユーザのアカウント名のようなもので”@”から始まる一意な文字列である

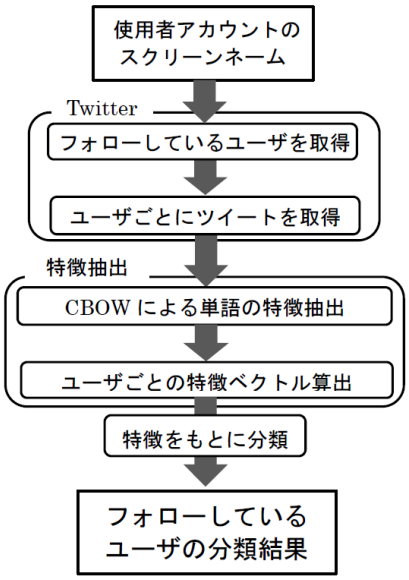

図 1: 提案手法の概要

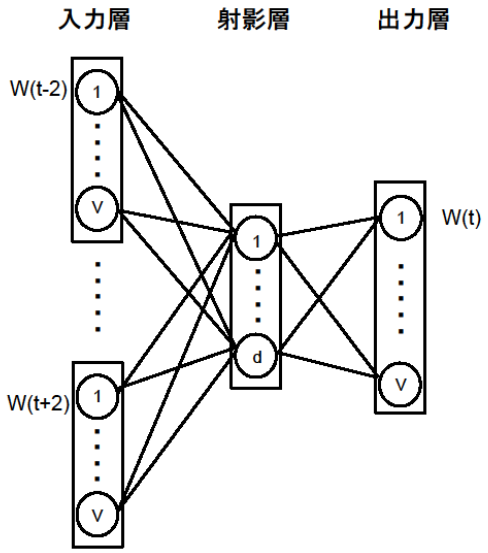

図 2: CBOW の概要

イートを取得する．ツイートから特徴を抽出することによりユーザの分類を行い，分類結果を本提案手法の出力とする．

3.1 単語の特徴抽出

本研究において単語の特徴抽出には Continuous-Bag-of-Words Model(CBOW)[4] と呼ばれるニューラルネットワーク言語モデルの一種を使用する．図 2 に CBOW の概要を示す．

図 2 のように CBOW は単語列 $w(t-2), w(t-1), w(t), w(t+1), w(t+2)$ のような順で存在するとき，単語 $w(t)$ を $w(t-2), w(t-1), w(t+1), w(t+2)$ から予測するようなモデルである．CBOW の入出力は単語の 1-of-V 表現であり，射影層では入力された各単語の 1-of-V 表現を $d(< V)$ 次元のベクトルに射影する，このとき射影された d 次元のベクトルを単語の特徴ベクトルとよぶ．単語の特徴ベクトルとは，学習コーパス上でその単語をよく表すベクトル表現のことであり，単語どうしの類似度などを計算することができる．CBOW では誤差逆伝搬法を用い，射影層への重みを学習する．学習終了後は入力単語の 1-of$V$ 表現と射影層への重みから，入力単語の特徴ベクトルを得ることができる．この特徴ベクトルを使用してユーザごとの特徴ベクトルを算出する．

3.2 ユーザの特徴ベクトル

Twitter を利用した研究では，特徴を抽出するための手段として LDA が多く使用されていることが奥村 [5] により報告されている．一方本研究ではユーザの特徴として，CBOW を使用して得られた単語の特徴ベクトルから算出したユーザの特徴を使用する．CBOW ではユーザが使用するすべての単語に着目して特徴を推定するため，LDA のようなトピックのみを特徴とするよりもユーザの特徴をうまくとらえることができると考えられる．CBOW のようなニューラルネットワーク言語モデルから得られる単語の特徴ベクトルは，単語間の類似度を使用して類義語などを判定するタスクなどに用いられる．本手法ではこれを応用し，異なる単語間の類似度ではなく，それぞれのユーザが使用する同じ単語の類似度からユーザの特徴を得る．以下ではユーザの特徴ベクトルを算出する方法を述べる．これ以降では単語の特徴ベクトルを CBOW ベクトルと表記する．CBOW ベクトルはすべて同次元であるとする．

全ユーザのツイートを使用して学習した CBOW ベクトル (全体の CBOW ベクトル) とユーザごとに学習した CBOW ベクトル (ユーザの CBOW ベクトル) からユーザの特徴ベクトルを算出する．ユーザ i の特徴ベクトルの t 番目の要素 x_{it} は

$$x_{it} = \begin{cases} 1 - sim(M(w_t), m_i(w_t)) & (w_t \in W_i) \\ 0 & (w_t \notin W_i) \end{cases}$$

と計算される．ユーザの数を k，全ユーザのツイートに含まれる語彙数を V とすると．$w_t(t = 1, 2, ..., V)$ は t 番目の単語，$M(w_t)$ は単語 w_t の全体の CBOW ベクトル，$m_i(w_t)(i = 1, 2, ..., k)$ は単語 w_t のユーザ i の CBOW ベクトル，W_i はユーザ i のツイートに含まれる単語集合，$sim(M(w_t), m_i(w_t))$ は $M(w_t)$ と $m_i(w_t)$ のコサイン類似度である．

ユーザの特徴ベクトルの各要素は同じ単語どうしの類似度をもとに計算される．単語 w_t の全体の CBOW ベクトルは全ユーザのツイートから得られた特徴であるので，単語 w_t の平均的な特徴が得られていると考えられる．一方単語 w_t のユーザ i の CBOW ベクトルはユーザ i のツイートのみで得られた特徴である．すなわち $sim(M(w_t), m_i(w_t))$ が 1 に近いほどユーザ i は単語 w_t を全体と同じように使い，0 に近いほど全体とは異なる使い方をしていると考えられる．1 からこの類似度を引いた値をユーザ i における単語 w_t の特徴値とする．すなわちユーザの特徴ベクトルの各要素は各単語の特徴値からなる．

3.3 リスト分類

はじめに述べたように，本研究では Twitter のリストを作成することを目的としたユーザ分類を行う．しかしながらリスト作成では事前にクラス数を知ることはできないため，本研究では事前に分類するクラス数を指定する必要のない Affinity Propagation 法 (AP 法)[6] を使用する．また分類をする際に，ユーザの特徴ベクトルは高次元でありそのままではうまく分類できないこともあるため主成分分析 (PCA) を使用して次元を削減する．AP 法および PCA には python ライブラリである scikit-learn[3] で実装されているものを使用する．

4 実験

4.1 実験データ

Twitter の認証済みアカウントを 80 人フォローしたアカウントを作成した．認証済みアカウントとは Twitter 社がそのアカウントが本人のものであると保証するもので，著名人，公的機関，企業など様々なアカウントが存在する．本実験では著名人，企業，スポーツチームなど多くの人が目にしたことがあると思われるアカウントをフォローした．

本研究において Twitter のデータはすべて Twitter-Api[4] を使用して取得している．今回は 1 アカウントあたり 800 ツイートを取得し，リツイートなども含めた．CBOW では入力を単語に区切る必要があるため，取得したツイートは形態素解析ツール MeCab[5] を用いて形態素解析を行った．Bengio ら [7] によると出現回数が少ない単語はよいベクトル表現を得ることができないため，本実験では出現回数が 10 回未満の単語は取り除いた．また以下のような文字列も取り除いた．

- 句読点
- @から始まる文字列
- http://から始まる文字列
- RT

4.2 実験方法

本提案手法に従って 4.1 節で作成したアカウントの分類結果を被験者に見てもらい以下の質問に 1 から 5 の 5 段階 (5 が最良) で答えてもらい評価とした．

質問 1 分類結果は妥当か

質問 2 各クラスに対してジャンル，内容のまとまりがあると思うか

質問 3 クラス数は妥当か

質問 4 この結果をリストとして使用したいか

また本研究で用いた単語の特徴がリスト分類において有用であることを確かめるために，単語の特徴の代わりに LDA を用いて抽出したトピックベクトルも分類に使用する．そのため本実験では次のような 3 通りの特徴で分類を行う．

- CBOW ベクトルのみを用いる (CBOW)
- トピックベクトルのみを用いる (LDA)
- CBOW ベクトルとトピックベクトルの両方を用いる (CBOW+LDA)

ここで CBOW ベクトルは 120 次元，LDA のトピック数は 500，分類する際の PCA のパラメータは n_components=8 とした．

[3] http://scikit-learn.org/stable/index.html

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

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

4.3 実験結果

実験の結果を図 3 に示す．この結果は被験者 7 名による評価であり，評価値はその平均値である．

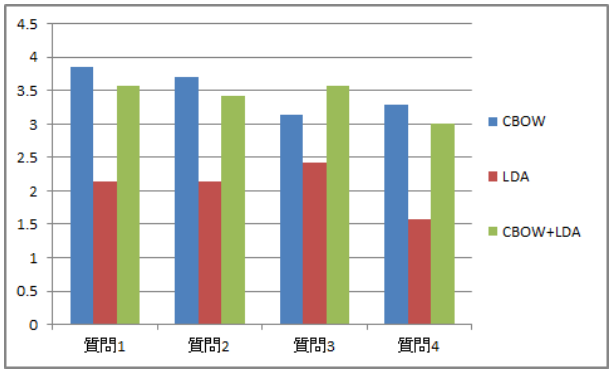

図 3: 被験者による評価

5 考察

図 3 よりすべての質問項目に対し，CBOW のみの評価が LDA のみの評価を上回っている．しかしながら CBOW の質問 4 に対する評価はそこまで高くない，これは CBOW の分類結果の 1 つには政治家やサッカー選手など同じジャンルには属していないようなアカウントが集まっているものがあるからだと考えられる．このような結果になる理由としては，違う内容をツイートしている場合でも例えば地名などを多くつぶやくアカウントは比較的同じクラスに分類されやすいことがある．例えば政治家なら街頭演説を行う場所，サッカー選手なら試合をする場所が挙げられる．今回の実験結果から CBOW を用いたユーザの特徴はリスト分類において LDA を利用したトピックよりも適していることがわかったが，評価値はそこまで高くないため改善法を考える必要がある．また単語の特徴とトピックを同時に使用した場合の分類結果の評価値も上がることはなかったが，トピックを使用した分類結果の評価値を上げることができれば単語の特徴とトピックを同時に使用した場合の評価値も上がると考えられるため，トピックを使用した分類も改善する利点はある．

本研究では Twitter 使用者ごとにリストの数は異なるという仮定のもと，分類アルゴリズムとして AP 法を使用した．しかし AP 法がリスト作成において最良のアルゴリズムであるかは本実験のみでは断定できない．他の分類アルゴリズムも使用した実験を行い評価値を比較することで，有用性を検証していく必要がある．

6 おわりに

本稿ではツイートから単語の特徴を抽出し，ユーザの特徴ベクトルを考えることでユーザをリストに分類する手法を提案した．今回の実験では Twitter 使用者の一人一人アカウントを使用せずに，意図的に作成したアカウントの分類結果の評価をしてもらったため，今後はユーザ個人のアカウントも使用した実験も行い評価の改善を目指したい．また本研究において使用しなかったフォロー関係やプロフィール情報などの特徴も使用し，リスト分類において有用な特徴の組み合わせを検証していくことが評価向上に繋がると考えられる．

7 参考文献

参考文献

[1] Marco Pennacchiotti and Ana-Maria Popescu, A Machine Learning Approach to Twitter User Classification, Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media, pp. 281-288, 2011.

[2] Takuya Yamashita, Haruhiko Sato, Satoshi Oyama and Masahito Kurihara, Classification of Twitter Users Based on Following Relations, Proceedings of the International MultiConference of Engineers and Computer Scientists, 2013.

[3] Shane Bergsma, Mark Dredze, Benjamin Van Durme, Theresa Wilson and David Yarowsky, Broadly Improving User Classification via Communication-Based Name and Location Clustering on Twitter, Proceedings of NAACL-HLT, 2013.

[4] Tomas Mikolov, Kai Chen, Greg Corrado and Jeffrey Dean, Efficient Estimation of Word Representations in Vector Space, http://arxiv.org/abs/1301.3781, 2013.

[5] 奥村 学, マイクロブログマイニングの現在, 電子情報通信学会第 3 回集合知シンポジウム, 2012.

[6] Brendan J. Frey and Delbert Dueck, Clustering by Passing Messages Between Data Points, Science 315, pp. 972-976, 2007.

[7] Yoshua Bengio, Rejean Ducharme, Pascal Vincent, and Christian Jauvin, A Neural Probabilistic Language Model, Journal of Machine Learning Research, 3:1137-1155, 2003.