Vol.2014-MPS-97 No.25
2014/3/4

マイクロブログのインフォーマルな書き込みに対する自動分類

原 正和[†1]　浅井 拓海[†2]　高橋 寛幸[†2]　但馬 康宏[†1]　菊井 玄一郎[†1]

マイクロブログに多数含まれる，公的な情報やメッセージを含まないインフォーマルな書き込みをその意味内容に基づいて自動分類する実験を行った．PMM(Parametric Mixture Model)と SVM(Support Vector Machines)を用いて分類したところ，前者の方が若干 F 値が高いことが分かった．さらに、SVM の結果をベースに PMM で分類カテゴリを追加する方法を試み、それぞれの単体より性能が改善できることが分かった．

Automatic Categorization of Informal Messages in Microblogs

Masakazu HARA[†1]　Takumi ASAI[†2]　Hiroyuki TAKAHASHI[†2]
Yasuhiro TAJIMA[†1]　Genichiro KIKUI[†1]

This work reports our experimental results on automatic categorization of twitter messages. Unlike existing work, we focus on ‘informal messages’ describing user’s mental or physical states and/or very local events. We applied PMM (Parametric Mixture Models), SVM (Support Vector Machines) and a combination of them. Experimental results suggest that the combination outperformed each single method.

1. はじめに

Twitter に代表されるマイクロブログは広く普及しており， 2014 年 1 月時点でユーザ数が 6 億人以上，一日あたりの投稿数は 5 千万件以上と言われている[a].

このような多くのユーザによる大量の書き込みを分析することにより，実世界の状況や人々の嗜好や行動等を推定することが試みられている．これらの試みは大きく２つに分けられる．一つ目は地震の発生や感染症の流行など特定の現象に限定して，それらの状況を推定しようとするものである([1]など)。二つ目は書き込みの集合全体を対象として，話題による分類，構造化を行うことにより，話題分布やその経時的変化を把握したり，ブラウジング支援などに活用したりしようとするものである （[2]など）.

ここで，マイクロブログ（特に twitter）の大きな特徴の一つは，ごく狭いコミュニティでのみ意味のある「インフォーマルな書き込み」が多数なされることにある．これらの書き込みは１つ目の用途には活用されてきたが，２つ目については分類軸の多くが「政治」「経済」などの「ジャンル」であることから，不要なものとみなされてきた.

しかしながら，「インフォーマルな書き込み」には人々の行動や心理状態など，社会の状態を考える上で重要な手がかりが含まれるため，これらの内容の分布や経時的な変化などを明らかにすることは意義深い. この課題に対して，我々は話者の発話行為タイプや発話目的に着目した書き込みの分類体系を提案した[3]. 本稿ではこの分類体系に従ってミクロブログのインフォーマルな書き込みを自動分類するための予備的な検討を行った結果について報告する.

†1 岡山県立大学　Okayama Prefectural University
†2 NTT レゾナント株式会社　NTT Resonant Inc.
a) http://www.statisticbrain.com/twitter-statistics/

2. 分類体系の概略

2.1 分類体系

本研究で用いる分類体系を表１に示す．これはマイクロブログの「インフォーマルな書き込み」を発話行為等に着目して 10 種類に分け，細分化したものである[3]．今回はこのうち「F.自問」,「G.メモ」を除いた８個の大分類を用いる.

表１：インフォーマルな書き込みに対する分類体系

大分類		中分類
記号	名称	
A	心情	感情，評価，願望，意見，推測・憶測
B	話者の状態	身体状態，身体感覚，所在，その他の状況
C	話者の行動・体験	過去，現在，予定
D	外界の状況	事実・他者の行動，伝聞，義務，自然現象,日付・時刻
E	他者への働きかけ	あいさつ，よびかけ，命令･依頼･勧誘,警告･威嚇，質問，感謝，謝罪，拒否，抗議
F	自問 (使用しない)	
G	メモ (使用しない)	
H	引用	
I	返事	
J	その他	

2.2 複数分類の許容

実際に個々の書き込みを分類してみると，複数のカテゴリに所属する，あるいは，それらのいずれとも決め難い場合がしばしばある．文献[3]には２名の作業者で１カテゴリのみを付与した場合のカッパ係数が 0.59 と報告されている．そこで，本研究ではひとつの書き込みが複数カテゴリ

に所属することを許容した（多重トピック）．実際に 300 記事を人手で分類した結果，64 記事(21%)に複数カテゴリが与えられ，その多くは１記事あたり２分類であった．

3. 分類手法

本研究では学習型の分類アルゴリズムを適用する．

3.1 分類のための素性（特徴量）

分類アルゴリズムを適用する素性（特徴）ベクトルの次元は形容動詞, 連体詞, 接続詞以外の単語の出現形であり，各次元の値は当該文書における当該単語の出現頻度である．

3.2 基本的な分類アルゴリズム

基本的な分類手法として PMM (Parametric Mixture Models)[4]と SVM(Support Vector Machine)[5]を用いた．

PMM はナイーブベイズを多重トピックがモデル化できるように拡張したものである．すなわち，カテゴリの組み合わせが与えられた時の文書の事後確率をモデル化し，この確率（尤度）[b]が最大になるようなカテゴリの組を求めるものである．この手法は１つの記事に複数カテゴリを許す本研究の課題に適していると考えられる．なお，今回は PMM のうち, より基本的な PMM1 というモデルを用いる．

SVM は過学習に頑健な手法として知られている. 基本的に２値分類を行うものであるから，今回は J（その他）以外のカテゴリごとに，「その分類に所属するかしないか(one-vs-rest)」を判別する学習を行い（すなわち，カテゴリ数だけのパラメータセットを作り）, 所属すると判定されたカテゴリ全てを出力する．なお，どのカテゴリにも所属しない場合は「J.その他」とした．

3.3 PMM と SVM の組み合わせ

予備実験の結果，F 値で見ると PMM が SVM より高いが適合率は SVM の方が高いことが分かった. そこで, SVM の(J 以外の)結果をまず採用し，PMM によって他のカテゴリを追加する方法を考えた．PMM でカテゴリを推定する(近似)アルゴリズムは，どのカテゴリにも属さない状態から始めて，尤度が最も上昇するようなカテゴリを一つ加える，という操作を尤度が上昇しなくなるまで繰り返す，というものである[4]．SVM の結果を取り入れる場合は，初期状態を「SVM で得られたカテゴリに所属する状態」とし，以降の処理はオリジナルのアルゴリズムと同様とする．

3.4 実験結果と考察

300 記事を評価用データとして 5 分割交差検定によって学習と評価を行い，適合率，再現率，F 値を算出した．それぞれの結果を表２に示す．なお、SVM のソフトマージンのペナルティ係数は予備実験の結果 2.0 とした．

まず PMM と SVM それぞれ単独で見ると, 上述のように, 全体的には前者の方がわずかながら F 値が高いが適合率は SVM が高い．カテゴリ別に見ると E および H の F 値において SVM が PMM1 を超えていて絶対値も高い．理由としては，これらのカテゴリが特定の言語表現と強い依存関係にあり, SVM ではこのことがより直接的にモデル化できたことがあげられる．E に属する記事殆どが挨拶や質問であり, また H に属する記事の殆どに引用記号が含まれている．

PMM と SVM を組み合わせた提案手法については全体の適合率, 再現率, F 値の全てで単体での数値を上回っている．すなわち PMM の再現率を保った状態で SVM の適合率を反映させることができたと言える．但し，データ数は十分とは言えず，より大規模なデータセットでの検証が必要である. またそれぞれの絶対値についても改善の余地は大きい．

表２　実験結果

	PMM			SVM			PMM+SVM		
分類	適合率	再現率	F 値	適合率	再現率	F 値	適合率	再現率	F 値
A	.36	.73	.48	.44	.22	.29	.36	.74	.49
B	.1	.5	.07	.24	.1	.14	.22	.10	.14
C	.39	.44	.41	.47	.27	.35	.38	.50	.43
D	.27	.2	.23	.5	.03	.05	.36	.23	.28
E	.38	.55	.45	.77	.37	.50	.41	.60	.48
H	.59	.73	.65	.93	.65	.77	.63	.87	.73
I	.0	.0	-	.63	.28	.38	.31	.28	.30
J	.0	.0	-	.14	.73	.23	.0	.0	-
合計	.35	.42	.38	.33	.30	.31	.38	.47	.42

4. おわりに

本研究では Twitter のインフォーマルな書き込み内容を分析するための自動分類を試みた．分類の単位は一つの書き込み（1 ツイート）とし, 複数のカテゴリに所属することを許容した．分類アルゴリズムとして SVM の結果を初期値として PMM(PMM1)で他のカテゴリの可能性があれば追加するという手法を新たに試み，PMM,SVM 単体より適合率, 再現率, F 値それぞれで若干の向上が確認できた．

今後はデータを増やして実験を行うとともに，SVM と PMM の組み合わせ方の改良や semi-supervised な方法について検討する．

参考文献

1 T. Sakaki, M. Okazaki and Y. Matsuo:Earthquake shakes Twitter users: real-time event detection by social sensors, WWW Conference (2010),
2 西田 京介, 坂野 遼平, 藤村 考, 星出 高秀: データ圧縮による Twitter のツイート話題分類, DEIM Forum 2.11, A1-6 (2.11)
3 菊井玄一郎：なにをつぶやいているのか？-マイクロブログの機能的分類の試み-, 言語処理学会年次大会, pp.(2012)
4 上田修功, 斉藤和巳：多重トピックテキストの確率モデル-パラメトリック混合モデル-, 電子情報通信学会論文誌 D-Ⅱ Vol.J87-D-Ⅱ No.3 pp872-883 2004-03-01
5 Thorsten Joachims: SVM-Light Support Vector Machine.

b 本来、カテゴリの組の事前確率を掛ける必要があるが、[4]によれば相対的に値が小さいため無視して良いとされている