カーネルSOMによる機能語辞書を用いたマイクロブログ記事における投稿者の意志の発見

濱田翔吾　黒澤義明　目良和也　竹澤寿幸
広島市立大学大学院　情報科学研究科

1. はじめに

近年，ウェブ上で普及しつつあるコミュニケーションツールとして，Twitter に代表されるマイクロブログが挙げられる．

マイクロブログの閲覧は，個人が投稿した記事を誰でも閲覧できるように掲載されており，ユーザがその記事を収集するという方式である．このため，閲覧しているユーザは，常に新しく有益な情報を入手しやすいという利点を持つ．そのため，ユーザ自身がこれから起こす行動の起因となる場合もある．しかし，マイクロブログ上ではリアルタイム性の高い記事，低い記事を含む多種多様な投稿がされている．そのため，ユーザが有益な情報を見逃すという欠点を持っている．そこで，ユーザにとって必要な情報，不必要な情報を分類する必要がある．過去に我々は，ユーザにとって有益な情報を取得する研究を行い，ある程度の分類は行う事はできた．課題として，素性が膨大で高次元であった点．ユーザが有益，必要だと感じる情報はユーザの興味，関心がある部分に委ねられるため，ユーザによって違いが生じる点があった．また逆に，投稿者も，同内容の投稿記事であっても，各々の考えや意志を込める事が多い．その投稿者の意志が，ユーザの興味，関心と合致した時，ユーザは有益，必要であると感じると考えられる．有益な情報取得のためには，投稿者の意志やユーザの興味，関心と言った部分を考慮して分類を行う必要がある．

本研究は，マイクロブログデータを用いて，投稿者の意志を反映させた情報として調査を行い，閲覧するユーザの興味と投稿したユーザの意志が合致するようなツイートを発見することを目指す．ところで，ある膨大なデータに関する分類に関する研究には数多くされている．その中で，大量のデータを使って分類を行う場合，計算量が膨大になる事や，算出された結果の考察が難しい場合がある．そこで，非線形の境界を求め，複雑で高次元なデータを解析，分類する必要がある．そこで本研究では，高次元なデータを解析できる手法の一つであるカーネル SOM を用いて実験を行う．

2. 関連研究

2.1. ユーザの知的欲求による選好に基づいたマイクロブログの記事分類

マイクロブログ上の記事の中からユーザの選好に基づいて有益な記事のみを取り出すことを目的とした研究がある[1]．そもそも，ユーザが有益であると感じる思考は，フォローしたユーザの知的欲求による取捨選択であり，この選好を持ったユーザが欲する情報は名詞や動詞等の内容語だけでなく，モダリティ，すなわち助詞や助動詞を含む文章表現からも区別する必要があると考えられる．内容語だけではなく，ツイートの意図にも注目する必要がある．例えば，アイドル，または特定の個人に興味があるユーザは，どんな服を買ったかとか，どんなことをしているのか，という情報に触れたい．つまり，傍観者としての立場でツイートを読むわけである．このようなユーザの欲求をツイートに含まれる内容語によって分類するのは困難である．

2.2. 言葉が紡ぐデザイン-意思抽出への認知言語学の構成論的アプローチ-

ツイートに関するユーザの「意志」を扱った研究がある[2]．言葉は記述されている内容だけでなく，別の意図を付与されている場合がある．例えば「門を開けてください」という文があったとする．文章中に「命令した」という記述はない．しかし，「命令」という発話行為であることは明ら

かである．この様に，文章中に含まれない行為や，「た」，「ようだ」のような確信さの表現を，この研究の著者は「意志」と呼び，記述された内容から「意志」の部分を抽出する．Twitter においては，似たつぶやきがあっても，末尾の表現の変更に伴い，興味を失うことがあると考えられる．この意志の特徴を抽出出来れば，投稿者の意志を明確化でき，嗜好と合致したユーザに提供ができると考えられる．

3. 提案手法

本研究では，ツイートデータから投稿者の意志を抽出し，後述する属性のグループに含まれる投稿者の意志を抽出し，同じ属性のグループからツイートを発見する研究である．まず，大まかな処理手順を説明し，追って詳細を述べる．

(1) ツイートデータに予め属性を付与
(2) ツイートデータから機能語辞書に基づき，機能語を抽出
(3) 機能語の意味を素性とし，分類

3.1. 属性の定義

本研究では，2.1 節を参考に，投稿者の意志を持つツイートを主に 4 種類の属性に分類することによって表現する．ツイート内容を中心として判断し，記述内容を，主体，対象に分ける．主体が投稿者主体か，投稿者以外か，対象が投稿者主体か，投稿者以外か，に着目して分類を行う．

1 つ目は傍観である．主体が投稿者，対象も投稿者である場合に分類する属性である．主に「～～なう。」というような表現や，主体の予定，行動のツイートがこの属性に含まれる．

2 つ目は追従である．主体が投稿者，対象がその他である場合に分類する属性である．主に，投稿者が何らかの話題の感想や，起きた事柄について述べたツイートが，この属性に含まれる．

3 つ目は先駆である．主体がその他，対象がその他である場合に分類する属性である．宣伝や伝達の記事の場合に分類される．

4 つ目は受身である．主体がその他，対象が投稿者である場合に分類する属性である．今回は対象外とする．

この 4 つの属性を元に分類を行うことで，誰が何もしているのか，何が起っているのかなどが分類される．しかし，投稿者の意志は，内容に記述されない．例えば，特定の個人に対するユーザの内なる情報であるとか，特定の何かに関する感想や意見の情報であるとか，更なる新情報といったものが判断できない．そのため，助詞，助動詞等に着目し，意志を抽出する必要がある．

3.2. 日本語機能表現辞書「つつじ」

日本語の文を構成する要素には，名詞，動詞のような，主に内容的な意味を表す要素(内容語)以外に，助詞や助動詞といった，主に文の構成に関わる要素がある．総称して「機能語」と呼ぶ．機能語辞書「つつじ」[3]には，機能語毎に，直前または直後に接続する品詞一覧や，「伝聞」，「推量」等の意味カテゴリーが存在する．接続する品詞と一緒に機能語を調査し，該当する語であれば，その機能語が持つ意味カテゴリーを利用し，ツイートから機能語とその意味を抽出する．本研究では，投稿者の意志を抽出するために機能語を用いる．

3.3. 素性の作成

ツイートの属性を付与した後，学習データの素性とするため，ツイート毎に機能語辞書を用いて総当りで調査をする．該当する言葉が存在した場合，機能語辞書にある接続品詞を確認する．必ずしも該当する言葉が機能語とは限らないためである．接続品詞を確認する際，形態素解析を行う．形態素解析には MeCab[1] を使用する．MeCab により，直前の語の品詞が辞書にある接続品詞に含まれていた場合は，機能語の意味カテゴリーを調べ，その意味を素性として作成する．

4. 実験

実験は 2 種類行い，3 章の手法を用いて行う．実験 1 は，3 文字以上の機能語を使用した実験で 3 章の手法を用いて行う．実験 2 では，2 文字以上の機能語を用いた実験で，3 章の手法を適用す

[1] MeCab，http://mecab.sourceforge.net/.

る．機能語を含むツイートデータ 660 件を用い，学習に関しては，後述するカーネル SOM(Self-Organizing Map)を用いる．カーネル関数はガウシアンカーネルを用いて，マップサイズ 26×26，学習回数は 1000 回(ただし，計算量の関係上評価の調査は 100 回まで)．また，学習の際，tfidf 法により重みをつける実験も行う．

4.1. カーネル SOM を用いた実験

カーネルSOM[4]は，カーネル法をSOMに適用し，データの特性に適した写像を行うことで，分類やクラスタリングの精度向上が見込める．

カーネル SOM では，勝者ノードの決定のために以下の非類似度 d_{lk} を随時更新していく

$$d_{lk}(t+1) = (1-\alpha)d_{lk}(t) - \alpha(a-\alpha)d_{lh}(t) + \alpha(K_{kk} - 2K_{kh} + K_{hh}) \quad \text{4-1}$$

4.2. 評価

本研究では，カーネル SOM を用いて，明確な境界線により分類が為される事を期待する．しかし，カーネル法を用いて高次元写像しただけでは容易に境界線を描いて分類できるデータではない．そのため，本研究では，「マップ上で同属性のデータが隣接しているかどうか」を評価指標とする．評価指標は二つ用意する．

1. あるノードのデータを A とする．隣接ノードの所持しているデータに存在する A の数
2. あるノードのデータを A とする．隣接ノードの所持しているデータの A の有無

4.3. 実験結果

得られた実験結果を以下，表 4.1 に，得られたカーネル SOM のマップの一部を図 4.1 に示す．ちなみに図 4.1 に関する対応表は以下の通り．

表 4.1：実験 2 における結果

実験 2	学習回数	値
重みなし 評価 1	26	0.215933
重みなし 評価 2	50	0.725341
重みあり 評価 1	73	0.237557
重みあり 評価 2	44	0.738998

図 4.1：実験結果 SOM マップの一部

表 4.2：対応表

伝 1	傍観
伝 2	追従
伝 3	先駆
伝 4	受身
呟	その他

5. 考察

紙上の都合により，実験 2 の考察のみ述べる．傍観，追従，先駆で得られた結果の一部を示す．

表 5.1：傍観に関するグループ

感嘆 否定 理由	学生書発見したー！よかったー．ゆーて，学部のカードで研究室とかは入れるからそんなに困らなかったけどね．
感嘆 否定	今度治らなかったら、原付に乗り換えだな。
感嘆 継続	大学に行ってきます．いつもより 2 時間近く早いな．

表 5.1 より，感嘆が含まれている．これは，投稿者の行動を述べた後，その内容に関する気持ちを述べる場合に含まれる．この情報は，ある人が興味を持った事実や行動に興味があるユーザ(例えば先述した，アイドルのファンや特定の個人に興味のあるユーザ等)にとってはこのようなツイートこそが必要となる．

表 5.2：追従に関するグループ

推量 不−可能 疑問 仮定	前ネット時代なら足並み揃えて無視することもできたでしょうが。
推量	シリコンといえばなんでウエハーの断面図って丸いんでしょう
推量 否定	轆轤は回さなかったけど、大事な一日だったように思う

表 5.2 より，追従に関して，推量の素性を含むグループが生成された．「～は～だろう」といった，事柄に対する展望について述べるためである．この情報は，例えばユーザが何らかの意見や感想について興味があるユーザにとって，事柄に対する，意見，予測等を参考にすることが出来るため，必要となる．

表 5.3：先駆に関するグループ

依頼 仮定 継続 理由	国際学部棟の 3F でフリマしてますので良ければ見に来て下さい！！
勧め 仮定 推量	EC ナビ社、社内にお洒落バーがあって社員はタダで利用できるらしい……受ければよかった

表 5.3 より，先駆に関して，依頼や勧めを含むグループが生成された．どちらも他の情報を示し，情報に関する自分の意志を表現しているためである事が分かる．この情報は，新しい情報として発信，または伝播しているため，新情報を欲しいユーザにとっては必要となる．

6. まとめ

今回，マイクロブログに投稿されている記事を学習させ，予め定義した属性である傍観，追従，先駆，受身の 4 種類に分類を行うことができる特徴を発見するための提案を行った．提案手法では，ツイートに属性を付与する．属性は主に 4 種類．人手でツイートデータ 4285 件に属性付与を行った．学習時に，日本語機能表現辞書「つつじ」を利用し，ツイートに含まれる機能語を素性とした．学習方法にはカーネル SOM を使用した．

提案手法の有効性を調べるために，前述したツイートデータを使用し，実験を大きく 2 種類行った．生成されたマップから 2 種類の評価を用いて値の一番高いマップを元に考察を行った．

実験結果を総合すると，傍観のツイートには付帯，確定，意志，仮定の意味を持つツイートが特徴としてある．追従のツイートには，推量等の意味を持つツイートが特徴として存在する．先駆のツイートには，依頼，勧め，伝聞の意味を持つツイートが特徴として存在することが分かった．

7. 今後の課題

今後の課題として，データ数の少な過ぎることがあげられる．今回，機能語が含まれるツイート 660 件で実験を行なっている．しかし，各属性のツイート数が少なすぎたため，それぞれの違いが判断できなかったと考えられる．よって今後，ツイートデータを増やして実験を行いたい．

実験では今回機能語のみを使用した．内容語と合わせる事によって，更なる分類精度向上が見込めると期待する．しかし，特定の機能語のみに重みをつける事前処理が必要であると考える．

比較実験で通常の SOM でも実験を行なって行きたいと考えている．今回実験したカーネル SOM と SOM における分類精度の違いなどを調べたいと考えている．

謝辞

この研究の一部は，平成 24 年度広島市立大学特定研究費(一般研究)の補助を得ている．関係各位に感謝申し上げる．

参考文献

[1] 濱田翔吾，黒澤義明，目良和也，竹澤寿幸，“ユーザの知的欲求による選好に基づいたマイクロブログの記事分類，”情報処理学会 自然言語処理研究会，NL-204，2011.

[2] 宇野良子，橋本康弘，岡瑞起，李明喜，荒牧英治，“言葉が紡ぐデザイン―意志抽出への認知言語学の構成論的アプローチ，”Cognitive Studies，17(3)，491-498，2010.

[3] 松吉俊，佐藤理史，宇津呂武仁，“日本語機能表現辞書の編纂，”自然言語処理，Vol.14，No.5，pp123-146，2007.

[4] 井口亮，宮本定明，”カーネル関数を利用した LVQ クラスタリングと SOM，”知能と情報(日本知能情報ファジィ学会誌)，Vol.17，No.1，pp.84-94，2005.