Vol.2011-NL-204 No.7
2011/11/21

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

濱田翔吾† 黒澤義明† 目良和也† 竹澤寿幸†

マイクロブログは，情報の伝播が速い点や，投稿が容易である点で，ユーザ数が急増している．しかし，その伝播の速さと容易さによって大量の記事が氾濫するため，ユーザが有益な情報を見逃すという問題が起きている．この問題を避けるために，ユーザの選好に合わせて記事を提示する必要がある．本論文では，必要な記事と不必要な記事を分割する手法を提案する．特にユーザの選好は，傍観者，追従者，先駆者の 3 種類あると仮定し，それぞれに記事を分類することを試みる．その際，助詞や助動詞に含まれている，ユーザの意志に焦点を当てる．上記の手法を元に実験を行った結果，意志の条件である，助詞や助動詞に重みをつけた場合の方が，重みを付けなかった場合に比べ，F 値が 6%程度上昇した．したがって，ユーザの選好に基づいた記事の分類に有効な手法であると言える．

Classifying micro-blog posts based on user preference motivated by an intellectual desire

Syougo HAMADA†, Yoshiaki KUROSAWA†, Kazuya MERA† and Toshiyuki TAKEZAWA†

Micro-blog users are increasing at a rapid pace because of easier procedure of posting and faster information transfer. However, a massive flood of posts made by the ease and speed becomes to result in the fault that the users miss important ones. For avoiding this issue the posts need to be selected according to their preferences. In this paper, we propose a technique to distinguish necessary posts from unnecessary ones. In particular, we assume three types of the users' preference; spectator type, follower type, and pioneer type. Our aim is to classify various posts into these three classes. When classifying them, we also focus on particles and auxiliary verbs to capture so-called users' intention including them. We performed an experiment based on the idea mentioned above. The condition with intention, i.e., using some weights to particles and auxiliary verbs, achieve approximately 6% lager F-mesure compared with the one without using them. Thus, our technique had an effect on the classification of posts as a basis for selecting based on user preference.

† 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*1に代表されるマイクロブログが挙げられる．マイクロブログは投稿が容易であるため，一日に何度でも投稿が可能である．また，ユーザの投稿した記事はオープンであるため，どのユーザからも見ることが出来，興味のある投稿をするユーザをフォローする事が出来る．フォローとは，自分のページ上にフォローしたユーザの記事を表示させるための登録である．フォローするユーザ数が多いと，自分のページ上に他のユーザの投稿記事が大量に表示される．そのため，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．その反面，リアルタイム性の高い記事，低い記事や，情報の伝達を目的としていない生活記録，独り言など，多種多様な発言が多く投稿されている．したがって，Twitter などのマイクロブログでは，ユーザが各々にとって有益な情報を見逃すという欠点を持つに至っている．

確かに，自分と同等の興味を持っているユーザや、自分の興味のある企業やメディアのアカウント等はユーザにとって有益な記事を投稿している可能性が高い．しかし，フォローしているユーザの記事の中にも必要な情報と不必要な情報が存在する．こうした必要／不必要の基準は，ユーザの嗜好あるいは選好に基づく．特定のドメインに関心がある場合には，そのドメインの記事を求めるはずである．また，いわゆる新しい物好きのユーザは，様々な分野の流行に感心を持ち，様々なドメインの記事も求めるはずである．これら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．

そこで，本研究では，Twitter 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する．例えば，あるユーザは様々な分野の流行に関する記事を閲覧したいかもしれない．別のユーザは生活記録のようなツイートを閲覧したいかもしれない．この様に，一人のユーザに対する他のユーザの選好に応じて記事を提示可能なシステムの構築を目指す．

2. マイクロブログサービス

本研究の目的は，閲覧するユーザにとって有益な情報をできるだけ見逃さずに取得することにある．通常のブログでも，有益な情報は取得できる．しかし，情報の取得から発信までに時間がかかる上，ユーザが自ら検索をしなければ，ユーザにとって有益な情報は取得できない．対して，マイクロブログでは，ユーザが興味を持ったユーザをフォローすることで，リアルタイムにフォローしたユーザのツイートを閲覧する

*1 Twitter，http://twitter.com/.

ことが出来るため，情報収集が容易になる．これらの点に関してマイクロブログサービスと通常のブログとの違いがある．よって本章では，マイクロブログの特徴として，リアルタイム性，フォローの 2 つについて述べる．なお，本研究はマイクロブログサービスの一つである Twitter を例に述べる．

2.1 リアルタイム性

Twitter で投稿する（ツイートという）場合，一日多数回の更新が一般的である．また，携帯電話等の端末を使用しての投稿が容易に行える．この特徴を生かして，イベントやテレビ番組の内容を逐一更新するようなツイートも行われる．このリアルタイム性によって，情報の入手，もしくは発信を短時間に行うことができる．従来のブログでも，ユーザがある話題に関して情報を入手し，またその情報を発信することは出来る．しかし，ブログは一般的に長文で書き込まれるため，投稿までに時間がかかる．そのため，異なるユーザにより同じ内容が書き込まれる場合において，書き込まれる時間に差異が生じる場合があった．よってブログと Twitter との比較では，マイクロブログの方がブログよりもリアルタイム性に長けているといえる．

2.2 フォロー

Twitter には，フォローという手続きが存在する．この関係を，図 2.1 より説明する．

まず，この図は，ユーザ A のページである．投稿 a はユーザ A が投稿したツイートで，続けて投稿 a の内容を示している．ユーザ A の友人や，ユーザ A が興味のある他のユーザに対し，フォローをする．すると，投稿 a とフォローしたユーザのツイートが，タイムラインと呼ばれる領域にフォローしたユーザの投稿が併せて表示されるシステムである．

タイムラインとは，ユーザが，フォローしているユーザと自分自身の投稿を時系列順に一覧表示する領域のことである．図 2.1 では，ユーザ A は，ユーザ B，C，D，E，F をフォローしている．このため，ユーザ A のタイムラインには，自分自身を含む A～F の全ユーザのツイートが表示されている．

このフォローという手続きが Twitter の大きな特徴であり，フォロー数が多いほど，タイムライン上に短時間で多数のツイートが表示されるため，大量の情報を入手できるものの，短時間に大量のツイートが表示されてしまうため，重要な情報を見逃す可能性がある．また，ユーザの嗜好や選好によって，フォローしたユーザの記事を読む際にも，有益な情報とそうでない情報が含まれているため，閲覧するユーザ自身の嗜好や選好によって，提示するツイートを変更できるよう，ユーザの属性を定義する必要がある．

本研究では，閲覧するユーザ自身の選好を考慮した，有益な情報を予め学習した上で，有益な情報とそうでない情報を自動的に分類する手法を提案する．そのため，次章では，本研究の手法に関連している研究を述べる．

図 2.1：マイクロブログサービス(Twitter)

3. 関連研究

本章では本研究に関連する主な研究を述べる．3.1 節では，マイクロブログに関して，ツイートを話題ごとの繋がりを抽出する研究である．3.2 節では，マイクロブログに関して，ユーザの嗜好や選好に着目し，ユーザの投稿のタイミングに合わせて，ユーザの嗜好にあった情報推薦を行う研究である．3.3 節では，本研究と同様の考えを持った，言語表現に着目した研究であり，言葉の中に含まれる「意志」を取り出す研究を行なっている．

3.1 文脈を考慮したリアルタイムなツイート繋がりの抽出・提示システム

この研究はツイートのタイムスタンプや，ツイート内の単語を抽出して，単語や時系列の類似度を算出し，類似度からツイートの話題ごとにクラスタ化する研究である[1]．災害や，ニュース速報，サッカーの試合やテレビ番組の放映などのような，「時間」や「期間」に関連した内容のツイートを，フォローしているユーザがほぼ同じタイミングで，投稿した場合，同じ内容の記事が大量に短時間にツイートされる場合がある．この研究では，ツイートと語彙を使用し，単語に関する類似度，時系列的な近さに基づいたツイート類似度，一定間隔にツイートされる時系列的な類似度を算出す

る．算出した類似度から話題ごとのクラスタにまとめる．話題別にツイートを整理できる利点を持つ.このため,ユーザの嗜好に反映した提示が可能であると考えられる．

3.2 マイクロブログの分析に基づくユーザの嗜好とタイミングを考慮した情報推薦手法の提案

本研究と同様に，ユーザの嗜好や選好に着目した研究がある[2]．この研究は，マイクロブログ特有のリアルタイムな投稿を活用し，過去の投稿によるユーザの嗜好分析によって，実際の商品データを用いて効果的な情報推薦を行う．例えば，好きなプロ野球チームが優勝した時点のツイートのタイミングに合わせ，祝杯のビールを推薦する．という情報推薦などがある．このため，ユーザの嗜好に応じた情報推薦，提示を行う事が出来ると考えられる．

これらの研究は，ユーザにとって有益な情報を提示するために，内容語に着目した上で,学習を行っている.その他にも,近年行われているマイクロブログの研究では，ユーザ同士の近接度に着目して，有用なツイートを探索する研究[3]や，マイクロブログにおけるユーザに着目した有用情報の分類手法の提案[4]，マイクロブログの返信行動に着目した投稿やユーザの分類[5]，表記揺れの緩和を行った上でツイートの集合をもとに，クラスタリングを行う研究もある[6]．これらの研究全て，ツイートの内容の内容語のみを情報としている．

しかし，マイクロブログでは，フォローしたユーザのツイート全てに興味をもつわけではない．閲覧するユーザによって，興味のあるフォローユーザの記事は異なると考えられる．このため，内容語だけでなく，別の要素に着目する必要がある．

3.3 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[7]．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．

このことから，3.1 節，3.2 節で述べた，内容語だけなく別の要素として 3.3 節で述べた「意志」を使う必要がある．「意志」は，ユーザが望ましいと感じる表現であり，そのユーザの属性とも言える．聞き手を意識しない「意志」の言明を収集することによって，ユーザの選好と「意志」が合致した情報を提示する．なお，3.2 節の研究では，ユーザの購入動機を決める目的のため，短期的な変化が予想される嗜好を扱っている．本研究では，より長期に渡って安定すると考えられる選好を扱う．

4. 提案手法

前項に上げたように，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．本研究で分類する有益な情報は,「投稿者が相手に意味のある情報を伝える内容が書かれているツイート」である．しかし，受け取るユーザにとって有益でないと感じるツイートも存在する．これが前述した選好である．したがって，その選好の違いを学習させるために，内容語だけでなく，ツイートの意図に関する情報を，助詞や助動詞といった文章表現を学習，分類させることによって，ユーザの選好に合わせたツイートの分類を行う研究である．

本研究では，選好における閲覧するユーザの属性を定義している．この点に関しては，後述する選好の定義で明確に述べる．まず，本研究の流れを説明する．

(1) Twitter からツイートのデータを取得
(2) 人手でツイートに属性を付与
(3) ツイートを形態素解析し，形態素ごとに素性を作成する
(4) 素性の品詞情報による学習データの重み付け
(5) N 分割交差検定を行い，精度，再現率，F 値を算出

本研究で提案する手法は 2 つある．一つは有益な情報とそれ以外のツイートの 2 値分類とし，(2)の属性の定義付けを行い属性の付与を行う手法．もう一つは，(3),(4)の，ツイートデータの形態素ごとに素性を作成した上で，ツイート内容に応じて新たな素性や，素性の値の調整を行う手法である．この 2 点に関して次節に述べる．

4.1 属性の付与

本研究では，選好のツイートとそれ以外のツイートに分類する研究であるため，選好かそれ以外かのツイートを学習させるために，正解データを作成する．正解データ作成は，ツイートの内容に応じて属性を人手で付与する．属性の種類を以下に示す．

- 選好……誰かにとって有益と感じるツイート
- その他……それ以外のツイート

また，Twitter の機能である，Retweet 機能を使用している場合，投稿者が，自分のフォロワーに知ってもらおうとする意図があるとみなす．よって，いかなる内容でも Retweet 機能を使用している場合は選好とする．また，返信（リプライ）機能を使用している場合は，返信者を対象とした会話であるとみなすので，内容がいかなる情報でも今回の実験から省いている．

しかし，属性の付与は人手で行うため，属性を付与する人によって違いが生じる可能性がある．選好の定義を人手で付与する際に，定義を明確にするため，選好をさらに 4 種類の属性に分類する．

4.1.1 選好の定義

本研究では，ツイート内容が選好かそうでないかを分類するため，選好の定義は非常に重要な位置づけとなる．本研究の行う学習は予め正解を与えるため，一つ一つのツイートに人手で選好か，その他かの属性を付与しなければならない．この場合，人によって，選好かその他かの判断は異なることが予想される．このため，予め人手で属性を付与する際，できるだけ明確に属性の定義付けを行わなければならない．今回重要であるのは，選好の定義である．選好の定義は，ツイートの内容において，記述主体と記述対象が投稿者自身か，投稿者以外かによって分類を行う．4 種類の定義をまとめた表を表 4.1 に示す．主体，対象の考え方は，「意志」を扱う研究にも扱われている[1]．本研究では，この考え方を拡張し，「意志」の記述内容が，投稿したユーザの事を指しているのか，その他を指しているのかで判断する．

表 4.1：選好 4 種類の定義

	記述主体	記述対象
傍観（S）	投稿者	投稿者
追従（F）	投稿者	その他
先駆（P）	その他	その他
受身	その他	投稿者

表 4.1 より，S は Spectator（傍観者），F は Follower（追従者），P は Pioneer（先駆者）を表している．ここで，投稿したユーザと閲覧するユーザを区別するため，それぞれの属性のツイートを ST(Spectator Type)，FT，PT と定義する．または，それぞれの属性を好んで閲覧するユーザのことを，STU(Spectator Type User),FTU,PTU と定義する

記述主体と記述対象について，本研究のイメージを図 4.1 に示す．

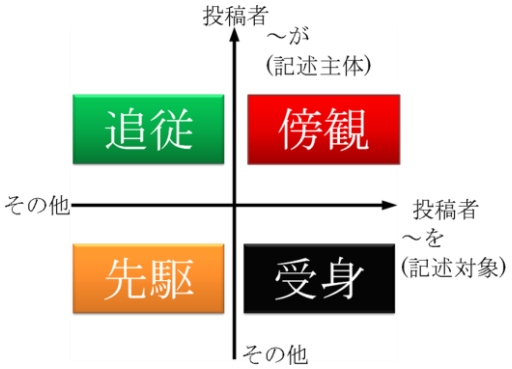

図 4.1：選好の定義

4.1.2 傍観(Spectator Type)

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，自分自身の予定，行動のツイートがこの属性に含まれる．

文章例

「渋谷なう。」

—投稿者が「渋谷にいる自分」についてツイートをしている．

「早く休憩したい」

—投稿者が休憩を希望する旨をツイートしている．

ユーザの選好がこの属性にある場合、そのユーザは STU である．有名人や芸能人や友人の行動を知りたいユーザにはこの属性を有するツイート（ST）を提供するべきであると考える．

4.1.3 追従(Follower Type)

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．

文章例

「Twitter 結構面白いよね」

—投稿者が Twitter（その他）の感想をツイートしている．

「今朝登校中に，事故現場に遭遇した」

—投稿者が登校中に事故現場に遭遇したこと（その他）をツイートしている．

ユーザの選好が、この属性のツイートにある場合、そのユーザは FTU である．他のユーザの意見、感想を重視するユーザにはこの属性を有するツイート（FT）を提供すべきであると考える．

4.1.4 先駆(Pioneer Type)

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．
宣伝や伝達の記事である場合に分類する属性である．
文章例
「今日，駅前でイベントがあるらしいよ」
—イベント（その他）が駅前で（その他）ある旨をツイートしている．
「○月×日，△△が発売するんだって！」
—△△（その他）が△△を売る場所（その他）で発売される旨をツイートしている．
ユーザの選好が、この属性のツイートにある場合、そのユーザは PTU である．会社やメディアの公式アカウントはこのツイートをする場合が多い．新情報や新製品を好む人は、この属性を有するツイート（PT）を提供すべきと考える．

4.1.5 受身

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．
投稿者が何らかの影響を受ける，被害を受けるという記事はこの属性に含まれる．
文章例
「さっき車に轢かれそうになった！」
—投稿者が車（その他）に轢かれそうになったことをツイートしている．
「店員に笑われたー…何故？」
—投稿者が店員（その他）に笑われたことをツイートしている．
なお，この属性はツイートを調べた際数が少なかったため，本研究ではこの属性を省いている．

上記の定義より,マイクロブログでは,フォローしたユーザの記事を閲覧する場合，基本的に傍観者（STU），追従者（FTU），先駆者（PTU）の 3 つの属性を閲覧するユーザが持っていると考えられる．有名人や芸能人のユーザの記事を閲覧する場合は，フォローしたユーザの生活等を現した傍観ツイート（ST）を好み，追従者（FTU）はフォローしたユーザの意見や感想の追従ツイート（FT）を好んで閲覧する．そして先駆者（PTU）は，フォローしたユーザの新情報の先駆ツイート（PT）を好む．この 3 つの特徴は，密接につながりを持つ場合があり，メディア等から情報を入手した PTU は，その情報を元に行動した結果を FT として投稿し，FT を収集した FTU が行動を起こす．その行動を含め，日常を行う行動である ST を閲覧するのを好むのが STU，という流れとなる．この様に，STU，FTU，PTU の記事の閲覧の流れが円滑になることで，ユーザはストレスなくユーザにとって有益な情報を得ることが容易になると考えられる．

しかし，ST，FT，PT の 3 種類それぞれの分類は，内容語のみでは難しい．例えば，
A「このゲーム，難しいよ！」
B「このゲーム，難しいらしいよ！」
という 2 文があったとする．A は，「このゲーム」が「難しい」事を「自分の意見や感想として」述べている事がわかる．しかし，B は「このゲーム」が「難しい」事を「相手に新情報を伝える」旨を述べている事が分かる．前述した 3 種類の属性に分類すると，A は FT，B は PT に分類される．しかし既存研究と同様に，名詞，動詞等の内容語のみで学習させると，A，B の違いである「らしい」は取得できないため，A，B は正しく分類されないと考えられる．このため本研究では，助詞や助動詞と言った表現に関する品詞も使用する．

このように定義を明確化し，ツイートに投稿する際，記述主体と記述対象によって，選好を 4 つに分割する．すると，人手で属性をつける際に，人によって付ける属性に差が出にくくなり，機械学習する際，それぞれのツイートの学習で共通点を発見しやすくなり，分類する際に，正答に分類されやすくなると考える．

4.2 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，ツイート毎に形態素解析を行う．形態素解析には MeCab*2を使用する．URL や，Retweet を示す RT の記述， Twitter 特有の表記が存在した場合，MeCab のシステム辞書には登録されていない単語（未知語とする）と判断され，形態素解析が正しく行われないという点．また，Twitter 特有の表記は，選好かそれ以外かを分類するために重要なデータであるため，Twitter 特有の表記においても学習の素性として使用する点．2 点の理由から，他の表記と重ならない特別な表記に変更して学習データの作成を行う．また，公式 RT には RT という文字列が存在していない．しかし，学習の都合上，通常のツイートと公式 RT を区別しなければならないため，予め公式 RT に関してはツイートの文頭に RT の文字列を付与しておく．また，MeCab での形態素結果が一部例外（そうだ，ようだ，らしい，のような伝達の表現に使われる言葉）を除く，非自立，接尾，接頭のような，単独では意味を持たない表現は処理の対象としない．

Twitter は，140 字の字数制限のため，「眠い」を「ねむ」と表記するような言葉の省略を使用する場合が多い．また，フォロワーとのツイートでの会話の場合等，くだけた表現を使用する場合も多い．さらに，Twitter 特有の表現，「～なう」（～している），

*2 MeCab，http://mecab.sourceforge.net/.

「～だん」(～した）なども存在する．このような表現は前述した，未知語と判断される．したがって，MeCab では正しく解析されない．そこで，Twitter の特有の表現や，一部固有名詞を含んだユーザ辞書[8]や，wikipedia のキーワードを登録した辞書を使用し，MeCab の形態素解析誤りの緩和を試みる．それに加えて，Twitter には「言う」と「いう」のような，表記揺れが多数存在している．このような表記揺れに対処するために，MeCab による形態素解析後に，JUMAN*3を用いて再度解析を行う．形態素解析結果に対して，代表表記が付与されている．前述したような，「言う」や「いう」のように同じ語で違う表記でも，代表表記としては「言う/いう」に統一される．MeCab，JUMAN の解析結果例を表 4.2，表 4.3 に示す．また JUMAN の解析結果に含まれるドメインと呼ばれる機能も素性として使用する．

このように解析された形態素毎に ID 番号を割り振り，単語 ID とする．そして，その ID 番号とその形態素の出現数を一組として 1 ツイート毎に学習データを作成する．学習データの例を表 4.4 に示す．なお，MeCab の解析した結果が未知語であると判断した場合でも，学習データとして登録は行うこととする．

本研究は，ツイートの内容を元に分類を試みている．しかしながら，ツイート内容のみだと，短文のツイート，つまり形態素数の少ないデータでは，適切な分類が行われない可能性がある.つまり,そのデータは分類するための情報が少ないからである．その誤分類を防ぐために，ツイート内容だけでなく，そのツイート内容に関連した素性も作成する．素性を増やすことによって，形態素数の少ないデータでも，分類するためのデータを増やすことで，正しく分類しやすくする．新しく生成する素性は 3 種類あり， 1 つ目は Twitter 特有表記である，Retweet や URL 等を素性とする．2 つ目は伝達に使われやすい言葉を素性とする．この 2 つを述べる．

表 4.2：MeCab の解析結果例

Mecab	
らしい	助動詞,*,*,*,形容詞・イ段,基本形,らしい,ラシイ,ラシイ

表 4.3：JUMAN の解析結果例

Juman	
らしい	らしい らしい 助動詞 5 * 0 イ形容詞イ段 19 基本形 2 NIL

*3 JUMAN，http://nlp.kuee.kyoto-u.ac.jp/nl-resource/juman.html/.

表 4.4：学習データ例

属性	単語ID	値（単語数）	単語ID	値（単語数）		単語ID	値（単語数）
選好	0	1	1	1	・・・	13	1
その他	14	1	15	2	・・・	20	1

4.2.1 Twitter 特有表記

Twitter には特有の表記が存在する. 2 章で述べた, Retweet の表記などを素性とする．また，Retweet に関しては，4.1 節で述べた通り，Retweet に関しては選好として扱うので，選好の重要な素性となる．よって，Retweet を示す表記は素性として加える．

4.2.2 伝達表現のラベル

伝達特有の表現をラベルとして付与する方法である．例えば，何らかの情報を伝える場合「～～らしい」や「～～だって」というような表現を使う場合がある．このような表現に,「L=伝達」というラベルを新たに作成し単語 ID を設定する. 例えば表 4.5 のような学習データがあったとする．この場合,「～だって」と「～らしい」という単語 ID は別であり，値も 1 ツイート中のそれぞれの単語の数になる．しかし，表 4.6 のようにラベル付与後の単語 ID は同一になり，値もラベルの数＋重み付けの値となるので，伝達表現として学習がしやすくなり，この表現が存在している場合に選好であると判断されやすくなると考える．値の重み付けに関しては次節で説明する．

表 4.5：ラベル付与前の学習データ例

単語	単語ID	値
だって	0	「だって」の数
らしい	1	「らしい」の数

表 4.6：ラベル付与後の学習データ例

単語	単語ID	値
L＝伝達	2	「L＝伝達」の数＋重み付け
L＝伝達	2	「L＝伝達」の数＋重み付け

4.3 学習データの重み付け

4.2 節で述べたような形式で学習データを生成しても，それぞれのデータの値は殆ど 1 や 2 程度となるため，1 ツイートの形態素数の違いで，学習に差が出る可能性がある．その可能性を減らすために，学習データに重み付けを行う． 重み付けを行うのは前述した, 4.2.1 節 4.2.2 節の 2 種類である. 重み付けに関しては, 4.2.2 節で述べた.

このラベルは伝達に関する表現をラベルとして新しく素性を生成するので，選好のツイートを分類する重要な素性となる．重み付けに関しては，表 4.6 のようにラベルの数＋重み付けの値とする．重み付けの値は(1)における値とする．

$$\frac{\text{全形態素数}}{\text{全ツイート数}} = \frac{57827}{3506} = 16.4934... \fallingdotseq 17 \quad (1)$$

この章では，本研究の提案手法についての説明を行った．特に，4.1.1 節の選好の定義は，属性の付与の際，人によって与える属性に違いが出ることを防ぐ点，選好かどうかを決める上で明確な指標となる点で重要である．また，4.2 節での新しい素性の作成では，選好のツイートに関わりのある表記，または，その他のツイートに関わりのある表記を明確に表す上で重要である．4.3 節での学習データの重み付けは，選好の属性となりやすい表現に，学習の重みをつけることは，選好のツイートをより正答しやすくするために重要である．学習データの重み付けによる効果については，次章で詳しく検討する．

5. 実験

本研究では，4 章で述べた手法を用いて実験を 2 つ行う．

実験 1 では，まず，選好のツイートと，それ以外のツイート全ての中から，どれだけ選好のツイートを正しく分類出来るかを調べる．

実験 2 では，選好のツイートのみのデータから，ST，FT，PT の 3 種類それぞれが正しく分類できるかを調べる．

まず，実験に使用するデータと，評価尺度を述べる．実験に使用するデータは実験 1 と実験 2 で異なる．評価尺度は実験 1,2 共に同じ評価尺度を用いる．実験に使用するデータは，Twitter から取得した本学の関係者と思われるユーザのツイートデータである．人手で付与した属性毎に表したツイート数を表 5.1，表 5.2 に示す．実験 1 では，学習を行う際，ST,FT,PT を 1 つの属性にまとめて，選好とする．

表 5.1 実験 1 における属性ごとのデータセット

属性	ツイート数
選好	2,132
その他	1,374
合計	3,506

表 5.2 実験 2 における属性ごとのデータセット

属性	ツイート数
傍観(ST)	460
追従(FT)	441
先駆(PT)	519
合計	1,418

次に，精度，再現率と，F 値の評価尺度である．システムが正解と認識したデータを S，人手で付与した正解データを H，システムと人手の判定が一致した正解データを C とする．評価尺度を示す計算式を表 5.3 に示す．

表 5.3：評価尺度

精度	再現率	F 値
$\frac{C}{S}$	$\frac{C}{H}$	$\frac{C}{\frac{1}{2}(S+H)}$

ツイートデータを形態素解析して生成した素性数の内訳を表 5.4 に示す．表より，ラベルとは，伝達に関連する表現，言葉に関して付与したものである．ドメインとは，JUMAN による形態素解析結果から登録された，ドメインを表している．

表 5.4：素性数の内訳

素性	数 (個)
自立語等	6938
ドメイン	86
ラベル	6
合計	7030

4 章で述べた提案手法より，学習データに重み付けを行う時，選好に出現しやすい品詞を見つけ，該当する品詞の単語の値に重み付けをする．学習方法の違いによる結果の差異と提案手法の有効性の確認を行う．

5.1 学習方法

生成したデータを使用し分類器で機械学習を行う．学習方法には NaiveBayes 法と，SVM(Support Vector Machine)を使用する．なお，Twitter のような，140 字の字数制限が設けられていると，記述主体や記述対象が省略されている場合が多い．また，日本

語はそもそも省略の多い言語である．このため，省略が行われた場合はツイートの内容で，どの選好に属するかを考えることとする．

NaiveBayes 法は，学習したデータの値の組み合わせや，頻度から，分類する属性の確率を求める学習方法である．SVM 法は，2 種類の属性のいずれかに分類する機械学習法の一種で，与えられた学習データで生成された点において，属性の境界近傍にあるそれぞれの点と，属性を識別する識別面と呼ばれる距離のマージンを最大化するように分離させる．本研究では broomie*4と TinySVM*5を使用する．

5.2 実験１

本節では，4 章で述べた提案手法より，実験を行う．このデータはツイートデータに関して，形態素解析を行った結果の形態素数がそのまま素性となる．5.1 節で述べた実験方法で実験を行う．実験方法は 5 分割交差検定を行う．

5.2.1 実験結果

実験 1 の実験結果を NaiveBayes 法，SVM 法合わせて，表 5.5 に示す．

表 5.5：実験結果

		精度	再現率	F値
with juman	NB	71.1%	75.4%	73.1%
	SVM	82.0%	61.6%	70.3%
w/o juman	NB	71.3%	75.3%	73.2%
	SVM	82.7%	60.5%	69.8%

この表における，with juman は，形態素解析器 JUMAN におけるドメインの使用の有無であり，w/o juman は JUMAN のドメインを学習の素性として使用していない，という意味である．また，NB は NaiveBayes を表す．

本研究では，選好を無駄なく抽出することに重点をおく．そのため，実験結果で一番重要視するべき部分は選好の再現率である．再現率は，システムが正解と認識したデータが，正解データである確率である．

NaiveBayes 法では，選好の結果が，約 70%の精度，再現率が得られた． SVM 法における結果は，再現率は NaiveBayes 法よりは低い値であるものの，精度が 80%と高い値となった. NaiveBayes 法と比べれば, 抽出結果は殆どが選好のツイートであるため，さらに分類の正確性を求める場合には SVM 法を使用した方が良い結果が得られるということがわかる．NaiveBayes 法と SVM 法において，様々なツイートの中から，選好のツイートとそれ以外のツイートに分類する場合は，どちらも大体 7 割程度分類が行えることが分かった．

5.3 実験２

本節では，4 章で述べた提案手法に関して選好のツイートの中から，ST，FT，PT それぞれの分類を行った実験について述べる．

実験 1 とは違い, データ自体は全て選好のデータを用いる. 選好のデータから, ST, FT，PT にどの程度分類できるかを調べる．実験手順は 5.1 節で述べた実験 1 と同様の実験方法で実験を行う．

5.3.1 実験結果

4 章で述べた提案手法に関して行った実験(実験 2)の結果として，5.1 節で述べた学習方法，NaiveBayes 法，SVM 法それぞれの実験結果を表 5.6，表 5.7，表 5.8 に示す．

表 5.6：実験結果(ST)

			精度	再現率	F値
傍観(ST)	with juman	NB	64.1%	49.2%	55.4%
		SVM	57.8%	66.6%	61.8%
	w/o juman	NB	63.1%	48.5%	54.6%
		SVM	57.5%	66.0%	61.3%

表 5.7：実験結果(FT)

			精度	再現率	F値
追従(FT)	with juman	NB	45.3%	52.1%	48.3%
		SVM	52.7%	36.9%	43.1%
	w/o juman	NB	45.0%	49.7%	47.2%
		SVM	50.6%	34.2%	40.7%

表 5.8：実験結果(PT)

			精度	再現率	F値
先駆(PT)	with juman	NB	62.7%	46.0%	53.0%
		SVM	57.9%	39.6%	46.7%
	w/o juman	NB	63.5%	48.0%	54.4%
		SVM	59.4%	40.1%	47.5%

4 broomie，http://code.google.com/p/broomie/wiki/broomie_turorial_ja/.
*5 TinySVM，http://chasen.org/~taku/software/TinySVM/.

Vol.2011-NL-204 No.7
2011/11/21

表 5.6 より，ST の結果では，NaiveBayes 法による再現率が 49%程度となった．SVM 法によるツイートの再現率が 66%となった．表 5.7 より，FT の結果では，NaiveBayes 法での再現率が 52%，SVM 法は，他の結果より極端に低く，37%程度となっている．表 5.8 より，PT の結果では，NaiveBayes 法での再現率が 46%，SVM 法は，約 40%程度となっている．

5.4 考察

5.2 節, 5.3 節で述べた実験結果を比較する．ここでは実験 2 の結果を中心に行なう．

表 5.6，表 5.7，表 5.8 より，選好のデータ内から，ST，FT，PT それぞれに分類結果から考察する．

ユーザの選好に基づいた記事分類のため，実験 2 の結果は非常に重要であるが，全ての品詞を使用していると，名詞，動詞，形容詞と言った内容語が多く，どの属性にも含まれる可能性があるため，既存研究によく用いられていた内容語中心のやり方では精度再現率の上昇は見込めないと思われる．

そこで，内容語ではなく，助詞や助動詞等の表現に重みを付けて行う．実験方法は実験 1，実験 2，と同様の手法をとる．これまでと違う点は助詞と助動詞それぞれに重みをつける点である．助詞や，助動詞に重みをつける理由は，ツイートに関して，PTU が欲する宣伝のツイートや，FTU が欲する，意見や感想のツイートに含まれやすく，助詞，助動詞の品詞に重みをつけると，その素性が重要視されるため，より詳細に分類できると考えられるからである．この重み付けの処理を含めて，3 種類の属性に関して考察を行う．

ST の場合，助詞，助動詞に Twitter 特有表記と同様の重みを付けた場合の結果と，表 5.6 との差分をグラフとして，図 5.1 に示す．

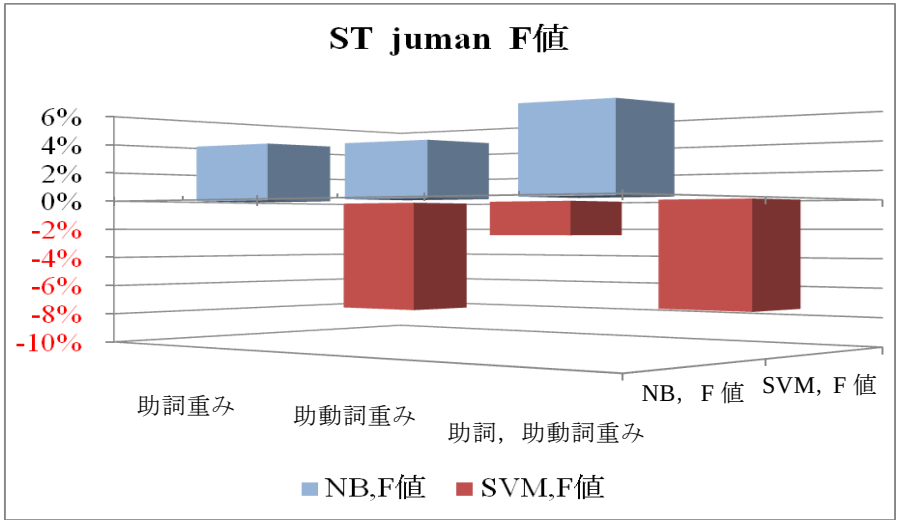

図 5.1：ST における，重みの有無による結果の差分

まずグラフの説明を行う．これは左から，助詞に重みをつける，助動詞のみに重みをつける，助詞，助動詞どちらにも重みをつける，という 3 種類の処理を行った場合の NaiveBayes と SVM の F 値のグラフである．表 5.6 の結果の差分をグラフとしている．図 5.1 では，手前のグラフである NaiveBayes 法の結果が，重み付け前の表 5.6 の結果より高かったため，グラフは上に伸びている．対して SVM 法の結果は，表 5.6 の結果より低かったため，グラフは下に伸びている．

図 5.1 より，助詞，助動詞に重みを付けた，NaiveBayes 法による値が，表 5.6 と比べ，値が 6%近く高くなっていることが分かる．一方で SVM に関しては重みをつけることにより，値が下がってしまっている．よって，選好のツイート中から ST のツイートを抽出する場合，助詞，助動詞に着目した上で，NaiveBayes 法によって分類をしたほうが抽出しやすくなることが言える．

次に FT において，図 5.1 と同様に処理を行った結果のグラフを図 5.2 に示す．

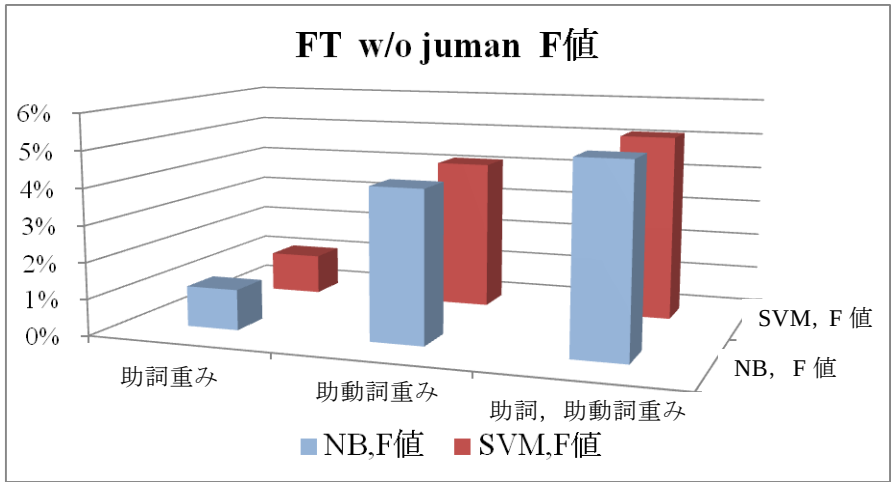

図 5.2：FT における，重みの有無による結果の差分

図 5.2 より，FT においては，助詞，助動詞共に重み付けを行うほど結果が 5%程上昇していることが分かる．しかし，ドメインを使用していない場合に限る．ドメインを学習データに使用すると，重みをつけるほどに悪くなっていってしまう．ドメインが付与されるのは名詞が殆どである．つまり，このドメインが付与されているツイートに ST，FT，PT それぞれの違いが学習器で判別できていないためだと考えられる．この事はデータセットのデータ量が少ない理由にも当てはまると考えられる．

最後に，PT の考察を行う．前述した 2 つと同様に，助詞や，助動詞に重みを付けて行った場合の結果を以下の図 5.3 に示す．

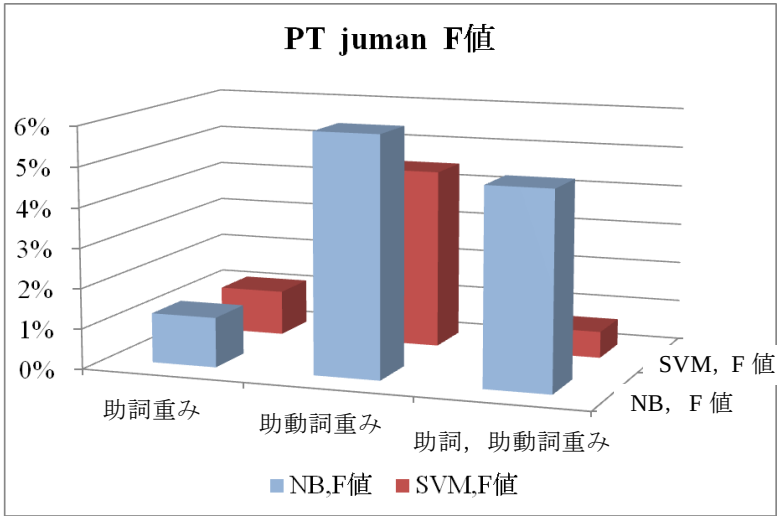

図 5.3：PT における，重みの有無による結果の差分

図 5.3 より，PT の場合だと，助動詞のみに重みを付与した場合に，NaiveBayes 法 SVM 法ともに，表 5.6 と比べ，5%程度高くなった．この要因は助動詞の伝聞，「ようだ」「そうだ」等の表現が PT に含まれるためだと考えられる．

選好ツイート内での分類は，実験 1 と同様の手法の場合，精度，再現率の上昇は見込めなかった．この原因は選好のツイートは 3 種類の属性同士の殆どが内容語であるため，それぞれのツイートの違いを学習できず，内容語だけでは分類することは難しいと言える．従って，選好ツイート内での 3 種類の属性の分類は，内容語だけでなく，3.3 節で述べた「意志」にあたる，助詞や助動詞と言った表現の品詞等の素性に着目し，重みを付与することで，精度や再現率等の数値の改善出来る．

6. おわりに

本研究では，ユーザにとって有益な情報の記事，選好の記事のみを表示するシステムの構築を目標にマイクロブログに投稿されている記事を学習させ，選好とそうでないツイートに分類するための提案を行った．提案手法では，まず，ツイートに一つの属性を付与する．属性は選好とその他の 2 種類．人手で 3506 件のツイートに属性付与を行った．属性付与の際は，選好の定義を ST，FT，PT，受身の 4 つに分けて付与し，学習時には受身以外の定義を一つにまとめて行なった．学習時に，Twitter 特有の表記，

JUMAN の形態素解析結果に含まれるドメイン，伝達に関連する品詞を素性としてデータに付与することにより，選好のツイートの再現率向上を計った．学習方法には NaiveBayes 法と SVM 法を使用した．

提案手法の有効性を調べるために，前述したデータを使用し，実験を 2 種類行った．

まず，全ツイートデータから，選好のツイートとその他のツイートに分類する実験を行った．実験の結果，NaiveBayes 法では，選好に関して，精度 71.1%，再現率 75.4%，F 値 73.1%となった．また SVM 法では，選好に関して，精度 82.0%，再現率 61.6%，F 値 70.3%となった．どちらも高い値であり，選好とその他の分類に関して，提案手法の有効性が確認できた．

2 つ目の実験は，データセットは全て選好のデータとし，その中で，選好の定義である，ST，FT，PT それぞれを分類する実験を行った．3 つそれぞれの結果は，どれも 50%程度の結果となった．しかし，この実験に，助詞や助動詞のような，品詞等に着目して重みをつけることにより，ST，FT，PT それぞれに値が上昇する特徴を持っていることが分かった．重み付けによって，3 種類各々の分類の結果は，重み付けを加える前よりも，6%程度上昇する結果となった．

7. 今後の課題

今後の課題として，前述したが，選好はユーザにとって有益な情報という定義であるため，3 種類の属性それぞれの違いを見出しにくい．そのため，例え大量のツイートから，選好のツイートが正しく抽出できたとしても，ユーザに提供するための選好の部分の精度再現率の値が低くなってしまう．精度再現率低下の対策に関しては 5.4 節で述べた ST，FT，PT それぞれの記事にある特徴に重みをつけることで，精度再現率等の値を上昇させることが出来る．今後さらに「～に違いない」等のモダリティ表記等に重みをつけることにより，更に ST，FT，PT の正確な分類が行う必要がある．

次に，投稿記事自体に関して，記事の長さはそれぞれ違い，一つの記事に 2 文掲載されている場合もある．本研究では 1 つの記事に対して 1 つのラベルしか付与していない．そのため，2 文の中身が全く別の内容，つまり，1 文は ST の内容であるが，もう 1 文は PT の内容，と言った記事が存在する場合がある．この場合の定義に関して，表示する単位はツイート単位であるため，今回は 1 つの正解ラベルを付与している．しかし，学習の際に弊害が起こる．したがって，今後の課題として一つの記事に複数のラベルを付与する方法を取り入れるなどして改めて実験を行う必要がある．

データセットに関しても，マイクロブログは表現が多岐にわたるため，今回行った実験に使用したデータセットの場合，素性数が約 7000 という値が，学習に十分な語彙数をカバーしたとは言いにくい．また，正解ラベル作成も一人で行ったため，客観的とは言い難い．よって，データセットも質，量共に高めていくべきだと考えられる．

そして，最終的には，ユーザがフォローしたユーザのタイムラインを閲覧する際，内容後に基づくユーザの嗜好に加えて，本研究の知的欲求に基づく選好により，フィルタリングを行うシステムの構築を実現したい．

謝辞

この研究の一部は，平成 23 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] 青島傅隼，横山昌平，福田直樹，石川博，“文脈を考慮したリアルタイムなツイート繋がり抽出・提示システムの試作，” 3rd Rakuten R&D Symposium，2010.
[2] 向井友宏，黒澤義明，目良和也，竹澤寿幸，”マイクロブログの分析に基づくユーザの嗜好とタイミングを考慮した情報推薦手法の提案，”言語処理学会年次大会，2011.
[3] 岩木祐輔，アダムヤトフト，田中克己，“マイクロブログにおける有用な記事の発見支援，” DEIM Forum 2009，2009.
[4] 田中淳史，田島敬史，“twitter のツイートに関する分類手法の提案，” DEIM Forum 2010 A5-4，2010.
[5] 黒澤義明，竹澤寿幸，”マイクロブログの返信行動に着目した投稿及びユーザの分類，”言語処理学会年次大会，2011.
[6] 青島傅隼，横山昌平，福田直樹，石川博，“マイクロブログを対象とした制限付きクラスタリングの実現，” DEIM Forum 2010 B1-3，2010.
[7] 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.
[8] 市川博通，黒澤義明，目良和也，竹澤寿幸，"マイクロブログを用いた音声認識用モデルの構築及び分析，"言語処理学会年次大会，2011.