価値の時間依存性に基づくマイクロブログ記事の分類

竹村　光†　田島　敬史††

† 京都大学工学部情報学科　〒 606–8501 京都府京都市左京区吉田本町
†† 京都大学大学院情報学研究科　〒 606–8501 京都府京都市左京区吉田本町
E-mail: †takemura@dl.kuis.kyoto-u.ac.jp, ††tajima@i.kyoto-u.ac.jp

あらまし　従来の Web やブログに比べ，Twitter を始めとするマイクロブログの記事の中には，投稿されてから短時間のうちに読んでこそ価値があるというものが多く含まれる．しかし，マイクロブログには日々膨大な量の記事が投稿されているため，全ての記事に目を通すには時間がかかり，ある記事を読んだ時には，すでにその記事の価値が下がっていたという事態が頻繁に発生する．そこで本研究では，まず，各記事の価値が，時間とともに減衰するようなものか，それとも，特に時間に依存しないようなものかを判定する手法を提案する．提案手法では，マイクロブログ記事の時間依存性を特徴付けるような様々な特徴量を SVM に学習させることにより，判定を行う．さらに，この判定結果に基づき，記事を，今見るべき記事，時間の経過によりすでに価値が減衰しているので見なくてよい記事，価値が変化しないのであとで見ても構わない記事の 3 つに分類して表示する方式を提案する．

キーワード　Twitter, ツイート, 自動分類, 緊急性

1. はじめに

ブログや SNS というソーシャルメディアの普及に伴い，誰もが簡単に Web 上で情報を発信できるようになった．特に近年では，マイクロブログと呼ばれる，SNS の性質を併せ持ったブログサービスが，爆発的な成長を遂げている．

マイクロブログとは，記事を短いテキスト形式で投稿するブログサービスであり，従来のブログサービスや SNS に比べて，手軽かつ簡易に記事を投稿でき，かつ情報発信を必ずしも主な目的としないというのが大きな特徴である．このような特徴のため，その時の思いつきで記事を投稿するといったことも多く行われ，その結果，マイクロブログでは日々膨大な量の記事が投稿されている．さらに記事の内容も，ニュースの記事のように他人に読んでもらうことを意識した記事もあれば，単なる自身の状況の説明や雑記のような記事もあり，多岐に渡っている．

また，このような投稿の手軽さのため，マイクロブログでは，リアルタイムな状況に付随した記事が非常に多くなっている．したがって，ユーザは「今どこで何が起こっているのか」「今誰が何をしているのか」といった，リアルタイムな状況を共有することが可能である．このような性質から，マイクロブログは現在，新たな情報発信メディアとして大きな注目を集めている．

このように，リアルタイムに記事が投稿され続ける状況の中で，全ての記事に目を通すのは，ユーザにとってかなり時間のかかる作業である．しかし，マイクロブログでは，リアルタイムな状況に付随し，時間とともに価値が失われていくような記事が非常に多い．例えば，あるユーザの「今から誰かご飯行こう」という記事を，それが投稿されてから 1 日後に読んでも，価値は残っていない．もちろん，いつ目を通しても情報の価値がほとんど変化しないような記事も含まれるが，マイクロブログでは，従来のブログサービスと比べると，前者の記事が圧倒的に多くなっている．そのため，ある記事を目にした時には，すでにその情報の価値はほとんど失われていたというような事態が頻繁に発生する．また，現時点で価値のある情報が，すでに価値を失ってしまった情報に埋もれてしまうというような事態も多い．

そこで本研究では，どの記事を優先的に見るべきかを判定するために，まず，各記事の価値が，時間の経過とともに減衰するようなものなのか，それとも，時間に依存しないようなものなのかを判定する手法を提案する．提案手法では，マイクロブログ記事の時間依存性を特徴付けるような様々な特徴量を SVM に学習させることにより，判定を行う．そして，この判定結果に基づき，記事を，今見るべき記事，時間の経過により既に価値が減衰しているのでもう見なくてよい記事，価値が時間に依存しないのであとで見ても構わない記事の 3 つに分類して表示する方式を提案する．このように分類することで，例えば，今忙しくて全ての記事を読むことのできないようなユーザの場合，今見るべき情報のみを見るといったことができるし，また，時間があるのでゆっくり記事を読みたいようなユーザの場合は，今見るべき情報に加えて，いつ読んでもよい情報も見るといったことが可能になる．これにより，ある記事に目を通した時にはすでにその価値が失われていたという事態を防ぐことができる．さらに，価値のある記事が，すでに価値を失った記事に埋もれてしまっているために見落としてしまったというような事態への防止にもつながる．

実験は，Twitter(注1)を用いて行う．Twitter は，現在最も普及しているマイクロブログサービスの一つであり，2011 年 4 月時点でユーザ数が 2 億人を超えている [1]．また，同年 10 月現在，1 日に 2 億 5,000 万以上の記事が投稿されていると言われている [2]．Twitter では，ツイートと呼ばれる，上限 140 字のテキストを記事として投稿する．また，Twitter には，フォ

(注1)：http://twitter.com/

図 1　Twitter ホームページのスクリーンショット

ローという，自分の読みたいユーザの記事をホームページに表示させる機能があり，図 1 のように，フォローをした全ユーザの記事が時系列順にソートされてホームページに表示される．これをタイムラインと呼んでいる．さらに，リプライと呼ばれる，ある記事への返信という形で記事を投稿する機能も有しており，これにより，ユーザ間のコミュニケーションも円滑に行うことができるようになっている．本研究では，Twitter から実際に収集したデータを用いて，提案手法の評価実験を行った．

2. 関 連 研 究

マイクロブログ記事の分類に関する研究は近年盛んに行われている．分類に関する研究としては，田中ら [3] による，ツイートが広く一般の人にとって情報として有用なものとそうでないものに分類する研究，Irani ら [4] による，ツイートがスパムであるかそうでないかに分類する研究，Sriram ら [5] や西田ら [6] による，ツイートをトピックにより分類する研究など，様々な観点からの分類に関する研究が行われている．本研究では，その有用性が時間に依存するかそうでないかでツイートを分類しており，これらの研究とは分類の目的が異なる．

岩木ら [7] は，マイクロブログにおいて，ユーザにとって有用性のある記事を効率よく発見するための支援手法を提案している．作成した感性辞書による記事のジャンルの特定と，ユーザ同士の近接度の計算により，ユーザと記事の近接度を計算し，記事の有用性を求めている．本研究では，記事の有用性を，静的なものとしてではなく，時間とともに変化するような動的なものとしてとらえている点で，岩木らの研究とは異なる．

マイクロブログのリアルタイム性に着目した研究も行われている．Sakaki ら [8] は，ユーザの投稿から地震や台風などのイベントをリアルタイムに取得し，さらにその発生地・中心地を観測する手法を提案している．Sakaki らの手法では，SVM により，記事がイベントに関するものかを取得することにより，イベントが発生しているかどうかを検出している．また，イベントが起きていると考えられる場合は，記事の投稿日時と位置情報を用いることによって，その発生地・中心地を求めている．Mathioudakis ら [9] は，取得したツイートからバーストキーワードを発見し，キーワードの共起性によってグルーピングを行うことで，リアルタイムに変動するトレンドの発見を行っている．さらに，その手法を用いたアプリケーション「TwitterMonitor」を作成している．本研究も，記事の有用性の時間による変化を考えており，マイクロブログのリアルタイム性に着目した研究であるといえる．しかし本研究は，リアルタイムに発生したイベントの発見や，トレンドの発見が主な目的ではなく，今見るべき情報かそうでないかを判定することを目的としている．よって，これらの研究とは目的が異なる．

3. 記事の価値の時間依存性

本章では，マイクロブログ記事の価値の時間依存性を定義し，それを生じさせる要因について述べる．さらに，価値の時間依存性に基づいて，どのように記事の分類を行うかについて説明する．

3.1 時間依存性の定義

本論文では，記事の価値の時間依存性を，ある時刻 t において，その時点で記事を見るのと，後で見るのとで有意に価値の差があるかどうかで定義する．価値は 0〜1 での範囲で定め，記事が投稿された時点では，価値は 1，すなわち価値が最も高い状態にあるものとする．記事の価値の時間推移の例を図 2 に示す．「地震だ」の場合，しばらく価値の高い状態が続くが，その後減衰し，最終的には価値はほとんどなくなる．よって，投稿された直後は時間依存性が高いが，しばらく経つと時間依存性は低くなるといえる．また，「金閣寺きれい」の場合，投稿されてからしばらく時間が経過しても，価値は投稿時とほとんど変わらない．よって，この記事は時間依存性が常に低いといえる．

また，時間依存性は，どの程度のスパンで記事の価値の変化を考えるのかにも依存する．例えば「春って気持ちいい」という記事が投稿された場合，数日間というスパンで見ると価値はほとんど変わらないが，数ヶ月というスパンで見ると価値は減衰している．よって，どの程度のスパンで時間依存性を考えるのかも考慮するものとする．

以下，時間依存性の形式的な定義を与える．現在時刻を表すパラメータを t_n，どの程度のスパンで時間依存性を考えるのかを表すパラメータを δ_s とした時，記事 a の時間依存性 $TimeDependency(a, t_n, \delta_s)$ を以下のように定義する．

$$TimeDependency(a, t_n, \delta_s) = \frac{|val(a, t_n) - val(a, t_n + \delta_s)|}{val(a, t_n)}$$

ここで，$val(a, t)$ は，時刻 t における記事 a の価値を表す．このように，時間依存性は，現在時刻での記事の価値と，一定時間後の記事の価値の変位との比で表す．

3.2 時間依存性を生じさせる要因

情報の価値に時間依存性が発生するのは，時間の経過に伴い，実世界に何らかの変化が生じるためである．以下では，価値の変動の仕方に基づいて大きく二つの場合に分け，それぞれの場合について，記事の価値に時間依存性を引き起こす要因について考察する．

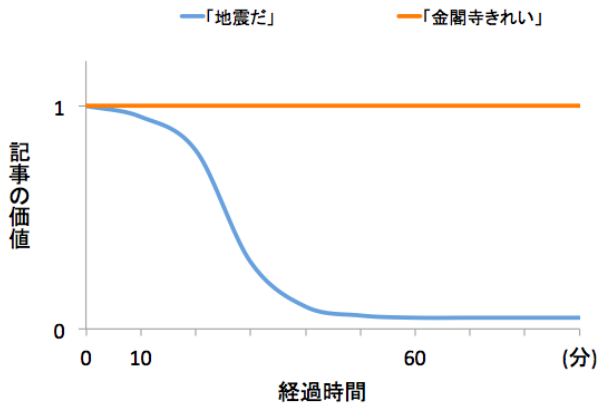


図 2　記事の価値の時間推移の例

(1)　記事の内容が実世界のイベント・状況のリアルタイムな発生・変動に付随している場合

実世界は時間とともに移り変わっていくものなので，実世界のイベントに付随する情報の価値は，時間の経過とともに減衰していくのが普通である．例えば「揺れた！」や「雨が降ってきた」という記事は，それぞれ地震の発生，天気の変動という実世界のリアルタイムなイベントに付随しているので，時間依存性が生じている．また,「今から飲みに行かない？」「いいよ」のような，リアルタイムに進行する会話も，状況が時間とともに変動していくので，この場合に当てはまると考えられる．このような記事の価値は，通常，リアルタイムに変動し続ける．

また，記事の投稿は，常に上記のようにリアルタイムなイベントと同時に行われるとは限らない．例えば，災害時の避難場所一覧のような記事は，投稿された時点では価値があまりなくても，その後，何か災害が発生した時には価値は上昇する．同様に，何らかの事象がある時点で社会的に話題になったような時にも，その事象に関する記事の価値は上昇すると考えられる．このように，そのトピックに関連する記事の価値が，ある特定の時点で高くなるようなトピックは，記事の価値の時間依存性を生じさせる要因となる．このようなトピックのことを，本論文では，その時点におけるホットトピックと呼んでいる．

(2)　記事の内容が実世界のイベント・状況の未来の発生・変動に付随していると思われる場合

記事の内容が未来のイベント付随していたり，記事に期限・期間などの情報が含まれているような場合，その価値は，ある特定の時点を過ぎると大きく変動し始めると考えられる．例えば「イベントは 17 時で終了です」という記事は，17 時を過ぎると価値は大きく減少し始める．よって，このような性質の記事は，時間依存性が生じているといえる．また，このような記事の価値は，ある特定の時点までは価値がほとんど変わらないが，その時点を過ぎると大きく変動し始めるものが多い．

4. 提 案 手 法

本章では，前章で述べた時間依存性の定義に基づき，マイクロブログ記事を分類する手法を提案する．

本研究では，Twitter を対象としている．また，ユーザがログインして自分のタイムラインを見る時に，そのタイムライン中のツイートの分類を行うことを想定している．提案手法では，ユーザのタイムラインを入力とし，まず，各ツイートに対して，価値の時間依存性を特徴付けるような様々な特徴量を求める．次に，投稿された直後 (ここでは 10 分以内とする) にはツイートに価値があると仮定した上で，投稿されて 1 時間後，6 時間後，1 日後，1 週間後，それ以降にそれぞれ価値があるかどうかを判定する，5 つの SVM を用意する．そして，各ツイートをこれらの SVM に判定させることにより，そのツイートが投稿されてから上記の各時間にそれぞれ読む価値があるかどうかを判定する．その判定結果と，各記事が投稿されてからの経過時間を基に，各ツイートを，今見るべき記事，時間の経過により既に価値が減衰しているので見なくていよ記事，価値が変化しないので後で見ても構わない記事のいずれかへ分類し，その結果を出力とする．

例として,「地震だ！」というツイートが，5 分前に投稿されていたとする．まず，そのツイートに対して様々な特徴量を求め，それらの特徴量からなる特徴ベクトルを，前述した 5 つの S VM に入力する．ここでは，1 時間後，6 時間後，1 日後，1 週間後，それ以降の全ての SVM で価値なしと判定されたとする．すると，投稿された直後には読む価値があるが，1 時間以上経過すると読む価値がなくなってしまうということが分かる．ここで再度ツイートを見ると，投稿されたのは 5 分前であることが分かるので，このツイートは，現在は読む価値があるが，時間が経過すると読む価値が減衰してしまうと考えられる．よってこのツイートは「今見るべき記事」に分類される．

以下，記事の価値の時間依存性を判定する手法と，その判定結果により記事を分類する手法について示す．

4.1　時間依存性の判定手法

本節では，記事の価値の時間依存性を，SVM に学習させることにより判定する手法を提案する．以下に，SVM で用いる特徴量として採用したものを示す．なお，特徴量による分類結果の偏りをなくすため，各特徴量は基本的に 0〜1 の範囲になるようにスケーリングを行なっている，

(1)　文字数

Twitter では，文字数は 140 字で制限されているが，その中でも，文字数にはばらつきがある．リアルタイムに起こったことや，思いついたことをその場で記事にする場合，文字数は短くなることが多いと考えられる．また，ある事象に対する自分の考えをまとめた記事のような，従来のブログサービスと同じような使い方をする場合，文字数は長くなることが多いと考えられる．ゆえに，一般的に，文字数が短い程，時間依存性が高い可能性が大きいと予想される．このように，文字数は時間依存性の判定に役立つと考えられるので，特徴量として採用する．

以下のように，投稿されたツイート a の文字数の特徴量 $TextLength_a$ を求める．

$$TextLength_a = \log len(a)$$

ここで $len(a)$ は，ツイート a の文字数を表す．また，ここで対

数をとっているのは，文字数が 5 文字と 10 文字のような小さな値での違いを，100 文字と 105 文字のような大きな値での違いと比べて明確にするためである．以後，特徴量で対数をとっているものは，これと同様の理由である．

(2)　ツイートしたユーザの性質

ユーザによって Twitter の使い方は様々である．有用な情報の発信のために Twitter を使っているユーザもいれば，他のユーザとのコミュニケーションのために使っているユーザもいる．そして，ユーザが普段どのような使い方をしているのかを考えることは，記事の時間依存性の判定に役立つと考えられる．例えば，他ユーザとのコミュニケーションを目的として，チャットのように Twitter を使っているユーザは，時間依存性の高い記事を多く投稿すると考えられるし，あるトピックに関する有用な情報を発信するために Twitter を使っているユーザは，時間依存性の低い記事を多く投稿すると考えられる．

ユーザの性質の特徴量としては，以下のものを採用している．

- ログインしているユーザと相互フォローかどうか

あるユーザが，自分と相互にフォローし合っていれば，そのユーザは知人である可能性が高いといえる．また，自分が一方的にそのユーザをフォローしていれば，そのユーザは自分にとって有用な情報を発信するユーザである可能性が高いといえる．このように，自分と相互フォローであるかどうかを判定することは，そのユーザの性質を特徴付ける要素になるので，特徴量として採用する．特徴量は，記事を投稿したユーザが自分と相互にフォローしていれば 1，そうでなければ 0 とする．

- 過去のツイートの傾向

あるユーザの過去のツイートを解析することにより，そのユーザが普段どのような記事を投稿しているかを知るのに役立つ．ユーザの最近のツイートを最大 200 件取得し，それを基に特徴量を求める．特徴量としては，ユーザの過去のツイートの平均文字数，文字数のばらつき (標準偏差)，会話頻度を取得する．平均文字数は，(1) と同様の理由により有用であると考えられ，文字数のばらつきは，bot(注2)のような，常に決まった形式のツイートを投稿し続けることが多いようなアカウントと，通常のユーザとを見分けるのに役立つと考えられる．また，会話頻度は，ユーザがどの程度 Twitter で他のユーザとコミュニケーションを行っているかを知るのに役立つ．

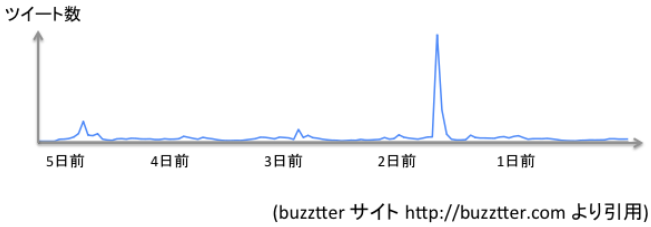
取得したユーザ u の最新ツイート最大 200 件の集合を A_u とする．まず，ユーザの過去のツイートの平均文字数の特徴量 $AvgTextength_u$ は，以下のように求める．

$$AvgTextLength_u = \log \frac{1}{|A_u|} \sum_{a \in A_u} len(a)$$

次に，ユーザの過去のツイートの文字数のばらつき $SdTextLength_u$ を，以下のように求める．

$$S_u = \log \sqrt{\frac{1}{|A_u|} \sum_{a \in A_u} (len(a) - m)^2}$$

(注2)：Twitter で自動的にツイートを投稿するアカウント及びそのプログラムのこと



(buzztter サイト http://buzztter.com より引用)

図 3　ある期間における「地震」を含むツイート数の時間変化

ここで，m は A_u 内のツイート a の文字数 $len(a)$ の平均であり，以下の式で表される．

$$m = \frac{1}{|A_u|} \sum_{a \in A_u} len(a)$$

最後に，ユーザの過去のツイートにおける会話頻度 $Conversation_u$ を，以下のように求める．

$$Conversation_u = \frac{|\{a \in A_u \mid \text{a はリプライである }\}|}{|A_u|}$$

ただし，リプライとは，あるツイートへの返信という形で投稿されるツイートのことを指す．

(3)　タームの出現頻度の上昇度

あるタームを含むツイートの数が，時間の経過とともに急増した場合，そのツイートは時間依存性が高く，その時点でホットなトピックであると考えられる．ここで取得するタームは，ツイートの内容を特徴付けると思われる，名詞，動詞，形容詞に限定する．以下，タームとは，これらのもののみを指すものとする．

例えば「地震」というタームは，地震が発生した直後に出現数は多くなる．図 3 に,「地震」を含むツイート数の時間変化を示す．このように，タームの出現頻度の急激な上昇は，リアルタイムなイベントに付随して発生することが多いので，時間依存性の判定に役立つと予想される．

タームの上昇度としては，(i) 現実世界全体でのタームの使用頻度の上昇度，(ii) 各ユーザのタイムライン内でのタームの出現頻度の上昇度の 2 つを取得する．以下，それぞれの場合について特徴量を求める．

(i)　現実世界全体でのタームの使用頻度の上昇度

まず，現実世界全体でのタームの使用頻度の上昇度を取得する．これを直接的に取得することは困難なので，ここでは，Google 急上昇ワード(注3)を用いて近似を行う．特徴量は，ユーザのログイン時に Google 急上昇ワードから取得したタームがツイート内に含まれていれば 1，そうでなければ 0 とする．これにより，現実世界全体でのホットなトピックに関するツイートを識別できる．

(ii)　各ユーザのタイムライン内でのタームの出現頻度の上昇度

次に，各ユーザのタイムライン内でのタームの出現頻度の上昇度を取得する．これにより，現実世界全体では出現頻度が上

(注3)：http://www.google.co.jp/m/trends

昇していないが，あるユーザのタイムライン内でのみ盛り上がっているようなトピックに関するタームも取得することができる．

以下，特徴量を求める．まず，ユーザのタイムライン内の全てのツイートを形態素解析し，取得された全てのターム w において，以下のように出現頻度の上昇度 $R(w)$ を求める．

$$R(w) = \sum_{a \in A} x_{w,a} \cdot idf_w$$

ここで，A はタイムライン中のツイートの集合，a はツイートをを表し，$x_{w,a}$ は，以下のように定める．ただし，c は定数であり，$E(a)$ は，ユーザがログインした時間と，ツイート a の投稿時間との差を表す．

$$x_{w,a} = \begin{cases} \dfrac{c}{E(a)} & (w \in a) \\ 0 & (\text{otherwise}) \end{cases}$$

また，idf_w は，あるツイート集合において，各ツイートをドキュメントと考えた時の w の idf 値であり，以下のように定める．

$$idf_w = \log \frac{|A|}{|\{a \in A \mid a \text{ は } w \text{ を含む }\}|}$$

すなわち，タームが含まれているツイートが最近投稿されたものである程，スコアの重み付けは大きくなり，同じタームがタイムライン内に複数出現している場合，それらのスコアは加算される．ただしこれだけだと,「今日」のような普段からよく出現するタームのスコアも大きくなってしまう可能性があるので，idf を用いることにより，それらのスコアを下げている．これにより，出現頻度が上昇しているタームを取得することができる．

この結果を用いて，ツイート a の特徴量 $HotTerm_a$ を以下のように求める．

$$HotTerm_a = \max_{w \in a} R(w)$$

つまり，各ツイートを形態素解析して得られたタームの中で，最もスコアが高いものを，そのツイートの特徴量として採用する．

(4)　会話の時間間隔

Twitter は，コミュニケーションのツールとして使用される場合も多く，様々な会話が行われる．会話の中には,「どこか行かない？」「いいよ」といった，チャットのようなものもあれば，ある内容に関して複数人で議論を行うような，BBS のようなものもあり，その使われ方は多岐に渡っている．チャットのように時間依存性が高いものは，リアルタイム性が求められるため，会話の時間間隔は短くなることが多い．逆に，BBS のように時間依存性が低いものは，リアルタイム性はあまり重要でないため，チャットのような使われ方の場合よりも会話の時間間隔は長くなることが多いと考えられる．このように，会話の時間間隔は，時間依存性を特徴付ける要素になり得る．よって，あるツイートが他のツイートへのリプライである場合，それらツイートの投稿時間間隔を取得する．以下のように，投稿時間間隔の特徴量 $Interval_a$ を求める．

$$Interval_a = \begin{cases} \Delta(a, \alpha) & (a \text{ が } \alpha \text{ へのリプライである}) \\ 1 & (\text{otherwise}) \end{cases}$$

ここで，$\Delta(a, \alpha)$ は，ツイート a とツイート α の投稿時間の差を表す．

(5)　日時情報

ツイートに期間・期限などの日時が含まれる場合，3.2 節で述べたように，時間依存性を特徴付ける要素になり得るので，正規表現を用いて取得する．現在以降の日時情報が時間依存性に役立つのはもちろん，過去の日時情報も，その日時が既に過ぎてしまっているというネガティブな情報として役立つと考えられる．したがって，ここでは，ユーザのログインした日時以降のものに限らず，全ての日時を取得することとする．また，現在以降の日時と過去の日時は，それぞれ別の特徴量として取得する．ログイン時の日時を d，ツイート a に含まれる日時を d_a とし，以下のように日時情報の特徴量 $Date_a$ を求める．

$$Date_a = \frac{1}{|d_a - d| + 1}$$

ツイート a に日時情報が含まれていない場合，現在以降の日時の特徴量は 0，過去の日時の特徴量は 1 とする．また，過去の日時，未来の日時ともに 2 つまで取得する．例えば「イベントは 13 時から 20 時までです」といった期間を含むツイートなどは，どちらの日時も役立つと考えられるためである．

(6)　URL, 写真が含まれるかどうか

Twitter には，URL や写真を含んだツイートが投稿されることも多い．URL を含んでいる場合，その時点で有用な情報を発信している記事である可能性が高い．また写真は，その時に撮ったものを記事に含めることが多いので，時間依存性が高いものが多いと思われる．よって，これらの特徴量は役立つと考えられるので，採用する．

特徴量としては，ツイートに URL が含まれるかどうかを表す特徴量と，写真が含まれるかどうかを表す特徴量の 2 つを考え，それぞれ URL，写真が含まれていれば 1，そうでなければ 0 とする．

(7)　タームごとの時間依存性を判定

Twitter に出現するタームの中には，時間依存性の高い記事によく現れるタームもあれば，低い記事によく現れるタームもあると考えられる．例えば，Twitter でよく用いられる「なう」というスラングは，内容がリアルタイムなイベントや状況に付随するような記事に現れることが非常に多い．また,「地震」「雨」といったタームは，それぞれ地震が発生した時，雨が降っている時に出現頻度が高くなると考えられるので，時間依存性の高い記事によく現れると予想される．逆に，数学で使われる専門的な用語のようなものは，時間依存性の低い記事によく現れると考えられる．このように，各タームが時間依存性の高い記事によく現れるか，低い記事によく現れるかを判定すること

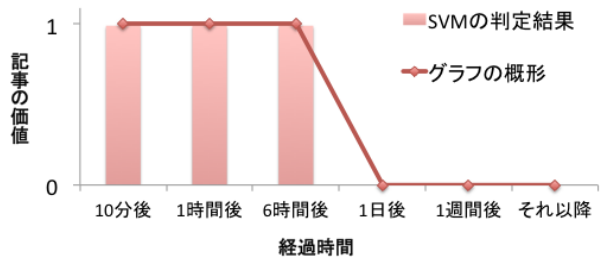


図 4　記事の価値の時間推移に関するグラフの概形の例

は，時間依存性の判定に役立つと考えられる．よって，特徴量として採用する．

タームの時間依存性を特徴量に変換するために，まず，投稿後いつまで価値があるか判定済みのツイート a に対して，値 v_a を，判定結果が 10 分後まで，1 時間後まで，6 時間後まで，1 日後まで，1 週間後まで，それ以降のどれであったかに応じて，それぞれ 0，0.2，0.4，0.6，0.8，1.0 で与える．そして，ターム w の時間依存性を表す値 v_w を．以下のように求める．

$$v_w = \frac{1}{|A_w|} \sum_{a \in A_w} v_a$$

ここで，A_w はターム w を含むツイートの集合を表す．

これを用いて，以下のようにツイートにおけるタームの時間依存性の特徴量 $TermDependency_a$ を求める．

$$TermDependency_a = \min_{w \in a} v_w$$

つまり，各ツイートを形態素解析して得られたタームの中で，最もスコアが低いもの，すなわち最も時間依存性が高いものを，そのツイートの特徴量として採用する．

これらの特徴量を，記事が投稿されてから 1 時間後，6 時間後，1 日後，1 週間後，それ以降にそれぞれ価値があるかどうかを表す 5 つの SVM に学習させることにより，時間依存性の判定を行う．各 SVM は，記事に対して価値がある状態，価値がない状態の 2 クラスへの分類を行い，これら 5 つの SVM の判定結果を総合して，時間依存性を判定する．例えば，1 時間後，6 時間後の SVM で価値がある状態，1 日後，1 週間後，それ以降の SVM で価値がない状態と判定されたとする．この場合，この判定結果をもとに，図 4 のようなグラフの概形を描けることが分かる．ただし，ここでは，記事に価値がある状態を 1，価値がない状態を 0 とし，記事が投稿させてから 10 分後は価値がある状態であると仮定している．これにより，いつ頃まで記事に価値があるかを判定することが可能となる．

4.2　記事の分類手法

本節では，前節における SVM の判定結果に基づき，各ツイートを，今見るべき記事，時間の経過により既に価値が減衰しているので見なくていよ記事，価値が変化しないので後で見ても構わない記事のいずれかに分類する手法を提案する．

ユーザのログイン時間とツイート a の投稿時間との差を表すパラメータを δ_a，SVM の判定結果に基づく，ツイートが投稿されてからいつまで価値があるかを表すパラメータを δ_v，時間依存性をどの程度のスパンで考えているかを表すパラメータを δ_s とおく．これを用いて，以下の条件で記事の分類を行う．

$$\begin{cases} \text{if } \delta_a > \delta_v & \text{もう見なくてよい} \\ \text{else if } \delta_a + \delta_s \geqq \delta_v & \text{今見るべき} \\ \text{otherwise} & \text{後で見てもよい} \end{cases}$$

例えば，ユーザがログインした時点において，投稿後 10 分後まで価値のある記事が，15 分前に投稿されていたとする．このような場合，記事の価値は現段階ではすでに失われてしまっていると考えられるので，もう見なくてよいに分類する．また，投稿後 6 時間後まで価値のある記事が，2 時間前に投稿されていたとする．このように，記事の価値が現段階でまだ残っている場合は，どの程度のスパンで時間依存性を考えているかにより，今見るべきか，後で見てもよいかを判定する．この記事は，あと 4 時間は価値があると考えられるので，もし 1 時間のスパンで時間依存性を考えている場合は，次にユーザが見た時にも価値は残っていると考え，あとで見てもよいに分類する．もし 6 時間のスパンで時間依存性を考えている場合は，次にユーザが見たときにはもう価値は失われていると考え，今見るべきに分類する．以上のようにして，3 つのクラスへの分類を行う．

5.　実験と考察

本章では，提案手法を実装することで評価実験を行い，得られた結果を基に考察を行う．また，提案手法を実装したアプリケーションについても述べる．

5.1　実験の概要

本実験では，Twitter API を用いて，Twitter から実験に必要なデータを取得した．以下，あるユーザがフォローしているユーザのことを，フォロイーと呼ぶ．まず，100 人の日本人ユーザをランダムに選択し，これらのユーザのフォロイー計 6,729 人が，2011 年 12 月 15 日から 12 月 22 日の間に投稿したツイートを全て収集した．ただし，100 人のユーザを取得する際，フォロイーが 0 人のユーザ，記事をフォロワー以外には非公開にしているようなユーザは除外した．これにより収集された 908,990 件のツイートを，本実験のデータセットとして用いている．

次に，上記の 100 人のユーザそれぞれに対して，12 月 16 日 0 時から 23 日 0 時までの 1 時間おきの日時の中から，ランダムに日時を 1 つ取得し，その日時に Twitter にログインした時に表示されるであろうタイムライン最大 100 件を再現した．そして，Twitter を使ったことのあるボランティア評価者 10 人を募集し，彼らに再現したタイムラインを見せ，タイムライン内の各ツイートが，いつまで読む価値があると思うかを，6 段階 (10 分後まで，1 時間後まで，6 時間後まで，1 日後まで，1 週間後まで，それ以上) の中から判定してもらった．これにより判定されたツイート 9,469 件を，SVM の訓練データとして用いている．ここで用いた SVM は 5 つあり，それぞれ投稿されて 1 時間後，6 時間後，1 日後，1 週間後，それ以降に価値があるかどうかを表すものである．また，各 SVM の訓練データは，

表 1　SVM の識別結果【1 時間後】

正解＼SVM	1	-1
1	5593	424
-1	708	2744

Accuracy = 88.045200

表 2　SVM の識別結果【6 時間後】

正解＼SVM	1	-1
1	3396	851
-1	734	4488

Accuracy = 83.261168

表 3　SVM の識別結果【1 日後】

正解＼SVM	1	-1
1	1788	1124
-1	528	6029

Accuracy = 82.553596

表 4　SVM の識別結果【1 週間後】

正解＼SVM	1	-1
1	1061	278
-1	428	911

Accuracy = 73.637043

表 5　SVM の識別結果【それ以降】

正解＼SVM	1	-1
1	620	187
-1	302	505

Accuracy = 69.702602

記事に価値がある状態を 1，価値がない状態を-1 としている．

本実験では，SVM のライブラリは LIBSVM(注3) を用いた．ガウスカーネルを使用して識別を行い，精度の測定には，10-fold クロスバリデーションを適用した．また，形態素解析ツールは MeCab(注4) を用いている．

5.2　実験結果と考察

5.2.1　SVM の精度

記事が投稿されてから 1 時間後，6 時間後，1 日後，1 週間後，それ以降にそれぞれ価値があるかどうかを判定する，5 つの SVM の識別精度を測定した．まず，これらの SVM に，4.1 節で提案した特徴量 (文字数，ツイートしたユーザの性質，タームの出現頻度の上昇度，会話の時間間隔，日時情報，URL・写真が含まれるか，タームごとに時間依存性を判定) を全て入れた時の識別結果を表 1，表 2，表 3，表 4，表 5 に示す．ここで，1 時間後，6 時間後，1 日後に価値があるかどうかを判定する SVM は，訓練データ 9,469 件全てを用いて学習させているが，1 週間，それ以降に価値があるかどうかを判定する SVM は，訓練データをそのまま用いると，訓練データの偏りによる不均衡データ問題が生じるため，訓練データのクラスのうち大きい方のクラスが，小さい方のクラスサイズと同じサイズになるようにランダムにサンプリングした後，学習を行っている．

1 時間後，6 時間後，1 日後に価値があるかを判定する SVM においては，80%以上の精度で正しく識別できていることがわかる．1 週間後，それ以降に価値があるかどうかを判定する SVM も，前者には精度が劣るものの，約 70%の精度で正しく識別できている，また，投稿されてからの経過時間が短い程，精度はよい傾向が見られるが，これにより，時間依存性の高い記事の方が，明確にその特徴が現れやすいということが分かる．

(注3)：http://www.csie.ntu.edu.tw/ cjlin/libsvm/
(注4)：http://mecab.sourceforge.net/

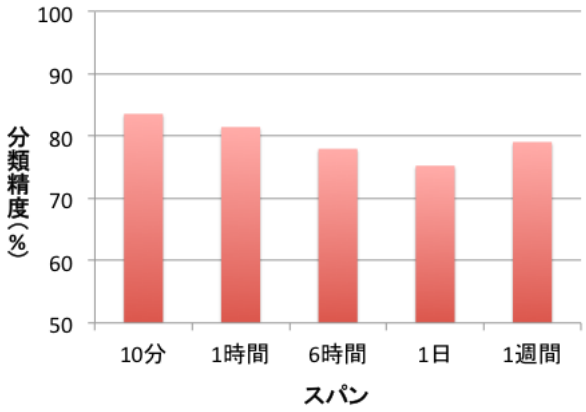


図 5　SVM の識別結果に基づくクラス分類の精度

次に，どの特徴量がどの程度識別精度に影響を及ぼしているかを判定するために，各特徴量をそれぞれ除いた場合の識別精度も測定した．この結果を，表 6 に示す．以下，この結果を基に考察を行う．

- タームごとの時間依存性を表す特徴量を除くと，識別制度が最も大きく下がっている．すなわち．この特徴量が，時間依存性の判定に最も影響を及ぼしていたことが分かる．この結果から，あるタームが時間依存性の高い記事によく現れるのか，低い記事によく現れるのかを判定することは，時間依存性の判定に大いに役立つといえる．

タームごとの時間依存性の具体的な調査を行った．タームの時間依存性の最も高い状態を 1，最も低い状態を 0 として，0〜1 の範囲で特徴量に変換した結果によると，時間依存性の高い例として「地震」が 0.91,「雨」が 0.70，低い例として「数学」が 0.33 といったものが挙げられ，これらの例に関しては，理想的な結果になっているといえる．

- 逆に，日時情報の表す特徴量を除くと，1 時間後，6 時間後，1 日後に価値があるかどうかを表す SVM において，識別精度は若干ではあるが上がっている．すなわち，日時情報は，これらの SVM の識別精度を下げているといえる．本研究では，日時情報は，期限・期日の判定に役立つと考えていたが，実際は，日時情報が含まれているからといって，期間・期日ではない場合も多いためではないかと考えられる．

5.2.2　クラス分類の精度

SVM の識別結果とツイートの投稿時間に基づいて，3 つのクラス (今見るべき記事，時間の経過により既に価値が減少しているのでもう見なくてよい記事，価値が時間に依存しないのであとで見ても構わない記事) へ分類した時の精度を測定した．図 5 に，測定結果を示す．ここでは，どの程度の時間間隔で時間依存性を考えるのかによって分け，それぞれの精度を測定している．その時間間隔を，時間依存性を判定する場合のスパンとして用いる．結果によると，全ての場合で 75%以上の精度で正しくクラス分類が行えていることが分かる．ここで，グラフの両端 (10 分，1 週間など) の精度が比較的高くなっている．この原因としては，例えば 10 分のスパンで時間依存性を考えて

表 6　各特徴量を除いた場合の識別制度

SVM / 除いた特徴量	1 時間後	6 時間後	1 日後	1 週間後	それ以降
なし	88.045200	83.261168	82.553596	73.637043	69.702602
文字数	87.369310	80.198543	80.019009	69.044063	66.480793
ツイートしたユーザの性質	87.696694	82.880980	81.919949	72.740851	68.897150
タームの出現頻度の上昇度	88.193051	83.292850	82.775372	73.076923	68.153656
会話の時間間隔	87.971275	83.166121	82.712008	72.292756	66.047088
日時情報	88.171929	83.092196	83.440701	73.300971	69.454771
URL・写真が含まれるか	88.098004	83.197803	82.743690	73.226288	69.392813
タームごとに時間依存性を判定	65.994297	64.610835	73.524131	54.144884	41.387856

図 6　実装したアプリケーションのスクリーンショット

いる場合だと，後で見てもよい記事の割合が，今見るべき記事に比べると圧倒的に多くなるというように，記事の分類に偏りが生じるため，分類が比較的容易になるからだと考えられる．

また，正しく分類できなかったものに関して具体的に調査を行った結果，時間依存性を考えるスパンが短いほど，後で見てもよい記事に誤って分類されるケースが多いことが確認された．これは，一般的に時間依存性が高いと考えられる記事でも，非常に短いスパンで考えると，ユーザが次にログインした時にも価値が残っているためだと考える．

5.3　アプリケーションの実装

本節では，本研究で提案した手法を実装したアプリケーションについて説明する．

まず，ユーザにどの程度のスパンで Twitter にログインするのかを入力してもらう，そして，そのスパンに基づき，記事を，今見るべき情報，後で見ても構わない情報，もう見なくてよい情報の三つのタブに分類し，それぞれを表示している．図 6 に，実装したアプリケーションのスクリーンショットを示す．

6. ま と め

本論文では，リアルタイムな状況に付随するような記事が非常に多いマイクロブログ記事の中から，優先的に見るべき記事を見つけるために，記事の価値が時間に依存するかどうかを判定する手法を提案した．提案手法では，記事が投稿されてから 1 時間後，6 時間後，1 日後，1 週間後，それ以降に価値があるかどうかをそれぞれ判定する 5 つの SVM を用意し，マイクロブログの時間依存性を特徴づけるような様々な特徴量をそれらの SVM に学習させることにより，判定を行った．さらに，その判定結果に基づき，記事を，今見るべき情報，時間の経過により既に価値が減衰しているので見なくてよい情報，価値が変化しないのであとで見ても構わない情報の 3 つに分類して表示する方式を提案した．そして，マイクロブログの 1 つである Twitter を用いて評価実験を行い．提案手法が時間依存性の判定と，それによる記事の分類に有効であることを示した．

今後の課題・展望としては，提案手法において，現段階では 5 つの SVM を独立なものとして扱っているが，それぞれの SVM の相互情報を用いることにより，精度をさらに向上させることや，本研究では，SVM を用いて記事を分類しているが，決定木などの他の手法でも同様に分類を行い，本手法との精度を比較することなどが挙げられる．また，本研究の評価実験では，SVM の訓練データは 9,469 件であったが，これは決して十分なデータ数とは言えないので，さらにデータ数を増やすことにより，評価実験を行うことも検討している．

謝辞　本研究の一部は科研費 (23650048) の助成を受けたものである．

文　献

[1] Nicholas Carlson. Twitter has less than 21 million ”active” users. http://www.businessinsider.com/twitter-has-less-than-21-million-active-users-2011-4, 4 2011.

[2] Alexia Tsotsis. Twitter is at 250 million tweets per day, ios 5 integration made signups increase 3x. http://techcrunch.com/2011/10/17/twitter-is-at-250-million-tweets-per-day/, 10 2011.

[3] 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. In *DEIM Forum 2010*, 2010.

[4] D. Irani, S. Webb, C. Pu, and K. Li. Study of trend-stuffing on twitter through text classification. In *Collaboration, Electronic messaging, Anti-Abuse and Spam Conference (CEAS)*, 2010.

[5] B. Sriram, D. Fuhry, E. Demir, H. Ferhatosmanoglu, and M. Demirbas. Short text classification in twitter to improve information filtering. In *Proceeding of the 33rd international ACM SIGIR conference on research and development in information retrieval*, pp. 841–842. ACM, 2010.

[6] 西田京介, 坂野遼平, 藤村考, 星出高秀. データ圧縮による twiter のツイート話題分類. *DEIM Forum 2011*, 2011.

[7] 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.

[8] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, pp. 851–860. ACM, 2010.

[9] M. Mathioudakis and N. Koudas. Twittermonitor: trend detection over the twitter stream. In *Proceedings of the 2010 international conference on Management of data*, pp. 1155–1158. ACM, 2010.