DEIM Forum 2012 F3-4

Twitterからの実生活情報の抽出法の提案

山本　修平†　　佐藤　哲司††

† 筑波大学情報学群知識情報・図書館学類　〒305–8550 茨城県つくば市春日 1-2
†† 筑波大学大学院図書館情報メディア研究科　〒305–8550 茨城県つくば市春日 1-2
E-mail: †yamahei@ce.slis.tsukuba.ac.jp, ††satoh@ce.slis.tsukuba.ac.p

あらまし　Twitterでは，多くのユーザがリアルタイムに，自分がしていることや考えていることなど，身近な「今」を投稿している．投稿する記事の中には，自分の経験や知識，身の回りの生活に関することなど，地域性が高く新鮮な情報がある．このような情報は，他のユーザの実生活における活動の支援に役立つと考えられる．例えば，電車の遅延情報や，スーパーの安売りに関する情報は，ユーザの実生活における活動に役立つ．本研究では，地域に生活するユーザを支援することを目的に，このような実生活情報をTwitterから抽出する手法を提案する．抽出した記事に生活の局面を自動的に付与した，実生活データベースを構築する．提案法を実装したプロトタイプシステムを用いて，実生活情報の抽出実験を行い，SVMを用いた手法による抽出精度と比較した結果，提案手法の有効性について確認したので報告する．

キーワード　Twitter，実生活，情報抽出

Real Life Information Extraction Method from Twitter

Syuhei YAMAMOTO† and Tetsuji SATOH††

† College of Knowledge and Library Sciences, School of Informatics University of Tsukuba
1-2, Kasuga, Tsukuba, Ibaraki, 305–8550 Japan
†† Graduate School of Library, Information and Media Studies, University of Tsukuba
1-2, Kasuga, Tsukuba, Ibaraki, 305–8550 Japan
E-mail: †yamahei@ce.slis.tsukuba.ac.jp, ††satoh@ce.slis.tsukuba.ac.p

Abstract

Key words　Twitter，Real Life，Information Extraction

1. はじめに

マイクロブログの代表としてTwitter(注1)は，最大140文字の短い文章からなる記事を投稿するサービスである．ブログでは1日から数日の出来事をまとめて投稿することが多いが，Twitterでは多くのユーザがリアルタイムに，自分がしていることや考えていることなど，より身近な「今」を投稿している．投稿する記事の中には，ユーザの経験や知識，また地域の生活に関することなど，地域性が高く新鮮な情報がある．例として，電車の運行情報や地域で行われているイベントといった，場所や時間が限定している有用性の高い情報がある．このような実生活情報が書かれた記事は，他のユーザの実生活における活動の支援に役立つと考えられる．

Twitterに投稿された実生活情報が実際にユーザを支援した場面として，2011年3月に起きた東日本大震災が挙げられる．地震の直後，震災の被害に遭った地域では断水や食料供給の不足や，電車等の交通手段に大きな乱れが生じた．この時，どこで給水や食料配布が行われているかや，電車が動いているかどうか，電車が何時に出発するかが書かれた記事がTwitterに投稿され，ユーザ同士で必要な情報を共有することにより，実生活の活動に役立った．

このように，ユーザにとって有用性の高い実生活情報が，Twitterにますます投稿されるようになってきたが，実生活に言及していない記事も多く存在する．特に，「ありがとう」や「そうなんだ」のような，誰かの投稿に対する相槌や共感といった，必ずしも共有する意義の高くない副次的な投稿も多い．このような記事の中に，実生活に言及している記事が埋没しているのがTwitterの現状であり，ユーザが各記事を読み，実生活に言及している記事を発見することは負担が大きく手間もかかる．

本研究では，ユーザの実生活における活動に有用な情報を提供することを目的に，Twitter上から実生活に言及している記事

(注1)：http://twitter.com/

を抽出する技術を確立する．抽出した記事に対して,「食事」や「交通」など，生活の局面に応じた属性を付与し，実生活データベースを構築する．尚，このような属性を用いることで，生活の局面に応じた実生活情報をユーザに提供する．例えば，ユーザがこれから電車に乗って移動する場合には「交通」の局面の実生活情報を提供する．昼食に何を食べるか迷っているときは「食事」の局面の実生活情報を提供する．このように，ユーザが求める実生活情報とは，ユーザのその時，その場所のコンテキストによって変化するため，ある局面に限定した実生活情報を提供することで，ユーザの実生活における活動をより一層支援できると考えている．本研究では「茨城県つくば市」に焦点を当てて実践的に研究を進めるため,「つくば市」を生活圏とするユーザの記事を対象とする．

本論文の構成を以下に示す．第 2 章で先行研究について述べ，本研究の位置付けを明確にする．第 3 章で本研究で扱う「実生活」の定義について詳細に述べた後，実生活に言及している記事の抽出法を提案する．第 4 章では提案した手法を実装し,「つくば市」を生活圏にするユーザの記事を用いて抽出精度を評価する．第 5 章で実験結果について考察と議論を行い，最後に第 6 章で本研究における結論を述べる．

2. 先 行 研 究

実生活情報の主要な部分は，ユーザの経験からなる．文書から経験を抽出する手法については，経験マイニングという研究が行われている．倉嶋ら [1] [2] は，人間の経験を｛状況，行動，主観｝からなる情報と捉え，文章中から｛時間，空間，動作，対象，感情｝を抽出する手法を述べている．また池田ら [3] や高野ら [4] は，ユーザの体験表現を 20 種類の品詞の組合せをルールとして定義し，文書中にルールに適合する文があった場合，体験表現として抽出している．これらの提案手法は，ブログなどの比較的長い文書を対象としたマイニング手法であり，一つ一つの文書が短い Twitter にそのまま適用することは出来ない．また，Twitter では主語や目的語が省略されることが多く，このことが経験マイニングを更に難しくしている．

有光ら [5] は，Twitter に断片的に投稿されているユーザ体験に関する記事を，体系化して検索する手法を提案している．この提案手法では，一つの体験が複数の行動の遷移であることに着目し，体験を順序関係を持ったキーワード列としてシステムに入力する．キーワード列からデータベース上の記事の内容関連度とコンテキスト関連度を計算し，入力に合ったユーザ体験コンテンツであるかどうかを判別する．Twitter の記事を体験単位で検索できるという利点があるが，その体験を表すキーワード列を自動的に発見することが現段階ではできていない．

上記以外にも，マイクロブログに関する研究は盛んに行われている．岩木ら [6] の提案手法では，ユーザと記事との近接度を過去の投稿における返信回数に基づいて計算し，感性辞書を適用することで，有用な記事の発見支援をする．Sakaki ら [7] は，Twitter ユーザをセンサーと想定し，地震などの現実世界でリアルタイムに起きるイベントを発見する手法を明らかにしている．松村ら [8] は，ある場所の名前で Twitter の記事を検索して収集し，その記事数から勢いを計測することで，その場所が盛り上がっているかどうかを判定するシステムを実装している．西田ら [9] は，ある Twitter 記事が着目する話題か否かを分類するため，ツイートの圧縮されやすさを応用した手法を明らかにしている．本岡ら [10] は Twitter の HashTag に着目し，入力した HashTag と類似するイベントを発見する手法を述べている．

経験マイニングに関する研究の多くは，ブログ記事のような比較的長い文書を対象としている．本研究の対象としている Twitter の記事は，一つ一つの文書が短く主語や目的語が省略されることが多い．また本研究はユーザの経験に限らず，ユーザの持っている生活における知識や見聞についても対象としている点で，先行研究とは異なる．

本研究では，Twitter の記事を対象とし，実生活に言及している記事を抽出する手法を提案する．抽出した記事から生活の局面に応じた属性を付与し，実生活データベースを構築する．更に，リアルタイムな実生活情報をユーザに提供するシステムの実装まで含めて提案している点で，より身近で新鮮な情報を提供することができるため，従来の研究とは大きく異なる．身近で新鮮な情報を提供することができるため，従来の研究とは大きく異なる．

3. 実生活情報の抽出方法の提案

3.1 実生活情報の定義

人々は，生活の様々な局面で情報を役立てている．例えば，通勤時の電車の遅延情報や，買物をする時のスーパーにおける安売りに関する情報，昼食時の飲食店に関する情報等がある．このような「人々の生活」にどのような局面があるかを，表 1 に示す．このような様々な局面がまとめられた，Wikipedia の「生活」ページ(注2)と「地域コミュニティ」ページ(注3)，Yahoo! カテゴリ(注4)を参考に，これらの 14 の局面を列挙した．本研究における実生活情報とは，表 1 にある 14 の局面を踏まえて，以下の二つの条件から判定する．

（1） 個人の生活に関する情報は，**実生活情報**である．

例えば，飲食店で出てきた料理の感想等の**食事**に関する情報や，自分の病状や通院等の**健康**に関する情報，買った商品の値段やその評価等の**消費**に関する情報がある．このような，個人の経験や意見，感想などの情報が，個人の生活に関する情報である．

（2） 生活圏に関する情報は，**実生活情報**である．

例えば，地域で行われる祭りや学校の運動会等の**行事**に関する情報や，電車の運行情報や道路の工事等の**交通**に関する情報，新規のレストランの開店や臨時休業等の**地域**に関する情報がある．このような，特定の場所でしか利用できない限定性の高い情報が，生活圏に関する情報である．

本研究は，Twitter に投稿された記事を情報源としたことから，以下の場合に対処する．

(注2)：http://ja.wikipedia.org/wiki/生活
(注3)：http://ja.wikipedia.org/wiki/地域コミュニティ
(注4)：http://dir.yahoo.co.jp/

（1） 文章を読んでも内容を汲み取れない記事

仮にその記事が投稿者本人にとって実生活情報であっても，投稿者以外の人間に内容が汲み取れないので，本研究では実生活情報として扱わない．

（2） リツイートされた記事、あるいはブログ等を引用している記事

投稿の形式が複雑になっているだけであり，記事中に実生活情報があることに代わりない．このため，本研究では実生活情報として扱う．

表 1 生活の局面の一覧

局面	説明
食事	料理，外食，調理方法，食べること
居住	住まい，宿泊先，家具，家事
服飾	服装，装飾，化粧，理容
健康	風邪，病気，怪我，通院，医療，ダイエット
交流	約束，勧誘，出会いに関すること
趣味	余暇や習慣的に行う行為，道楽，芸能，ホビー
行事	祭り，冠婚葬祭，企画，記念，イベント
消費	買物，注文，広告活動，割引，混み具合，値引き
災害	風水害，地震等，二次的被害
交通	道路，電車，移動手段
気象	天候，気温，湿度，風
地域	観光，案内，地図，地理，建造物
労働	職業，アルバイト，就職活動
学校	授業，勉強，宿題，レポート，テスト

3.2 記事の抽出方法

3.2.1 概　　要

実生活情報は，ユーザの経験だけでなく，地域に関する知識や見聞などの情報も含まれているため，品詞の組合せによるルールを用いて実生活情報を抽出することは難しい．また，いくつかの実生活らしい単語を用いて Twitter 上から検索する場合についても，3.1 節で述べたように実生活が含む話題の範囲は幅広い．そのため，数十種類の単語を定義し検索するだけでは，より多くの実生活に言及している記事を選択的に抽出できない．

そこで，本論文では実生活に言及している記事には実生活らしい特徴的な単語があると仮定し，実生活らしい単語を自動的に登録した**実生活辞書**を用いた，新たな記事の抽出法を提案する．

提案法を実現する実生活記事抽出システムの概要を図 1 に示す．この図は，「つくば市」を生活圏にするユーザの投稿記事を，Twitter API により収集し，実生活に言及している記事を抽出し，実生活データベースに格納する例を示したものである．図中の点線で囲っている部分が本論文で提案するシステムであり，抽出した記事は即座にシステムによって提示することで，リアルタイムな情報提供をする．実生活データベースには生活の局面や日時等の属性値も格納するため，ユーザが過去の実生活情報を知りたいときも柔軟に対応できる．

以降，3.2.2 節では実生活辞書の生成方法について述べる．3.2.3 節では実生活辞書による実生活情報の抽出方法を提案し，3.3 節で実生活データベースの構築手法について説明する．

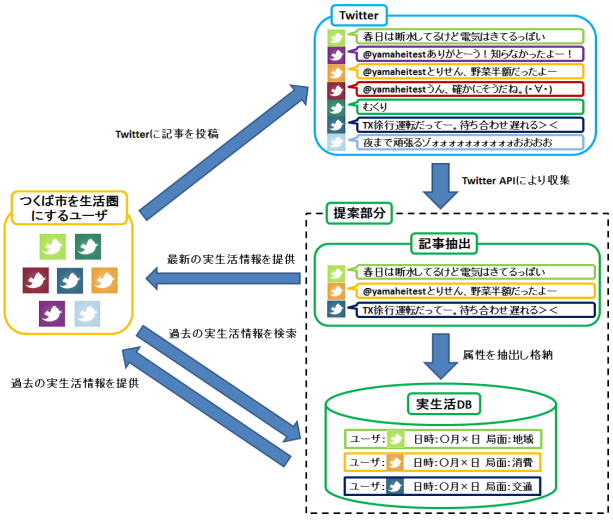

図 1 提案するシステムの概要

3.2.2 実生活辞書の生成方法

実生活に言及している記事は特徴的な単語を持つと仮定し，局面毎に特徴的な単語を集めた実生活辞書 $LD(asp)$ を以下の手順で生成する．このときの asp は，表 1 の 14 の局面である．

（1） 記事集合に対して，各記事が実生活情報と非実生活情報のいずれであるかを人手で分類する．実生活情報と分類した場合は，更に 14 の局面のいずれかを判定する．

（2） 分類した記事集合に対して，形態素解析器 MeCab[注5] で形態素解析し，{名詞，動詞，形容詞，形容動詞} のいずれかの品詞に該当する単語を抽出する．得られた単語集合を $T = \{t_1, t_2, \cdots, t_n\}$ とする．

（3） 単語集合 T について，実生活情報における局面毎に，個々の単語 t_i が出現する記事の出現頻度 $RF(asp, t_i)$ と，非実生活情報における単語 t_i が出現する記事の出現頻度 $NF(t_i)$ を求める．

（4） 単語 t_i の，その局面における**実生活らしさ** $L(asp, t_i)$ を，次式で求める．ある局面に出現する単語の頻度を，非実生活情報に出現する単語の頻度で除すことで，その局面における単語の特徴を表している．ある局面に頻出するが，非実生活には頻出しない単語ほど実生活らしさの値が大きくなる．

$$L(asp, t_i) = \frac{\log\left(RF(asp, t_i) + 1\right)}{\log\left(NF(t_i) + 1\right)} \tag{1}$$

（5） 単語の実生活らしさを正規化した値 $\bar{L}(asp, t_i)$ を，次式で求める．ある局面の単語の実生活らしさを，全ての局面の全ての単語の実生活らしさの総和で除すことで，実生活らしさを 0 から 1 の間で表す．

$$\bar{L}(asp, t_i) = \frac{L(asp, t_i)}{\sum_{j=1}^{14} \sum_{k=1}^{n} L(asp_j, t_k)} \tag{2}$$

(注5)：http://mecab.sourceforge.net/

（6） $\bar{L}(asp,t_i)$ が閾値 θ (0≦θ≦1) 以上なら，単語 t_i を $LD(asp)$ に加える．

閾値 θ は，4 章で行う実験で最適な値を選択する．

3.2.3 実生活辞書を用いた抽出

3.2.2 節の手順で生成した実生活辞書 LD_{asp} を用い，以下の手順で実生活に言及している記事を抽出し，どの局面についての情報かを分類する．

（1） 入力された未分類の記事 *Tweet* を形態素解析し，{名詞，動詞，形容詞，形容動詞} のいずれかの品詞に該当する単語を抽出する．得られた単語集合を $W = \{w_1,w_2,\cdots,w_n\}$ とする．

（2） 次式から *Tweet* の局面毎の実生活らしさ R を求める．LD は局面×単語で表される，実生活らしさの行列である．その局面に含まれていない単語の列には，0 が入る．W は 1 ×単語の行列であるので，転置行列をとることで，局面毎に W と一致する単語の実生活らしさの合計を求めている．

$$R = LD \times W^T \tag{3}$$

（3） 以下の条件を満たすとき，*Tweet* は**実生活に言及している記事**であるとする．

$$|R| > \theta \tag{4}$$

（4） 特に *Tweet* は，局面毎の実生活らしさ $R = \{r_1,r_2,\cdots,r_{14}\}$ について，最大値をとる局面の情報である．

3.3 実生活データベースの構築

3.2.3 節で述べた手法を用いて抽出した記事から，表 2 に示す属性を抽出する．記事 ID，ユーザ ID，投稿日時，本文については，Twitter API により抽出が可能である．

生活局面については，3.2.3 節により，一つの記事には 14 のいずれかの局面が付与されているので．その局面を抽出する．局面をデータベースの属性値として与えることで，その局面に限定した情報をリアルタイムに提供することができる．また，過去の実生活情報を検索する際も有用な属性となると考えられる．

動詞と形容詞については，記事の文章を形態素解析した時にその品詞に該当するものを全て抽出する．ただし，「する」や「いう」などの頻出する動詞は削除し，「サ変接続名詞」については動詞として抽出する．この 2 つの属性により，例えば「美味しい」で検索すると，美味しいレストランであったり，コンビニ等で売られている美味しい商品などに関する情報が，「閉店」で検索すると，店を閉じた飲食店に関する情報が得られると考えられる．

4. 評価実験

4.1 実験概要

3 章で提案した実生活情報の抽出手法を実装し，評価実験を行う．機械学習による実生活情報の抽出結果と比較することで，提案手法の有効性を確認する．以下，4.2 節では，提案手法と比較する機械学習について説明する．4.3 節では，実験用データについて説明した後，本実験で使用する評価尺度について述べる．テストデータから実生活情報が抽出できるか実験した結果を示す．4.4 節は，抽出した実生活情報に適切な局面を付与できるか実験により確かめる．実験方法について述べた後，局面毎の正答率を示す．

表 2 記事から抽出する属性

属性名	説明
記事 ID	記事に必ず付与される属性.
ユーザ ID	記事を投稿したユーザの ID.
投稿日時	記事を投稿した日時.
本文	記事の本文.
生活局面	表 1 の局面のいずれか.
動詞	本文中に存在する全ての動詞の原形.
形容詞	本文中に存在する全ての形容詞の原形.
実生活語	本文中に存在する実生活辞書にある全ての単語.

4.2 機械学習による抽出

提案手法の有効性を確認するために，機械学習による実生活情報の抽出をする．本研究では，高い分類性能と汎化能力を有しているという特徴から SVM を用いる．SVM は，教師あり学習を用いる識別手法の一つで，現在知られている多くの手法の中で，一番認識性能が優れた学習モデルの一つである．

実験ではツールとして libsvm(注6)を使用し，線形カーネル，RBF カーネルを用いてテストデータから記事を抽出する．SVM の素性は，記事を形態素解析して得られる，名詞，動詞，形容詞，形容動詞とする．これは，提案手法と同様の条件で実験をするためである．

4.3 実生活情報の抽出精度評価

4.3.1 実験用データ

本実験では，「つくば市」を生活圏にするユーザ 100 人を収集した．ユーザは，プロフィールや過去に投稿した記事から，現在も「つくば市」を生活圏にしているかを人手で選択した．これらのユーザは，過去に投稿した記事が数百件から数万件まで範囲が広いので，ユーザあたり上限 2,000 件を収集した．収集した記事の合計 155,403 件の中からランダムに 4,000 件抽出し，その記事が実生活に言及しているか否かを人手で分類した．また，Twitter の public timeline (注7)の記事についても，2010 年 10 月 24 日から 2011 年 4 月 23 日までの期間の記事を収集し，文章が日本語で書かれたものをランダムに 4,000 件抽出し同様に分類し正解データとした．これらの各 4,000 件の記事を対象に実験を行う．表 3 に示すように，学習用の記事と，テスト用の記事にランダムに分割する．実生活の局面毎の記事数を表 4 に示す．「学習用」とは SVM では教師データのことを意味し，提案手法では実生活辞書生成用データのことを意味する．

4.3.2 実験方法

提案手法と SVM の手法により，実生活に言及している記事が正確に抽出できているかを評価する．提案手法の有効性を議

(注6)：http://www.csie.ntu.edu.tw/ cjlin/libsvm/

(注7)：世界中のユーザの公開されている記事のタイムラインのこと

表 3　実験用データの記事数

	つくば		public	
	学習用	テスト用	学習用	テスト用
実生活	1,384	465	1,070	264
非実生活	1,616	535	1,930	736
合計	3,000	1,000	3,000	1,000

表 4　局面毎の記事数（つくば市）

	つくば		public	
局面	学習用	テスト用	学習用	テスト用
食事	116	50	147	34
居住	63	25	31	5
服飾	25	9	30	14
健康	74	35	62	18
交流	97	31	62	8
趣味	195	54	120	26
行事	112	35	104	25
消費	140	63	168	47
災害	47	9	2	3
交通	52	18	40	10
気象	89	18	68	18
地域	99	49	98	21
労働	102	25	86	18
学校	173	44	52	17
合計	1,384	465	1,070	264

論するには，抽出した記事集合がどれだけ正解しているかという正確性と，抽出した記事集合が全ての正解のうち，どれだけ正解を含んでいるかという網羅性の 2 つの観点からの評価が必要となる．本論文では，正確性を**適合率** (*precision*)，網羅性を**再現率** (*recall*)，適合率と再現率の調和平均である **F 値** (*F-measure*) によって提案手法の抽出精度を評価する．それぞれの計算方法について，以下に示す．正解記事数とは，実生活に言及している記事のことである．

$$precision = \frac{\text{抽出した正解記事数}}{\text{抽出した記事数}} \tag{5}$$

$$recall = \frac{\text{抽出した正解記事数記事数}}{\text{全ての正解記事数}} \tag{6}$$

$$F\text{-}measure = \frac{2 \times precision \times recall}{precision + recall} \tag{7}$$

提案手法による実験では，実生活辞書を生成する際に用いる閾値 θ を 0 から 0.004 の間で変動させることで，実生活辞書に登録する単語数を変化させながら，それぞれの値の変化を分析する．実験は，以下の 4 通りを行う．

項目 1　つくば市の実生活辞書を用いて，つくば市の記事集合から実生活情報と判断された記事を抽出.

項目 2　つくば市の実生活辞書を用いて，public timeline の記事集合から実生活情報と判断された記事を抽出.

項目 3　public timeline の実生活辞書を用いて，つくば市の記事集合から実生活情報と判断された記事を抽出.

項目 4　public timeline の実生活辞書を用いて，public timeline の記事集合から実生活情報と判断された記事を抽出.

SVM の手法による実験では，SVM が実生活と分類した記事を実生活情報として抽出し，提案手法と同様に，適合率（式 (5)），再現率 (式 (6))，F 値 (式 (7)) を計算する．SVM の手法では項目 1 の組合せについて実験をする．

4.3.3　実 験 結 果

提案手法により，F 値が最大となった時の閾値，適合率，再現率，F 値，実生活辞書の単語数について，表 5 に示す．また，各実験の適合率，再現率，F 値の遷移を図 2 から図 5 に示す．

提案手法における項目 1 と，SVM の手法の実験結果との比較を表 6 に示す．SVM の各パラメータについては，一般的な値である Gamma は 0.5，パラメータ C は 1.0 を用いた．

表 5　F 値を最大とする条件（提案手法）

項目	閾値	適合率	再現率	F 値	単語数	図番号
1	0.0014	**0.684**	**0.819**	**0.746**	**3,736**	2
2	0.0016	0.407	0.805	0.541	3,592	3
3	0.0012	0.645	0.817	0.721	4,139	4
4	0.0021	0.567	0.565	0.566	717	5

表 6　提案手法と SVM の手法の比較

手法	適合率	再現率	F 値
提案手法	0.684	**0.819**	**0.746**
SVM(線形)	**0.733**	0.570	0.656
SVM(RBF)	0.624	0.518	0.566

4.4　実生活情報の局面正答率の評価

4.4.1　提案手法を用いた実験方法

提案手法により抽出した記事には，抽出の過程で自動的に 14 のいずれかの局面を付与している．その付与された局面の正答率を評価する．実験では，全ての局面に記事数が十分存在するという前提をおき，つくば市の実生活辞書で，つくば市の記事に局面を付与する．

実験は，全ての局面から 10 件ずつ記事をランダムに選択し，合計 140 件をテストデータとする．残りの記事は，実生活辞書を生成するために用いる．実生活辞書を生成する際の閾値は，4.3.2 節の，項目 1 の実験で F 値を最大とする値 0.0014 とした．

テストデータに対して，全ての局面の実生活らしさを算出する．実生活らしさの値が大きい上位 3 位以内の局面について，人手で付与した局面との正答率を計算する．

4.4.2　SVM を用いた実験方法

一般的な機械学習として知られる SVM を適用することで，記事は 14 の局面と非実生活の合計 15 個の局面いずれかに分類される．その際に，個々の記事には 15 個の局面全てに値が付与され，値が大きいほどその局面に適合していることを表している．この値の降順に局面を並び替え，上位 3 位以内の局面について，人手で付与した局面との正答率を計算する．テストデータと学習データは，提案手法と同じものを用いる．

4.4.3　実 験 結 果

提案手法によってテストデータに付与された各局面の上位 3 位以内の正答率を，図 6 に示す．ここで 1 位とは，人手で付与

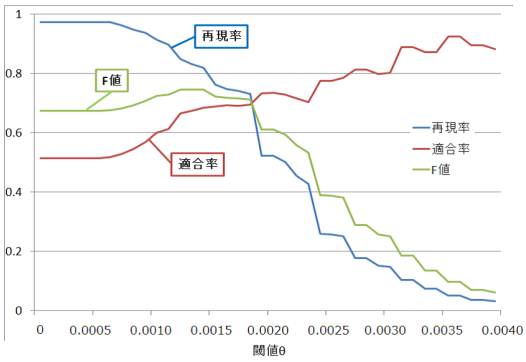

図 2　抽出実験（つくば市の辞書－つくば市の記事）

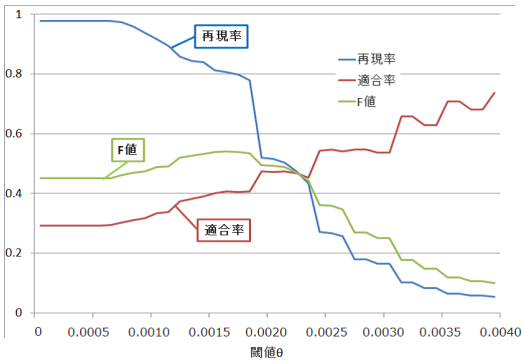

図 3　抽出実験（つくば市の辞書－ public timeline の記事）

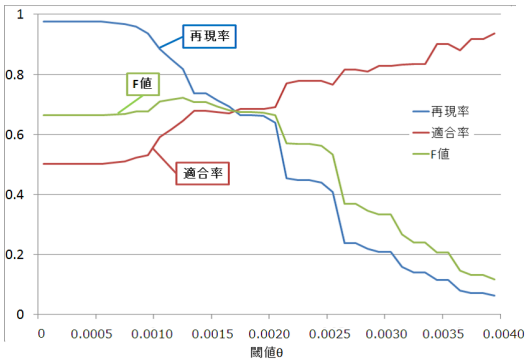

図 4　抽出実験（public timeline の辞書－つくば市の記事）

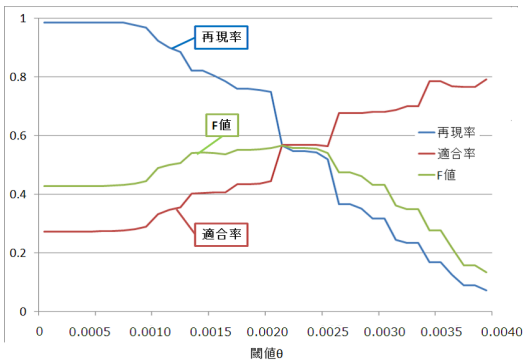

図 5　抽出実験（public timeline の辞書－ public timeline の記事）

した局面と，算出した実生活らしさの値が 1 位の局面と一致した時の正答率を求めている．2 位は，人手で付与した局面と，1 位か 2 位の局面と一致した時．3 位は，人手で付与した局面と，1 位か 2 位か 3 位の局面と一致した時の正答率である．各順位における平均は，それぞれの場合のマクロ平均値である．各局面について，1 位で最大正答率となったときの閾値についてまとめたものを，表 7 に示す．閾値が連続する場合は「〜」で表し，閾値が不連続な場合には「,」で表している．SVM を用いて局面を付与した場合の結果を図 7 と図 8 に示す．SVM では，使用するカーネル関数によって精度が異なることから，線形カーネル（図 7）と RBF カーネル（図 8）での結果を示す．SVM の各パラメータについては，一般的な値である Gamma は 0.0315，パラメータ C は 8.0 を用いた．

5.　考　　察

5.1　抽出精度の評価に関する考察

表 5 より，提案手法について，項目 1 と項目 3 では F 値が 0.7 以上を示しているが，項目 2 と項目 4 では F 値が 0.6 以下となっている．項目 1 と項目 3 では，つくば市の記事を対象に実生活情報を抽出していることが共通しており，項目 2 と項目 4 では，public timeline の記事から実生活情報を抽出している．つくば市と public timeline のテスト用データ各 1000 件について，一つの記事から抽出できる名詞，動詞，形容詞，形容動詞の品詞に該当する単語数について，表 8 に示す．この表は，つくば市と public timeline の記事について，実生活と非実生活に分けて平均単語数を算出したものを示している．表より，public timeline の記事について，実生活の平均単語数が 7.71 であるのに対し，非実生活の平均単語数は 22.27 と非常に多かった．このように単語数の多い記事は，実生活に言及していない記事でも，実生活辞書に存在する単語を含んでしまう可能性が高くなる．このことから，public timeline の記事では，閾値が低い時に非実生活の記事を多く抽出し，つくば市の記事の場合に比べて適合率が低くなったと考えられる．

また，いずれの項目においても，閾値 0.002 付近で再現率が大きく下がっていることが分かる．これは，実生活辞書に登録されている多くの単語が，同一の実生活らしさの値となっていることが原因である．この問題については，実生活らしさの値の算出式を改良することで対応できると考えている．例えば，単語数の少ない記事でも実生活情報であれば，その記事を構成する各単語の実生活らしさは大きいと考える．このことから，単語の頻度を計算する時に，記事の単語数の逆数を重みとして計算することで対応できると考えている．

表 6 より，提案手法と SVM を用いた手法を比較すると，線形カーネルを用いた場合に適合率が提案手法を上回っていた．提案手法では，実生活辞書に存在する単語が一つでも記事中にあれば実生活情報として抽出するため，非実生活の記事も抽出することも多いが，より多くの実生活情報を抽出できる．そのため，SVM と比べて適合率で下回っているが，再現率で大きく上回ったのだと考えられる．また，F 値でも SVM より上回っていた．

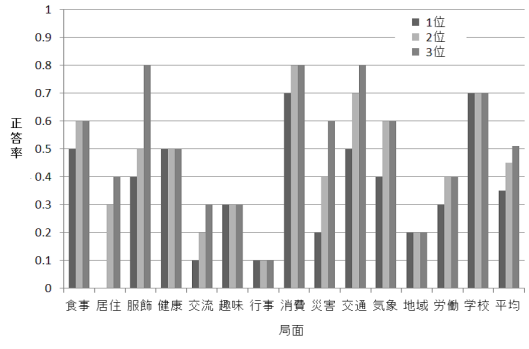

図 6　局面毎の正答率（提案手法）

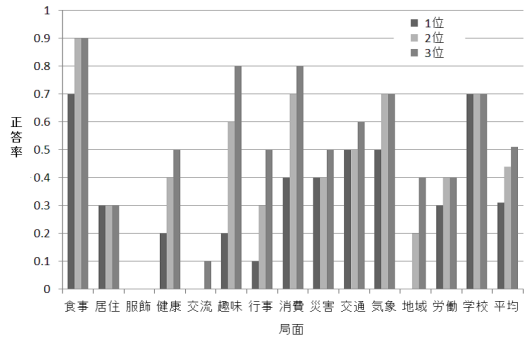

図 7　局面毎の正答率（SVM-線形カーネル）

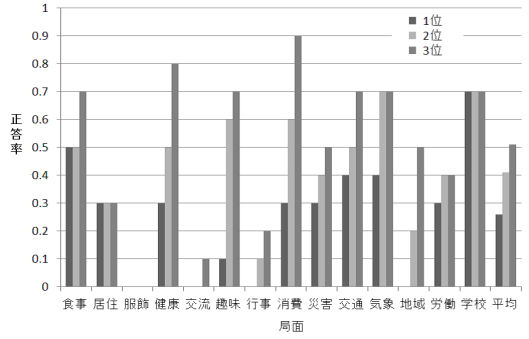

図 8　局面毎の正答率（SVM-RBF カーネル）

表 7　正答率が最大となる閾値 （提案手法）

局面	正答率	閾値
食事	0.50	0.0011〜0.0015
居住	0.10	0.0016〜0.0026
服飾	0.40	0.0011〜0.0015
健康	0.50	0.0011〜0.0014
交流	0.10	0.0007，0.0011〜0.0015，0.0021〜0.0026
趣味	0.30	0.0007，0.0012〜0.0015
行事	0.40	0.0009
消費	0.70	0.0008〜0.0015
災害	0.30	0.0006〜0.0012
交通	0.50	0.0011〜0.0015
気象	0.60	0.0008〜0.0010，0.0015，0.0021〜0.0023
地域	0.40	0.0008
労働	0.40	0.0000〜0.0010，0.0018〜0.0020
学校	1.00	0.0004〜0.0006

以上のことから，提案手法では単語数の少ない記事の実生活情報を抽出する場合に有効性が高いと考えられる．単語数が多い記事に対して提案手法を適用する場合においても，記事の単語数を考慮して，実生活情報として抽出する閾値を動的に変化させることで対応できると考える．例えば，単語数に比例して閾値を高くするなどの方法が考えられる．

表 8　テスト用データの平均単語数

	つくば	public
実生活	13.68	7.71
非実生活	15.68	22.27

5.2　局面正答率の評価に関する考察

図 6 より，提案手法において，服飾，消費，交通，学校の局面について，3 位の時に正答率が 0.7 以上という値を示した．

提案手法と SVM を用いた手法を比較すると，図 7 と図 8 より，提案手法が 1 位，2 位の正答率の平均値で SVM を上回っている．特に服飾の局面では提案手法が 3 位で 0.8 の正答率を示しているのに対して，SVM では 3 位までに一度も正答ができていない．表 4 より，服飾の局面の記事数は 34 件であり，テスト用データである 10 件を除いた 24 件が SVM の学習データとなっている．全ての局面において服飾が学習データの数が最も少なく，このような理由から，SVM を用いた手法では服飾の局面で一度も正答出来なかったと考えられる．逆に，提案手法で 0.3 と低い正答率となっている趣味の局面は，全ての局面において学習データ数が最も多く，SVM を用いた手法では正答率が高くなっている．このように，SVM を用いた手法では，正答率の高さは学習データの記事数に依存していると考えられる．

しかし，交流，行事，地域などの局面については，学習データが十分あるにも関わらず，3 位でも高い正答率を示していない．SVM だけでなく提案手法についても同様である．これらの局面については，他の局面に誤分類されやすいと考えられる．実際に Tweet 記事を見てみると，交流であれば誰かと食事をしたことに関する記事や，友人と休日に遊びに行ったことに関する記事などがあり，これらの記事は食事や趣味などの局面の話題を含んでいる．趣味に関しても，自分が応援している芸能人のイベントに参加するといった記事や，好きなアーティストが発売した CD を買うといった記事などがあり，行事や消費の局面の話題を含むことがある．このように，正答率の低い局面については，別の局面の話題を含みやすいという特徴がある．この問題については，一つの記事に複数の局面を付与することで解決できると考えている．友人と休日に遊びに行ったことに関する記事については，交流と趣味の局面を付与する．好きなアーティストが発売した CD を買ったという記事については，趣味と消費の局面を付与する．このように，複数の局面を付与することで，ユーザが求める局面に応じて実生活情報を提供できると考えている．また，今ある 14 の局面の内，いくつかを併合して適切な局面へと編集していく必要もあると考える．

提案手法について，閾値 0.0014 のとき正答率の低かった行

事，地域，労働の局面は，表 7 にあるように 1 位のとき最大で 0.4 の正答率を示していることから，今回用いた閾値が適切でなかったことが分かる．実生活辞書を生成する際は，全ての局面に同一の閾値を設けて単語を登録していくのでなく，局面毎に適切な閾値を設定することで，正答率が向上すると考えられる．

以上のことから，提案手法では学習データの数に依存せずに正しい局面を付与できるが，局面毎に適切な閾値を設定する必要がある．SVM を用いた手法では，局面毎に多くの学習データを用意することで，正答率が向上すると考えられる．

5.3 実装したシステム

提案手法により，実際に実生活情報をユーザに提供するため，システムを実装した．システムは Web ブラウザの Google Chrome(注8) 上で動作する拡張機能(注9)として実装した．実行画面を図 9 に示す．

システムは，画面右上の拡張機能のアイコンをクリックすることで実行することができる．実行中は，提案手法で抽出した最新の Twitter 記事を，左側にテロップしながら流す．記事に付与された局面は左上に画像アイコンとして示し，その横に記事を投稿したユーザ名，投稿日時，記事中の実生活辞書に存在する単語を示している．また，その横の「Next」をクリックすることで，テロップ中の記事をスキップして最新の記事を表示することができる．ユーザがある局面に限定して実生活情報を得たい場合は，オプションページから指定が可能である．指定した場合は，その局面における最新の記事を表示する．ユーザがまた見たいと思った実生活情報は，右側のお気に入りボタンを押すことによって，オプションページからいつでも閲覧可能である．

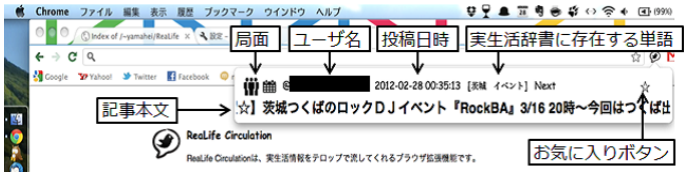

図 9　システムの実行画面

6. おわりに

本論文では，ユーザの実生活における活動に有用な情報を提供することを目的に，Twitter 上から実生活に言及している記事を抽出する技術を確立した．抽出した記事に対して,「食事」や「交通」等の生活の局面などの属性を付与し，実生活データベースを構築した．このような属性を付与することで，生活の局面に応じた実生活情報をリアルタイムにユーザに提供できる．また，本研究では「茨城県つくば市」に焦点を当てて実践的に研究を進めるため,「つくば市」を生活圏とするユーザの記事を対象とした．

提案手法では,「実生活」を 14 の生活における局面からなる情報と定義した．局面毎に実生活に特徴的な単語を集めた「実生活辞書」を生成し，実生活辞書に基づいた抽出法を提案した．抽出した記事には，14 のいずれかの局面を付与した．

つくば市を生活圏にするユーザから収集した記事 4,000 件と，public timeline から収集した記事 4,000 件を対象に実験を行った．実験では，3,000 件を実生活辞書を生成するために利用し，残りの 1,000 件をテストデータとして実生活情報の抽出を行った．つくば市の辞書でつくば市の記事を抽出する場合において，F 値が最大 74.5%という結果を示した．また，実生活に言及している記事に正しい局面が付与されるか実験を行った．その結果，5 つの局面で 50%を超える結果を得られた．SVM の手法との比較をし，実生活情報の抽出，実生活に言及している記事に対する局面の付与のどちらについても，提案手法が高い評価値を示し，有効性を確認できた．

今後の課題は，単語数と局面に考慮して動的に閾値を変動させる手法を実装し，抽出精度の変化を分析すること，実生活情報へ複数の局面を付与することが挙げられる．

謝　　辞

本研究の一部は科研費 (21500091) の助成を受けたものである．ここに記して謝意を示す．

文　　献

[1] 倉島健, 藤村考, 奥田秀範．大規模テキストからの経験マイニング．電子情報通信学会 第 19 回データ工学ワークショップ（DEWS2008）論文集 A1-4, 2008.

[2] Takeshi Kurashima, Taro Tezuka, and Katumi Tanaka. Extracting and geographically mapping visitor experiences from urban blogs. *WISE 2005*, pp. 496–503, 2005.

[3] 池田佳代, 高村大地, 奥田秀範．体験表現を手がかりにした blog の体験情報の抽出．電子情報通信学会 第 18 回データ工学ワークショップ（DEWS2007）論文集 A8-1, 2007.

[4] 高野太希, 井上潮．文章構造に基づいて blog からの体験情報抽出方法の提案．第 3 回データ工学と情報マネジメントに関するフォーラム（DEIM Forum 2011）A4-2, 2011.

[5] 有光淳紀, 馬強, 吉川正俊．ユーザ体験指向の twitter 検索手法．第 3 回データ工学と情報マネジメントに関するフォーラム（DEIM Forum 2011）F5-2, 2011.

[6] 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．第 1 回データ工学と情報マネジメントに関するフォーラム（DEIM Forum 2009）A6-6, 2009.

[7] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes twitter users: Real-time event detection by social sensors. *In Proceedings of 18th International World Wide Web Conference (WWW2010)*, pp. 851–860, 2010.

[8] 松村飛志, 安村通晃．街に着目した twitter メッセージの自動収集と分析システムの提案と試作．情報処理学会 インタラクション 2010, 2p, 2010.

[9] 西田京介, 坂野遼平, 星出高秀．データ圧縮による twitter のツイート話題分類．第 3 回データ工学と情報マネジメントに関するフォーラム（DEIM Forum 2011）A1-6, 2011.

[10] 本岡亮, 湯元高行, 新居学, 高橋豊, 角谷和俊．Twitter ハッシュタグを用いた類似イベント検索．第 3 回データ工学と情報マネジメントに関するフォーラム（DEIM Forum 2011）A1-5, 2011.

(注8)：http://www.google.co.jp/chrome

(注9)：http://code.google.com/chrome/extensions/docs.html