DBSJ Journal, Vol.10, No.1
June 2011

データ圧縮によるツイート話題分類

Tweet-Topic Classification using Data Compression

西田 京介 ♥　坂野 遼平 ♦
藤村 考 ♠　星出 高秀 ♣

Kyosuke NISHIDA　Ryohei BANNO
Ko FUJIMURA　Takahide HOSHIDE

本研究では，**Twitter** に日々投稿される膨大なツイート（口語的で短く，リアルタイム性の高いテキスト）の中から着目する話題に関するツイートを分類するため，ツイートの圧縮されやすさを応用した手法を提案する．評価のためハッシュタグ付きの日本語ツイートを用いて実験を行い，圧縮法として **gzip** で用いられる **Deflate** を利用した提案手法が，形態素や文字 N**-gram** を素性とした **confidence-weighted linear classification** よりも優れた分類精度を実現したことを示す．

We propose a new method that uses data compression for classifying topics of tweets (conversational, short, and real-time messages) in Twitter. Experiments with Japanese tweets assigned hashtags demonstrate that our proposed method using the Deflate data compression method, which gzip uses, achieved higher precision and recall rates than the confidence-weighted linear classification method, which used the character n-grams or morphemes of tweet texts as input features.

1. はじめに

マイクロブログサービス，特に Twitter は，世の中の「今」を知るための情報基盤として驚くべき成長を遂げている．利用者が主に自身の状況や雑感などを短いテキスト（Twitter ではツイートと呼ばれる上限 140 文字のテキスト）で投稿する形式が更新の容易さと高いリアルタイム性を産み出しており，2011 年 2 月には世界中で 1 日あたり約 1 億 4000 万ツイートが投稿された[1]．また，ハドソン川で発生した米旅客機の不時着事故（2009 年 1 月）の第一報が Twitter の投稿であった様に，Twitter は誰もが情報発信・情報収集できる非常に重要なサービスに発展している [1].

このような状況の中で，日々生成される膨大な情報量の中から，利用者にとって有益な情報のみを収集する需要が高まっている．現状の Twitter における主な情報収集手段にはフォロー，キーワード検索，ハッシュタグ検索がある．

まず，フォローとは，自身の友人や，著名人，政治家，企業，メディアなど，興味のあるアカウントを登録することで，他ユーザが発信した情報の収集を行うものである．フォローは重要な情報収集行動であるが，ユーザレベルでの情報収集であり，ツイートレベルでの細かな情報収集能力は有しない．一方で，キーワード検索は，ツイートレベルで興味のある情報を収集できるが，入力したクエリが含まれているツイートしか収集できず，検索の網羅性は低くなる．

そこで，ハッシュタグと呼ばれる，話題を明示的に表現しグループ化する「#」から始まる文字列による検索の利用が広まっている．例えば，ユーザは「育児」に関するツイートに対して，#ikuji や#kosodate というハッシュタグを自発的に挿入することで,「育児」に関するツイートであることを明示し,「育児」について興味を持つ他のユーザと情報を共有しやすくなる．しかし，ハッシュタグは自動的に付与されるものではないため，ハッシュタグのみで着目する話題に関するツイートを全て収集することはできない．

本研究では，着目する話題に関するツイートをより網羅的に収集するため，ツイートの話題分類に取り組む．ツイートの，他のメディアのテキストとは異なる特性には，

1. テキストが短い
2. リアルタイム性が高い（最新情報を多く含む）
3. 新語を多く含む
4. 口語・俗語・文法の誤りを多く含む

などがあり，特に日本語ツイートの分類においては形態素解析の精度が低下するため，単純に bag-of-words を素性とした機械学習の適用では高い分類精度を得ることが難しい．そこで，我々は，形態素解析に依存せず，学習対象の変化に素早く追従可能なアルゴリズムとして，データ圧縮によるテキストの圧縮されやすさを応用した分類を提案する．

2. ハッシュタグ付きツイートの解析

本章では，着目する話題に関するツイートを収集する手段のうち，現在の主流であるハッシュタグについて解析を行った結果を示す．

2.1 解析対象

2010 年 12 月 26 日に Twitter 社が提供する Streaming API[2] の statuses/sample（Gardenhose レベル，全ツイート

♥ 正会員　日本電信電話株式会社　NTT サイバーソリューション研究所　nishida.kyosuke@lab.ntt.co.jp
♦ 非会員　北海道大学大学院情報科学研究科複合情報学専攻　r_banno@complex.ist.hokudai.ac.jp
♠ 正会員　日本電信電話株式会社 NTT サイバーソリューション研究所　fujimura.ko@lab.ntt.co.jp
♣ 非会員　日本電信電話株式会社 NTT サイバーソリューション研究所　hoshide.takahide@lab.ntt.co.jp

[1] http://blog.twitter.com/2011/03/numbers.html
[2] http://dev.twitter.com/pages/streaming_api_methods

表 1　収集したツイートの統計（2010/12/26 00:00–23:59 GMT）．日本語ツイートはツイートテキストに日本語が含まれるもの．

Dataset	Number of Tweets	
	Worldwide	Japanese
sample	8,336,022	1,712,422
filter1 (#ikuji)	709	709
filter2 (#eiga)	1,086	1,084
filter3 (#seiji)	3,393	3,384
filter4 (#m1gp)	86,738	86,106

表 2　ハッシュタグが付与されたツイート数（sample データセット）．

Domain	Number of Tweets with Hashtags
Worldwide	902,692
Japanese	105,699

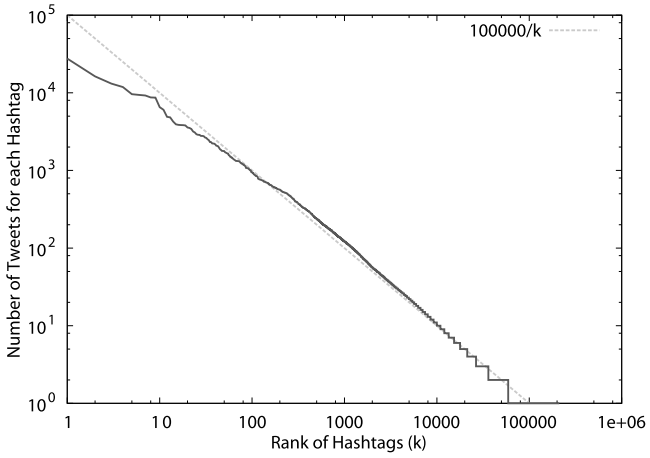

図 1　ハッシュタグ毎のツイート数の分布（sample データセット）．

の約 10% が収集可能）と statuses/filter（指定した文字列が含まれるツイートが収集可能）を利用して，表 1 に示すツイートを収集した．

2.2　解析結果

初めに，表 2 に sample データセットにおけるハッシュタグが付与されたツイート数を示す．ここから，全ツイートの約 10.8%，日本語ツイートに限ると約 6.17% にしかハッシュタグが付与されておらず，ハッシュタグだけでは話題抽出が十分に行えないことが分かる．

次に，図 1 に示すハッシュタグ毎のツイート数の分布から，1 日分のツイートだけでも 100,000 種類を超える，非常に多くのハッシュタグが存在していることがわかる（なお，ハッシュタグ毎のツイート数は Zipf の法則に従っている）．ハッシュタグはユーザが自由に作成・付与することができるため，#ikuji と#kosodate や，#m1 と#m1gp の様に同じ話題を表すハッシュタグが複数個存在しており，着目する話題に関するツイートの収集を難しくする一要因となっている．

また，図 2 と図 3 に示す様に，時間帯によって各ハッシュタグのツイート数は大きく変化することがある．特に，#m1gp のようにテレビ番組の実況用途で使われるハッシュタグは，1 時間毎のツイート量の比が大きく変化する．

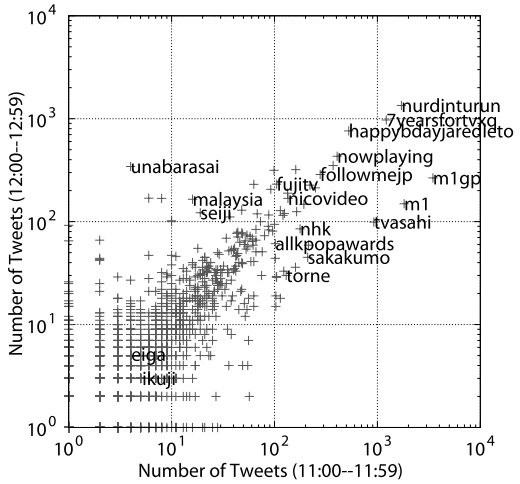

図 2　ハッシュタグ毎のツイート数の時間変化（2010/12/26 11:00–11:59 GMT vs. 2010/12/26 12:00–12:59 GMT，sample データセット）．

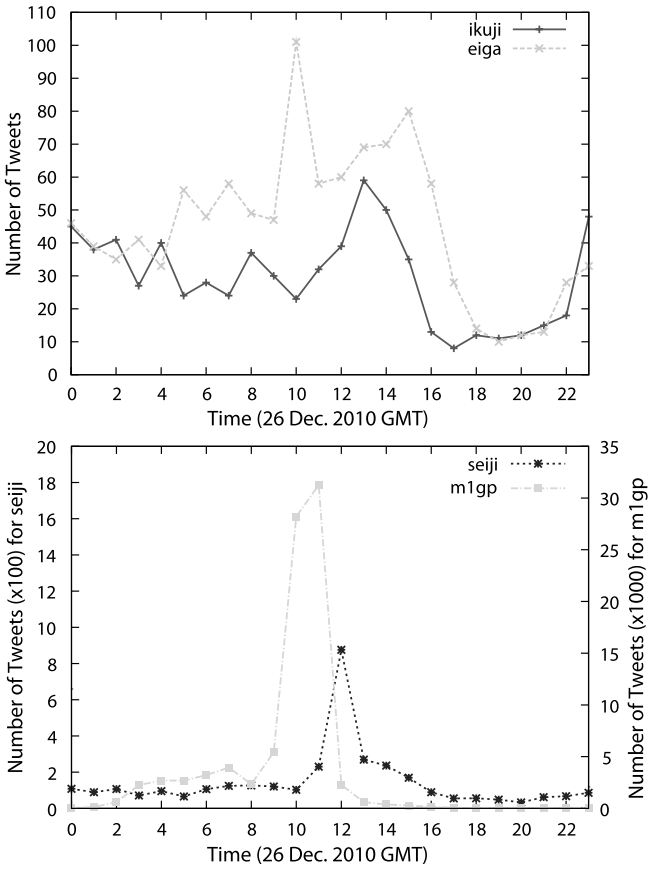

図 3　ハッシュタグ毎のツイート数の時間変化（filter データセット）．

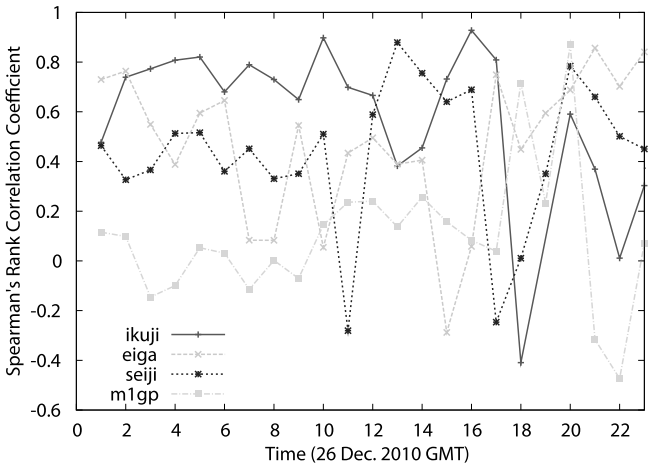

図 4　$t-1$ 時台と t 時台のツイート中における文字 3-gram 出現頻度に関する，Spearman の順位相関係数の時間推移（filter データセット）.

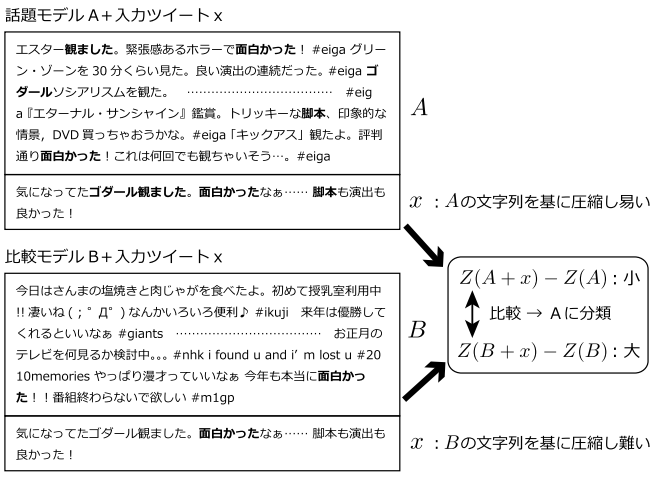

図 5　提案手法の概念図.

最後に，図 4 に，$t-1$ 時台のツイート中の文字 3-gram の出現頻度と，t 時台のツイート中の文字 3-gram の出現頻度に関する，Spearman の順位相関係数の時間推移を示す．なお，順位相関係数は，$t-1$ 時台と t 時台で合計して 10 回以上出現した文字 3-gram のみを対象として計算した．解析結果より，#m1gp の様にリアルタイム性の高い（テレビ番組の実況用途で用いられる）ハッシュタグは，他のハッシュタグと比べて，ツイート内容も時間と共に大きく変化していく傾向が強いことが分かる．

以上の解析結果より，ツイートの話題分類を行うためには，同時刻における話題量の偏り，話題量・内容の時間変化について十分に考慮しなければならないと考える．

3.　提案手法

データ圧縮は，データの冗長性を削減することで，データの転送や蓄積の際の資源を節約する目的で本来は利用されている．しかし，近年では，データの類似性に基づく分類手段としてデータ圧縮を応用する研究が進んでいる [2, 3, 4]. その基本的な概念は，あるデータ x が情報源となる他のデータ A を基に十分圧縮できる場合，x と A は類似しているというものである．データ圧縮は，テキストの言語に依存せずに（日本語テキストの形態素解析を行わず）高い分類性能を実現できる手法として期待できる．なお，スパムフィルタリングにおいては，従来の機械学習に基づくシステムよりもデータ圧縮が効果的であるという報告が為されている [4].

提案手法では，新しいツイートを，着目する話題に関するツイートの集合（話題モデル）と，それ以外のツイートの集合（比較モデル）の両方を基にして圧縮する．このとき，話題モデルを基にした方が圧縮されやすい場合に，新しいツイートは着目する話題に関連する可能性が高いとみなす．具体的には，指定した文字列（キーワード，ハッシュタグ，URL など）が含まれるテキストを時間順に連結したものを話題モデル A，それ以外のテキストを時間順に連結したものを比較モデル B と定義したとき，入力ツイート x の圧縮されやすさ $C_A(x)$ と $C_B(x)$ を，Benedetto らの手法 [2] に基づき以下のように計算する．

$$C_A(x) = Z(A+x) - Z(A) \tag{1}$$
$$C_B(x) = Z(B+x) - Z(B) \tag{2}$$

ここで，$A+x$ はテキスト A の後に x を連結したものであり，$Z(\cdot)$ は，入力テキストの圧縮後サイズを返す関数である．なお，**2.** の解析で得た知見に基づき，話題モデル A と比較モデル B はツイート量の偏りと内容の時間変化に対応するため，それぞれ最新の N ツイートから構築する．

そして，ツイートの分類スコア

$$f(x) = \frac{C_A(x)+\gamma}{C_B(x)+\gamma} \tag{3}$$

を計算し，$f(x)$ が θ よりも小さいときに，x が指定した話題に関するツイートであると分類する．ここで，γ はスムージングパラメータである．図 5 に提案手法の概念図を示す．

式 (1) で定義される $C_A(x)$ は，A に対する x の条件付の Kolmogorov 複雑性（データの複雑さを，そのデータを出力可能なアルゴリズム長の最小値で記述する指標 [5]）$K(x|A)$ を，圧縮プログラム $Z(\cdot)$ を用いて近似したものと見ることができる [3]. つまり，$A+x$ を記述するには，まず A を記述し，それを用いて x を記述するのが妥当な方法であるから，$K(A+x) = K(A) + K(x|A)$ が成り立っていると考えられ，Kolmogorov 複雑性 $K(\cdot)$ を実際の圧縮プログラム $Z(\cdot)$ で近似[3]することにより，式 (1) が得られる．

なお，圧縮アルゴリズム $Z(\cdot)$ には，gzip で用いられる Deflate（LZ77 とハフマン符号の組合せ）[6]，prediction by partial matching（PPM）アルゴリズム [7]，dynamic Markov compression（DMC）アルゴリズム [8] などが使用可能である．

[3] Kolmogorov 複雑性は一般に計算不能である．

4. 評価実験

提案手法の分類性能を評価するため，新しいツイートが着目話題（指定文字列）に関するツイートか否かを分類する実験を，指定文字列をハッシュタグとして実施した結果について示す．

4.1 実験設定とデータセット

表 1 に示す sample/filter データセットのうち，ハッシュタグが 1 つだけ付与されている日本語ツイートのテキストを用いて，sample データセットと filter データセットを組み合わたデータセットを 4 種類（それぞれ，ハッシュタグごとに ikuji, eiga, seiji, m1gp と呼ぶ）作成した．本データセットでは，ユーザ名などツイートのテキスト以外の情報は一切用いていない．なお，リツイート（他のユーザのツイートを再投稿すること，RT と省略される）に対する分類は，引用文がデータ圧縮を利用する提案手法にとって大きく有利に働くため，公式 RT（コメントの挿入不可）・非公式 RT（コメントの挿入可能）を問わず，作成したデータセットに含めていない．また，テキストの文字コードは UTF-8 とした．

各データセットはランダムに 5 分割し，1 つをテストデータ，残り 4 つを学習データとした 2 クラス分類実験を 5 回繰り返して手法の性能評価を行った（5-fold cross validation）．なお，分類器に対しては，学習データとテストデータを混在させてツイートの投稿時刻順に逐次的に与え（オンラインで学習とテストを行う），テストデータからはハッシュタグを削除した．

4.2 実験結果

提案手法と，現在のオンライン学習器の中で最も性能の良いものの一つである confidence-weighted linear classification（CW）[9] について性能を比較した．ここで，提案手法は，話題モデルと比較モデルをそれぞれ最新の 200 ツイートから構築し，$\gamma = 30$ と設定した．そして，圧縮アルゴリズム $Z(\cdot)$ には，Deflate（Ruby の Zlib::Deflate ライブラリ[4]の実装）を利用した．また，CW の素性には，文字 2-gram，文字 3-gram，形態素の 3 種類を利用した．なお，形態素は，MeCab 0.98[5]により名詞・動詞・形容詞と判定されたものとした．形態素解析に用いる辞書はオリジナルの IPA 辞書を利用した．

まず，表 3 に示す seiji データセットにおける提案手法のツイート分類例より，#seiji 以外の他のハッシュタグが付与されたツイートであっても，ツイートの内容に基付いて正しいスコア $f(x)$ を付与できたことが分かる．また，式 (3) におけるスムージングパラメータ γ の導入により，着目する話題（政治）に関するツイートのうち，テキストが長いものに対して低いスコア $f(x)$ を付与する傾向が生まれる．これにより，情報量の多いツイートを優先して精度良く分類することが可能になる．

次に，各データセットに対して，分類閾値（θ）を変更しながらテストデータに対する識別率と再現率を評価した．図 6 に示す結果より，データ圧縮を用いる提案手法が機械学習器 CW よりも優れた識別率と再現率を実現したことがわかる．特に，ikuji, eiga, seiji データセットでは，着目するハッシュタグが付与された学習ツイート数が少ないため，CW の分類性能が非常に悪くなっていた．これに対して，データ圧縮による提案手法は，着目する話題に関するツイート数が少ない場合でも，高い識別率と再現率を実現できた．また，テレビ番組に関する m1gp データセットでも，提案手法は CW に比べて高い識別率と再現率を実現していた．これは，提案手法では話題モデルと比較モデルをそれぞれ最新の 200 ツイートから構築することで，リアルタイム性が高く時間変移する話題に素早く追従できたからと考える．なお，m1gp では形態素を素性とした CW の分類性能が最も低かった．これは，m1gp が他のデータセットに比べて新語・口語・俗語が多いために形態素解析の性能が低下したためと考える．データ圧縮による分類は，テキスト中のタームの出現順序を考慮

表 3　提案手法の出力値 $f(x)$（式 (3)）によって昇順に整列した際の上位 15 ツイート（seiji データセット）．なお，ユーザ名と URL は論文記載時に@username，[URL] に変更した．

Hashtag	$f(x)$	Tweet
seiji	0.198	犯罪者は、もはや根本的に非社会的な存在、社会のなかによび入れられた一種の寄生的な要素、すなわち同化しえない異物などではなく、まさしく社会生活の正常な主体としてあらわれる〈デュルケム〉
seiji	0.201	○－６　一方､『超省エネのトランジスターを開発 起動時間ゼロのパソコン実現なるか-物質・材料研究機構 ([URL])』などの技術も加味した『エコスーパーコンピュータ』の開発を期待したい。
seiji	0.217	ほぼ全世帯のサラリーマンで負担増の試算。当然わかってたことなんだけど、昨年夏に民主党に投票した連中は納得なんだよな？ - [URL]
seiji	0.296	うーん、西東京市市議選、やはり民主が候補者多すぎだ。得票数を見たが、候補者を 7 人ではなく、6 人にしていたら、5 人当選できていた。たった 1 人多かったせいで総崩れ。菅グループの選挙の弱さ、読みの甘さが如実に出ている。
seiji	0.296	今の日本は借金漬け。責任取りなさいよ自民党!!(総理大臣経験者と所属議員全員と 55 年体制下で自民党に籍を置いていた人。理由:55 年体制下で国債を増発した。文字から怒りのオーラ放出:紫色)
seiji	0.349	公職選挙法か議員法か分りませんが前にも言ったことが有ります。比例当選者が党から外れるときは議員辞職と同じ党の繰り上げを認めない。悪用すれば１人で２議席を取ることが可能です。議員にとって不利なことも積極的にやりましょう。政治に信頼を取り戻すなら議員自ら示す事です。
seiji	0.359	休日の分散化、反対。「祝日」を軽視していると思う。ハッピーマンデーも止めてほしい。
seiji	0.374	キタ――――(∀)――――!!!@username 【政治】自民、ついに菅首相への問責決議案提出へ…可決される可能性大 [URL] 来たんちゃう!? 来るか？来たか？キタ――――(∀)――――!!
seiji	0.377	日本政府が国連から突っ込まれそうな恥ずかしいネタは数えたらきりが無い。人権・労働関係等々…。
2ch	0.449	ニュー速+:【政治】自民、ついに菅首相への問責決議案提出へ…可決される可能性大 [URL]
seiji	0.472	【抗議ツール有】緊急です！男女共同参画についてご確認下さい！ — FreeJapan [URL] via @username
googlenewsjp	0.481	【共同通信世論調査】内閣支持２３％、不支持６７％予算案７６％評価せず　７０％が小沢氏国会説明を [URL]
seiji	0.503	【米軍再編交付金１６億８０００万円停止　名護市民、諦めと不信】「稲嶺市長も市職員時代は移設案に賛成だった。なぜ反対ばかりするのか理由が分からない」と首をかしげた [URL]
followmejp	0.515	【2 ちゃんねるで話題のスレ】:【政治】自民、ついに菅首相への問責決議案提出へ…可決される可能性大 [URL]
seiji	0.534	こんな時だからこそ、自衛隊に激励の手紙を送ろう！彼等には国民の良心が共にあることを伝えよう！今、一番国民が期待しているという事実と、彼等に祝福と感謝を伝えよう！

[4] http://ruby-doc.org/core/classes/Zlib/Deflate.html
[5] http://mecab.sourceforge.net/

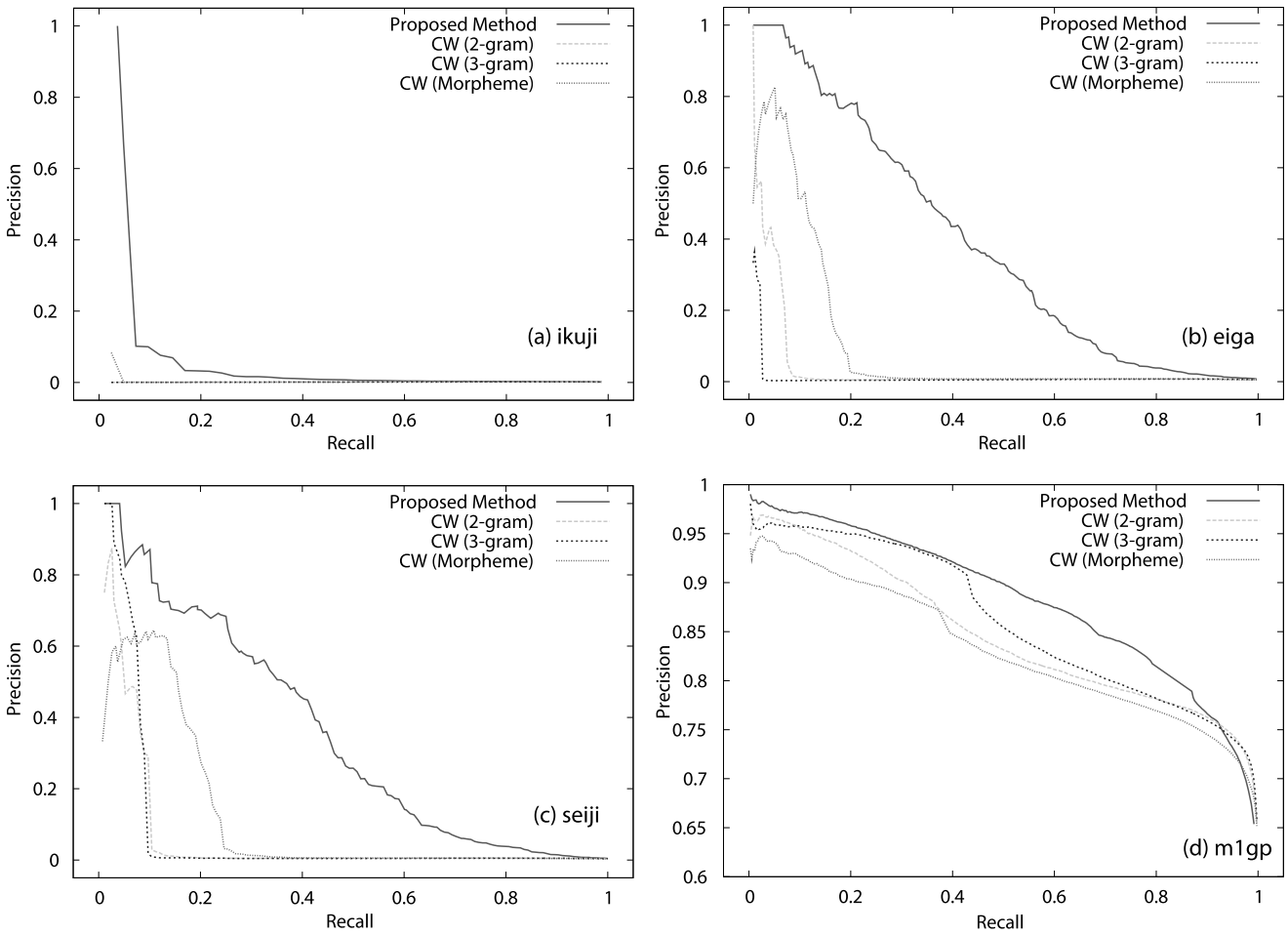

図 6　（a）ikuji（育児）（b）eiga（映画）（c）seiji（政治）（d）m1gp（テレビ番組）の各データセットにおける，提案手法と confidence-weighted linear classification（CW; 素性: 文字 2-gram, 文字 3-gram, 形態素）の識別率と再現率.

しない bag-of-words を素性とした機械学習による分類に比べて，テキストの文脈（タームの前後関係）が考慮されやすい．このことは，テキストが短く，分類に用いることのできる情報量の少ないツイートの分類においては大きな利点になると考える．

5. 関連研究

近年，ツイートの分類に関連する研究が進んでいる．Irani らは，ツイートがスパムであるか否かについて，ツイートのテキストと，ツイートからリンクされた Web ページの内容を用いて，機械学習により分類を行った [10]．Sriram らは，ツイートのタイプ（ニュース，イベント，意見，Deals，プライベートメッセージ）の分類を行うため，bag-of-words に加えて，著者情報とテキスト情報（@username がツイートの初めにある，通貨記号がある，など）を素性として機械学習を実施した [11]．Sankaranarayanan らは，Twitter に基づくニュースの配信システムを構築するため，bag-of-words を素性とした機械学習によりニュースとノイズのフィルタリングを行っている [12]．また，Go らによるツイートの感情分類 [13] や，Sakaki らによるリアルタイムイベント検出 [14] でも機械学習が用いられているが，データ圧縮をツイート分類に応用した例は無い．

6. おわりに

我々は，マイクロブログサービスである Twitter に日々投稿される膨大なツイートの中から，着目する話題に関するツイートを分類するため，ツイートの圧縮されやすさを応用した手法を提案した．提案手法では，新しいツイートを，着目する話題に関するツイートの集合（話題モデル）と，それ以外のツイートの集合（比較モデル）の両方を基にして圧縮する．このとき，話題モデルを基にした方が圧縮されやすい場合に，新しいツイートは着目する話題に関連する可能性が高いとみなす．提案手法はデータ圧縮を利用することで，形態素解析に依存せず，新語や口語・俗語が多く含まれるツイートを精度良く分類できる．

評価実験では，ハッシュタグ付きの日本語ツイートを用いて，新しいツイートが着目するハッシュタグに関するものか否かを分類する実験を実施し，現在提案されているオンラ

イン分類器の中で最も優れたものの一つである confidence-weighted linear classification と比較して，提案手法が優れた識別率と再現率を実現したことを示した．なお，提案手法は，ハッシュタグ分類に限らず，汎用的な話題分類に使用可能である．また，提案手法は日本語ツイートのみならず，どのような言語で記述されたツイートにも適用可能である．

本論文ではさらに，800 万件を超えるツイートを利用してハッシュタグに関する解析を行い，各ハッシュタグのツイート数に関する分布，ツイート数の時間変化，ツイート内容の時間変化について明らかにした．リアルタイム性の高いハッシュタグではツイート内容が時間と共に激しく変化することと，各ハッシュタグのツイート数に大きな偏りが存在することが，学習を難しくする大きな要因であることが分かった．提案手法はこの知見を利用して，それぞれ最新の N ツイートから構築した話題モデルと比較モデルを用いて新しい入力ツイートの圧縮されやすさを比較することで，高い分類性能を実現した．

なお，従来のツイート分類に関する研究の大部分は機械学習を用いており，データ圧縮をツイート分類に初めて応用した本研究は，口語的で短くリアルタイム性の高いテキストの分類に関する研究において大きな貢献を果たしたと考える．

今後は，将来的な実運用を見据えて，手法の分類精度と分類に要する時間に関してブラッシュアップを行いたい．また，パラメータ N や γ が分類精度に与える影響や，他の圧縮法を用いた場合の分類精度についてさらに詳細に評価したい．

[文献]

[1] H. Kwak, C. Lee, H. Park, and S. Moon, “What is twitter, a social network or a news media?,” Proceedings of 19th International Conference on World Wide Web, pp.591–600, 2010.

[2] D. Benedetto, E. Caglioti, and V. Loreto, “Language trees and zipping,” Physical Review Letters, vol.88, no.4, 28 Jan. 2002.

[3] M. Li, X. Chen, X. Li, B. Ma, and P.M.B. Vitányi, “The similarity metric,” IEEE Transactions on Information Theory, vol.50, no.12, pp.3250–3264, 2004.

[4] A. Bratko, G.V. Cormack, B. Filipič, T.R. Lynam, and B. Zupan, “Spam filtering using statistical data compression models,” Journal of Machine Learning Research, vol.7, pp.2673–2698, 2006.

[5] M. Li, and P. Vitányi, An Introduction to Kolmogorov Complexity and Its Applications, Springer, 2nd edition, 1997.

[6] L.P. Deutsch, “RFC1951: DEFLATE compressed data format specification version 1.3,” http://tools.ietf.org/html/rfc1951, 1996.

[7] J.G. Cleary, and I.H. Witten, “Data compression using adaptive coding and partial string matching,” IEEE Transactions on Communications, vol.COM-32, no.4, pp.396–402, 1984.

[8] G. Cormack, and R.N.S. Horspool, “Data compression using dynamic markov modelling,” The Computer Journal, vol.30, no.6, pp.541–550, 1987.

[9] M. Dredze, K. Crammer, and F. Pereira, “Confidence-weighted linear classification,” Proceedings of 25th International Conference on Machine Learning, pp.264–271, 2008.

[10] D. Irani, S. Webb, C. Pu, and K. Li, “Study of trend-stuffing on twitter through text classification,” Proceedings of 7th Annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference, 2010.

[11] B. Sriram, D. Fuhry, and M. Demirbas, “Short text classification in twitter to improve information filtering,” Proceedings of 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp.841–842, 2010.

[12] J. Sankaranarayanan, H. Samet, B.E. Teitler, M.D. Lieberman, and J. Sperling, “Twitter stand: News in tweets,” Proceedings of 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, pp.42–51, 2010.

[13] A. Go, R. Bhayani, and L. Huang, “Twitter sentiment classification using distant supervision,” Technical Report CS224N Project Report, Stanford University, 2009.

[14] T. Sakaki, M. Okazaki, and Y. Matsuo, “Earthquake shakes twitter users: Real-time event detection by social sensors,” Proceedings of 19th International Conference on World Wide Web, pp.851–860, 2010.

西田 京介 Kyosuke NISHIDA

日本電信電話株式会社 NTT サイバーソリューション研究所研究員．2008 年 北海道大学大学院情報科学研究科博士課程修了．2007–2009 年日本学術振興会特別研究員．2009 年日本電信電話株式会社入社．機械学習，Web マイニングの研究開発に従事．博士 (情報科学)．電子情報通信学会正会員．

坂野 遼平 Ryohei BANNO

北海道大学大学院情報科学研究科複合情報学専攻博士前期課程在学中．2010 年北海道大学工学部情報エレクトロニクス学科卒業．主に広域分散処理技術の研究に従事．情報処理学会学生会員．

藤村 考 Ko FUJIMURA

日本電信電話株式会社 NTT サイバーソリューション研究所主幹研究員．1989 年北海道大学大学院工学研究科博士課程修了．同年日本電信電話株式会社入社．トランザクション処理記述言語，電子価値流通システム，ソーシャルメディアからの知識抽出と可視化の研究開発に従事．工学博士．情報処理学会，電子情報通信学会，各会員．

星出 高秀 Takahide HOSHIDE

日本電信電話株式会社 NTT サイバーソリューション研究所主任研究員．1993 年九州大学大学院総合理工学研究科修士課程修了．同年日本電信電話株式会社入社．遠隔教育システム，Web マイニングの研究開発に従事．情報処理学会正会員．