経験を述べたツイートの抽出と行動カテゴリ分類

六沼　元貴†　　安江　駿亮†　　杉本　徹‡
芝浦工業大学大学院理工学研究科†　　芝浦工業大学工学部‡
{ma17130, ma17121, sugimoto}@shibaura-it.ac.jp

1. 研究背景と目的

近年，Facebook や Twitter など，ユーザが自由に短文をインターネット上に投稿できるサービスが多く利用されている．投稿されたこれらの情報は，企業のマーケティングやトレンド分析を目的として収集される機会も多い．特に Twitter 上で呟かれるツイートには，ユーザが何をした，どこへ行ったなど，行動や経験に関する情報を含むものが多く見られる．これらの情報を解析することは，コンピュータがユーザの行動パターンや嗜好情報を理解し，正しく記録するために重要な処理である．またこれらの情報は，例えば飲食店の推薦システムにおいて，ユーザ一人ひとりの好みを考慮した上で適切な情報を提供したり，対話システムにおいて，ユーザが親近感を抱きやすい応答を行う技術などに応用することができる．

本研究でははじめに，Twitter に投稿されるツイートを，ユーザの経験情報を含むツイートとそれ以外のツイートに分類する．ここで経験情報とは，ユーザが何をした，どこへ行ったなど，ユーザ自身が行った行動や経験に関する情報である．その後，経験情報を含むツイートから行動の時間や場所といったライフログデータを抽出する．さらに，経験情報を含むツイートを行動カテゴリへと分類する．それぞれの処理には，機械学習の手法である SVM(Support Vector Machine)とロジスティック回帰分析を用いる．

2. 経験情報を含むツイートの分類手法

経験情報を含むツイートの分類は以下の手順によって行う．

① ツイートの収集

分類器を作成するにあたり，必要な訓練用ツイートと評価用ツイートを Twitter Streaming API を用いて収集する． 収集するツイートは日本語に限定し，リプライやリツイート，URL の記述を除外する．

② 行動ツイートの分類

①で収集したツイートを，経験情報を含むツイートとそれ以外に人手で分類する．

③ SVM で学習し，経験情報を含むツイートの分類実験を行う．

④ 実験結果を分析し，手法の改良を行う．

⑤ ③，④を何度か繰り返した後，実験により分類の精度を評価する．

3. ライフログデータの抽出

3.1 ライフログデータの定義

本研究では，経験情報を含むツイートには表 1 のようなライフログデータが含まれていると考える．ここでライフログとは，個人の行動の記録を表す．

表 1. ライフログデータの定義

名称	定義	例
行動情報	ユーザの行動 ツイート中の動詞	行く,遊ぶ, 仕事する
時間情報	行動した時間	2018-01-01
場所情報	行動した場所 場所を表す名詞	東京,家, アメリカ
感想情報	行動に関する感想 形容詞	楽しい,悲しい
対象情報	行動の対象 動詞の目的語である名詞	サッカー, プレゼント
理由情報	行動の理由 理由となり得る文節	雨なので, 暑いから
手段情報	行動に用いた手段 手段となり得る名詞	ハサミ,電車

3.2 ライフログデータの抽出手法

ライフログデータの抽出は以下の手順で行う．

① ツイートの解析

2 節で述べた手法で収集した経験情報を含むツイートに対し形態素解析，係り受け解析，格解析を行う．これらの解析には形態素解析器 JUMAN および構文解析器 KNP を用いる．

② ラベル付け

各形態素が時間情報を除く 6 情報のどの情報を持つか，または何も情報を持たないかを人手でラベル付けする．

③ SVM で学習し，ライフログデータの抽出実験を行う．

④ 実験結果を分析し，手法の改良を行う．

⑤ ③④を何度か繰り返した後，実験により抽出の精度を評価する．

4. 行動カテゴリ分類

4.1 行動カテゴリの定義

本研究では，行動カテゴリという概念を定義し，収集したツイートをそれらに分類する．行動カテゴリとは，ユーザが経験した行動の種類を表す分類である．総務省統計局が『社会生活基本調査』という国民アンケートで使用している分類[1]を雛形としており，人手による分類実験を行うことで，不要なカテゴリの削除，および必要なカテゴリの追加を行った．結果として，本研究では 21 個の行動カテゴリを定義した．行動カテゴリの一覧を，表 2 に示す．このうち，Twitter における出現回数が多かった上位 3 カテゴリである『食事』，『サブカルチャー』，『買い物』について，本研究の手法を実践する．

表 2. 行動カテゴリの一覧

睡眠	食事	通勤・通学
移動	仕事	学業
学習・訓練	家事	買い物
テレビ・ラジオ	読書	サブカルチャー
趣味・娯楽	休養	旅行・外出
スポーツ	社会参加活動	医療
育児	交際・付き合い	その他

4.2 行動カテゴリ分類の手法

本研究では，機械学習を用いてツイートを自動的に適切な行動カテゴリへと分類することを目標としている．具体的な分類の手法としてロジスティック回帰分析を用いる．ロジスティック回帰分析は，説明変数の値から目的変数の発生確率を予測する統計学的な手法である．本研究では，1 つのツイートに含まれる単語を説明変数とし，各行動カテゴリを目的変数として設定する．

4.3 教師データの作成

2 節で述べた手法で収集した経験情報を含むツイートは全部で 127,032 件となった．これらのツイート群を本手法でのコーパスとして扱う．このコーパスから，『食事』，『サブカルチャー』，『買い物』カテゴリへと分類されるツイートを抽出し，ロジスティック回帰分析の教師データとする．この教師データの作成手法として，ブートストラッピング[2]を用いた．本研究での教師データの作成手法を図 1 に示す．

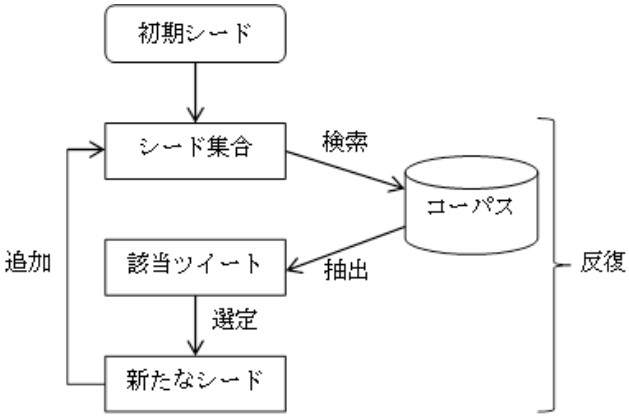

図 1. 教師データの作成手法

まず初めに，初期シードとしていくつかの単語を人手で与える．今までに人手で収集した，『食事』，『サブカルチャー』，『買い物』に分類されるツイートを MeCab[3]を用いて形態素解析し，多く出現している単語の中から，特にカテゴリの特徴を表すと考えられる単語を 6 つずつ選び出し，初期シードとする．次に，これ

らのシード（単語）をキーワードとしてツイートコーパスから検索し，複数あるシードのうち 1 つ以上が含まれているツイートを抽出する．

抽出したツイートを形態素解析し，出現数上位の単語から新たなシードを選び出す．ここで，単語の一覧から『する』，『てる』など多くのツイートに共通して含まれる語はストップワードとして除外する．残った単語の中で，その単語を見たときにカテゴリを連想しやすいものを，そのカテゴリの特徴を表す単語と判断し，人手で選び出した．

そして，1 回の試行で新たに得られたシードをシード集合に追加する．この手続きを 5 回反復し，シード集合を順次拡張していく．最後に，得られたすべてのシードのうち 2 つ以上を含んでいるツイートのみをコーパスから抽出し，それらを正例の教師データとして扱う．例として，『食事』カテゴリの初期シードを表 3 に，ブートストラッピングを 5 回反復したとき，各回に得られた『食事』カテゴリのシード集合を表 4 に示す．

表 3. 『食事』カテゴリの初期シード

食べる	飲む	ご飯
マック	ラーメン	味

表 4. 『食事』カテゴリの獲得シード

反復数	得られたシード
1 回目	食う,お腹,お昼,味噌
2 回目	空く,飯,夜,朝,腹,カレー,酒,ごはん,アイス,丼,肉,賞味,ミント,パン,晩,チョコ,料理,たべる,ケーキ,美味,麺,味噌汁,寿司,ビール,焼肉,焼く,炊く
3 回目	肉,夕飯,御飯,昼飯,居酒屋,おなか,野菜,牛,腐る,胃,メニュー,コーヒー,弁当
4 回目	牛乳,セブン,コンビニ,飲める,スタバ,米,魚,うどん,サラダ,パンケーキ,イカ
5 回目	スイカ,食,ランチ,ショートケーキ,食い,塩,小腹,そば,喉,ジュース

5. 評価実験と考察

5.1 経験情報を含むツイートの分類精度の評価

評価用ツイートとして新たに 500 件のツイートを用意し，訓練した分類器で分類した．さらに，分類結果を人手で判断し分類精度を評価した．分類器は RBF カーネル，パラメータ γ=0.001，C=1000 の SVM を使用した．素性はツイートに含まれる品詞が普通名詞，サ変名詞，固有名詞，動詞，形容詞，未定義語である単語を bag of words でベクトル化したものと，表 5 に示す表現の有無を用いた．

表 5. ツイート分類で用いる素性

ツイートに含まれる表現	例
過去	「した」,「だった」
否定	「ない」,「なかった」
希望	「たい」
疑問	「？」
受身	「られる」

実験の結果，正解率は 0.936 となった．実験結果の混同行列を表 6 に示す．

表 6. ツイート分類の混同行列

		分類結果		
		経験情報を含む	経験情報を含まない	合計
正解データ	経験情報を含む	71	0	71
	経験情報を含まない	32	397	429
	合計	103	397	500

5.2 ライフログデータの抽出精度の評価

訓練した分類器で評価用ツイート 300 件から時間情報を除くライフログデータを抽出した．分類の対象となる単語は，形態素解析の結果品詞が名詞，動詞，形容詞，未定義語であった単語である．分類器は多項式カーネル，パラメータ γ=0.1，d=2，C=10 の SVM を使用した．素性は，分類対象の形態素とその前後 1 形態素の原形とそれらの品詞，係り先の主辞の品詞，動詞の直後に「たら」，「ので」，「から」などの理由を示す可能性のある単語の有無を用いた．上記分類器を使用後，手段情報抽出用の分類器を使用した．実験結果を表 7 に示す．

5.3 行動カテゴリ分類の実験

ブートストラッピングによって得られたツイ

表 7. ライフログデータ抽出の実験結果

	適合率	再現率	F 値
無	0.70	0.86	0.77
行動	0.92	0.94	0.93
場所	0.76	0.76	0.76
感想	0.97	0.87	0.91
対象	0.88	0.80	0.84
理由	0.96	0.63	0.76
手段	0.95	0.63	0.76

ートを教師データとした，行動カテゴリ分類の精度を確認する実験を行った．本実験では，教師データに含まれる名詞および動詞を対象とした bag of words ベクトルをロジスティック回帰分析の特徴量として用いる．ここで，次元の増大を防ぐため，コーパス中での出現数が 1 である単語は除外している．総単語数は 23,739 個となった．教師データに用いる各ツイートに対して，(分類される:1，分類されない:0)の正解カテゴリと，bag of words ベクトルのペアが与えられる．

ロジスティック回帰分析の学習と予測分類には LIBLINEAR[4]を用いる．テストデータとして正例 30 個，負例 30 個のツイートを各カテゴリにそれぞれ用意する．

ロジスティック回帰分析に使用した正例の教師データの数，および予測分類で求められた正解率の値を以下に示す．表 2 は教師データをすべて人手で作成した場合の実験結果であり，表 3 はブートストラッピングを用いて半自動的に教師データを作成した場合の実験結果である．

また，この場合の各カテゴリの再現率，適合率，F 値を表 10 に示す．人手での作成はコストがかかることから教師データの数は少なくならざるを得ない．これに対して，提案手法を用いた場合，大量の教師データを使って学習し予測できることから，従来の手法に比べて高い正解率が得られた．さらに『食事』，『サブカルチャー』，『買い物』のいずれのカテゴリについても適合率が 0.85 を上回る結果となっており，行動カテゴリ分類において本手法が一定の精度を確保していることが示された．

表 8. 従来手法を用いた分類実験の結果

	正例データ数	正解率
食事	112	0.550
サブカルチャー	118	0.383
買い物	115	0.433

表 9. 提案手法を用いた分類実験の結果

	正例データ数	正解率
食事	1414	0.817
サブカルチャー	1318	0.717
買い物	738	0.783

表 10. 提案手法での再現率，適合率，F 値

	再現率	適合率	F 値
食事	0.767	0.852	0.807
サブカルチャー	0.467	0.933	0.622
買い物	0.567	1.000	0.724

6. まとめと今後の課題

本研究では Twitter に投稿されるツイートが経験情報を含むか否かを機械学習で分類し，経験情報を含むツイートからライフログデータを機械学習で抽出する手法を提案した．さらに，経験情報を含むツイートを行動カテゴリへと分類する方法を提案した．ユーザによって投稿されたツイートを分析し，適切な行動カテゴリへ分類することは，コンピュータがユーザの行動パターンや嗜好情報を理解し，正しく記録することにつながる．評価実験の結果，分類精度を確認し，また提案手法の優位性を示すことができた．今後は，分類精度の向上を目指すとともに，ユーザの情報を正しく理解する対話システムへの実装を実現したい．

参考文献

[1] 統計局ホームページ 平成 23 年社会生活基本調査 用語の解説 http://www.stat.go.jp/data/shakai/2011/pdf

[2] 加藤誠, 大島裕明, 小山聡, 田中克己,『ブートストラップ法による語の共起を用いた Web からの類似関係抽出』, DEIM Forum 2009 A5-6

[3] MeCab http://taku910.github.io/mecab/

[4] LIBLINEAR https://www.csie.ntu.edu.tw/~cjlin/liblinear/