カテゴリー分類を用いたツイートデータの特徴語抽出の評価

2013SE038 廣瀬史明　2013SE177 佐野文也　2013SE215 田中努
指導教員：河野浩之

1　はじめに

Twitter は毎年アクティブユーザが増加し，2016 年 6 月では 3 億 1300 万人にものぼる [1]．ユーザの嗜好情報が Twitter に投稿されていても必ずしもユーザに提示されないという問題が存在する．その大きな理由としては，世界で 1 分間で約 34 万ツイートされ続けており，その溜まっていく莫大な情報量がユーザの的確かつ有益な情報の取得を困難にしている [2]．次に言語の問題である．「サッカー」と「soccer」のように表記の違いから嗜好情報を見つけることができないという理由である．

本研究では Twitter API を用いてユーザごとのツイートを収集し，どの嗜好情報とも関連度が高い語句の除去や現代語の変化に対応できるように考慮される語句の数を拡大し，多くの特徴語の抽出を目的とする．また，表記の違いで一緒に考慮されないという理由より，同じ情報として統一させることで正確な特徴語の抽出を目指す．

本研究における論文は全 6 章で構成されており，各章の構成は以下のようになっている．まず，第 2 章では Twitter 情報推薦に関する先行研究 [3][4] で実験内容，結果，及び課題を比較し，本研究の目的を提示する．次に，第 3 章では第 2 章で取り上げた先行研究の課題に対しての解決方法とその方法で用いるアルゴリズムを提案し，さらに本実験で使用する技術の紹介をする．第 4 章で先行研究の課題解決を踏まえた本研究の実験の流れを説明する．そして，第 5 章で本研究の実験結果を報告し，最後に第 6 章で結論と今後の課題を述べる．

2　Twitter 情報抽出に関する先行研究

本章では，それぞれの先行研究 [3][4] の手法，結果，課題を簡潔に明示し，先行研究 [3][4] について比較する．

2.1　Twitter における語の関連性に着目したユーザ興味語抽出手法の提案 [3]

渡邉らの研究では，語句の共起関係と逆文書出現頻度を用いて作成した関連語辞書を用いることで語句の関連性を考慮したユーザの嗜好分析を行うことで，ユーザの嗜好情報を表す興味語が抽出されることを可能とした．語句の共起関係と逆文書出現頻度を利用した関連語辞書を用いることで新しい言葉の発生や語句の関連性の高速な変化に対応することが可能となり，語句の表記のぶれや分析の対象情報が少ないという問題に対してもユーザの興味語を抽出することができると渡邉らは考えた．

結果，提案手法の方が各被験者の中で正答数が多く，沢山の興味語が抽出できた．

2.2　嗜好に基づく時事情報推薦システムの構築 [4]

山本らの研究では，情報収集の効率を上げる手段としてテレビの視聴履歴および Twitter の被験者のツイートを用いて関連性のある情報を収集，分析することで自立語の属性や重みを獲得し重要度を付与することでユーザの嗜好情報の選別の精度向上を図った．

実験の結果，無作為の時事情報を選出する方法より平均で 2.3 倍の個人の趣味・嗜好に沿った時事情報の選別，提供が可能になったと報告している．

2.3　Twitter 情報抽出の課題

表 1 に先行研究の比較を示す．我々は，渡邊らの研究の各手法に精度の差がついた原因としてツイートを分析する中で正式名称と略語などが同一の関連性を持っていても別の情報として認識され，一緒に考慮されなかったことが原因ではないかと推測した．これらを同じ情報として認識させ考慮する必要がある．また，150 位より下位の語句は興味がある語句であっても関連性の考慮がされていないので，関連語辞書の範囲を拡大することが挙げられる．

山本らの課題としては，嗜好情報とも関連度が高くなる語句の除去が求められる．

表 1　Twitter を用いた特徴語抽出の先行研究比較

先行研究	手法	実験結果	今後の課題
渡邊ら 2012[3]	投稿の分析，関連語辞書の構築，ユーザ嗜好分析	興味語抽出性能の向上，ユーザへの興味語を抽出	対象とする語句の範囲の拡大，正式名称と略語，愛称を同じ情報として考慮
山本ら 2013[4]	テレビの視聴履歴，Twitter, web 記事を用いた重要度付加	平均 2.3 倍の趣味嗜好にそった情報の選別，提供が可能	どの嗜好情報とも関連度が高くなってしまう語句の除去

3　特徴語抽出の提案

本章では，前章で上げた問題点の改善方法の提案と本実験に用いる技術やツールの説明をする．

3.1　問題点と改善方法

我々は，対象とする語句の範囲の拡大は名詞にカテゴリーを付与し，カテゴリーに分けて考慮するという形で改善を図る．これによりカテゴリー考慮できる語句の範囲を広げることができる．また，正式名称と略語などが一緒の

情報として考慮されない問題で，TF-IDF を実行する前に略称などを正式名称に置換することで改善を図る．どの嗜好情報とも関連度が高くなってしまう問題でカテゴリーと属性を名詞に付与することで改善を図る．取り扱うデータを特定のカテゴリーに限定することで関連性のない語句の除去が可能になると考えた．

3.2 特徴語抽出システム構成図

図 1 に本実験の特徴語抽出システムの構成図を示す．

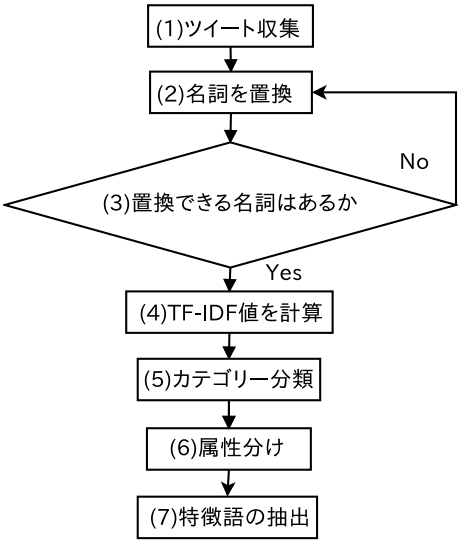

図 1 本実験における特徴語抽出システム構成図

次の (1) から (7) は図 1 の (1) から (7) と対応している．
(1) 1 人につきツイートを最新から最大 200 件収集する．
(2) 特定のカテゴリーに含まれる名詞を対象に正式名称に置換するプログラムを実行する．
(3) 置換可能な名詞が全て置換完了するまで繰り返す．
(4) 1 人ずつ各カテゴリーに分けて TF-IDF 値を算出する．
(5) カテゴリーごとに名詞を分類をする．
(6) (5) で各カテゴリーに分類したものをさらに細かく分類するために属性分けをしていく．
(7) 見られた属性がそのユーザのツイート全体を特徴付けるものであるため，特徴語として抽出できる．

本研究での属性分けとは，カテゴリーに属している名詞をさらに細かく分類することを指す．カテゴリーに分類した名詞をさらに細かく分類することにより，より細かい特徴語の抽出が可能となる．

3.3 Twitter API を用いたツイート収集

Twitter API はもともとあるプログラムを呼び出すことができるもので簡易に扱えるメリットがある．Twitter API の機能の中でタイムライン関連，ユーザ関連，DM 関連，フレンド関連などの機能が存在している．これらの機能の中で，タイムライン関連の API を採用する．この機能では，指定したユーザのツイートを最新から最大 200 件まで取得することが可能である．

3.4 形態素解析ツール

Mecab は未知の語句に対して定義の変更を行うことが可能なため，ノイズの除去に活かすことが可能であり，ChaSen よりも平均 3 倍から 4 倍ほどの解析速度で解析できる[*1]．Mecab による形態素解析の出力結果に示される品詞細分類では，解析した単語の色々な属性が出力される．そこで Mecab の辞書に特定のカテゴリーに属している名詞に登録することができれば特定のカテゴリーに絞り込んで名詞の抽出ができると考えた．我々の提案に対して Mecab は最も有効であるので Mecab を採用する．

3.5 特徴語抽出

情報検索や文章要約などの分野で活躍している TF-IDF 法と属性分けを用いて特徴語の抽出を行う[*2]．
TF-IDF 法は特定のカテゴリーに含まれる名詞の頻出度を表す TF と被験者の人数分の取得した総ツイート数と特定のカテゴリーに含まれる名詞が見られるツイート数から IDF を算出してその 2 つの指標に基づいて計算されるため文章を m，名詞を n と置くと (1) 式のように表せる．

$$tfidf_{m,n} = tf_{n,m} \times idf_n \tag{1}$$

(1) 式を使って計算された値が高ければ高いほど TF-IDF 値が高く，それぞれのカテゴリーに分けて単語をさらに属性分けすることによって，どの属性がそのユーザにとって興味のある属性になるかが判断可能になる．TF-IDF 法から特徴語抽出までの流れについて例を用いて表したものを図 2 に示す．

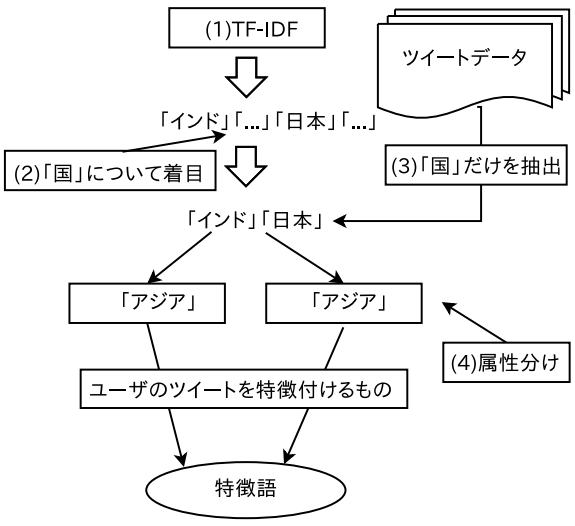

図 2 TF-IDF から特徴語抽出までの流れ

[*1] https://ja.wikipedia.org/wiki/MeCab
[*2] https://ja.wikipedia.org/wiki/Tf-idf

次の (1) から (4) は図 2 の (1) から (4) と対応している．今回は「国」というカテゴリーの特徴語の抽出を例として示す．
(1)TF と IDF に基づいて TF-IDF 値を算出する．
(2)「国」というカテゴリーに含まれる名詞に着目する．
(3)「国」というカテゴリーの名詞に絞って抽出する．
(4)(3) で抽出できた名詞に属性分けをしていく．そのユーザの特徴語は「北アメリカ」と「ヨーロッパ」になる．

他のカテゴリーの名詞がある場合，(2) から (4) までの流れを着目してないカテゴリーがなくなるまで繰り返す．

4　特徴語抽出の性能評価実験

本章では，実験環境と実験の流れを示し，我々の提案について詳しく説明する．

4.1　特徴語抽出システム実験の流れ

CPU:Intel Corei5-3320M CPU @ 2.60GHz，メモリ:4GB，OS:Ubuntu14.04 のパソコンを使用した．提案手法と比較手法を比較し提案に対する評価をした．また，[5] を参考に「芸能人・有名人」,「社長・実業家」,「政治家・議員」,「クリエイター」,「スポーツ系 (選手，団体含む)」の計 5 分野のそれぞれのフォロワ数上位 10 位までの合計 50 名を本実験の被験者とした．5 つの各分野の 1 位から 10 位までの 5 グループに分けて行った．使用言語は Python を使用し，ツイート収集には Twitter API 専用のライブラリ tweepy を用いた．

ここで提案手法の手順を (1) から (6) に示す．
(1) 端末からプログラムを実行する．
(2)10 名分のツイートを txt ファイルに保存する．
(3) 特定のカテゴリーの名詞を正式名称に置換する．
(4) それぞれのカテゴリーごとに TF-IDF を行う．
(5) それぞれのカテゴリーごとに属性分けを繰り返す．
(6) 現れた属性全てがそのカテゴリーの特徴語となる．

そして (1)，(2)，(4)，(5)，(6) を比較手法とする．実験は提案手法と比較手法を 5 グループ分繰り返す． 例として収集したツイートの一部を形態素解析した結果を図 3 に示す．最新から最大 200 件のツイートを取得することができる．

```
インド共和国 名詞, 国有名詞, 地域, 国, アジア,*, インド共和国
日本国 名詞, 国有名詞, 地域, 国, アジア,*, 日本国
```

図 3　形態素解析した一部

4.2　カテゴリーと属性の追加

wikipedia の情報量が多く，多くの人の手によって編集されているので日本の中でも注目度が高いと判断したため，実験で扱うカテゴリーとして「国」[*3],「ペット」[*4],「スポーツ」[*5]にした．属性はカテゴリーをさらに細かく分類したものである．例えば，「日本」は「国」のカテゴリーに属するがさらに細かく分類すると「アジア」に分類できる．このカテゴリーと属性の追加は Mecab の辞書を利用して実現できるものであると考えた．wikipedia を参考にカテゴリーと属性を追加登録した．表 2 にカテゴリーと属性を追加した辞書の一部を載せたものを示す．表 2 のスポーツはカテゴリー，球技は属性を表している．

表 2　Mecab の辞書にカテゴリーと属性を追加した表

サッカー	*	*	*	名詞	一般	スポーツ	球技
フットサル	*	*	*	名詞	一般	スポーツ	球技
ビーチサッカー	*	*	*	名詞	一般	スポーツ	球技
ラグビー	*	*	*	名詞	一般	スポーツ	球技

4.3　置換プログラム

正式名称と略語，愛称が同じ情報として考慮されない問題は正式名称に統一することで同じ情報として考慮することで解決できると考えた．具体的な方法としては，Wikipedia を参考にして作った略語などを正式名称に置換するプログラムを方法を採用した．今回追加したパターンは 3132 個である．置換プログラムの一部を以下に示す．
401.data = data.replace(”アメリカ”,”アメリカ合衆国”)
402.data = data.replace(”USA”,”アメリカ合衆国”)
403.data = data.replace(”米国”,”アメリカ合衆国”)

例としてツイートデータに置換プログラムを用いた結果の一部を図 4 に示す．これにより略称や愛称，英語，正式名称と別々に考慮されるはずのものが一緒の情報として考慮できるようになり，TF-IDF 法による重み付けの精度の向上が図れる．

```
置換前：　インド　日本
置換後：　インド共和国　日本国
```

図 4　置換した名詞の一部

4.4　TF-IDF 法を用いた属性分け

今回 3 つのカテゴリーに含まれる名詞を対象として重み付けを行い，TF-IDF 値が高い順に名詞を並び替えたものを txt ファイルに出力した．TF-IDF の表に現れたカテゴリーがそのユーザのツイートを特徴づけるものが含まれている可能性があるので，そのユーザの取得したツイートデータを対象にカテゴリーごとに分けて名詞を全て抽出す

[*3] https://ja.wikipedia.org/wiki/国の一覧
[*4] https://ja.wikipedia.org/wiki/ペット
[*5] https://ja.wikipedia.org/wiki/スポーツ競技一覧

る．特定のカテゴリーに含まれる単語のみを抽出するプログラムの一部を以下に示す．これはスポーツというカテゴリーに含まれる名詞のみを抽出することを表している．

```
65.while node:
66. if node.feature.split(",")[2] == u"スポーツ":
67. keywords.append(node.surface)
68. node = node.next
69.return keywords
```

上記のプログラムで抽出できた名詞を Mecab の辞書を用いてカテゴリー分類と属性分けをしていく．属性分けをしたときに見られた属性全てがそのユーザのツイート全体を特徴付けるもの，つまり特徴語となる．

5 特徴語の抽出の実験結果

本実験で提案手法と比較手法の 2 つの手法を 1 グループ 10 名でそれぞれ 50 名分行った．実験結果として比較手法で抽出できた名詞数と提案手法で抽出できた名詞数の差分を各グループごとに算出した結果を表 3 に示す．

表 3 比較手法と提案手法の名詞抽出数の差

	芸能人 有名人	クリエイター	社長実業家	スポーツ	政治家 議員	合計	合計の平均
比較手法	240	139	524	614	836	2353	470.6
提案手法	401	216	708	884	1175	3384	676.8
差分	161	77	184	270	339	1031	206.2
差分の平均	16.1	7.7	18.4	27	33.9	103.1	20.62

この表は正の数ほど提案手法の抽出数が比較手法より多く，負の値ほど比較手法の抽出数が提案手法より多いことを表している．比較手法と提案手法は，各グループの TF-IDF の計算対象となった名詞の抽出した数を表している．差分は，提案手法の名詞抽出数と比較手法の名詞抽出数の差を表している．差分の平均は，提案手法の名詞抽出数から比較手法の名詞抽出数を引いた一人当たりの差を表している．置換プログラムを用いた提案手法の方が置換プログラムを用いなかった比較手法より多くの名詞を抽出している．提案手法の方が比較手法より抽出した名詞の数が 1031 個多く TF-IDF 値が比較手法より細かく分散されていたのでより多くの名詞を考慮することができた．2 つの手法の各カテゴリーの名詞抽出数を表 4 に示す．

特徴語抽出をした結果，全体的に国に関係する特徴語が多かった．提案手法で国が 1000 以上あるのに対し，ペット，スポーツは少なく国に関心があるのが読み取れる．ペットは特徴語の抽出ができなかった箇所が少し存在する．スポーツも同様であった．比較手法より提案手法が多くの特徴語を抽出している．政治家・議員が「国」に関する名詞が特に多く，これは外交に関することや日本での自然災害が多かったことが考えられる．これにより，提案手法は語句の表記の違いから一緒に考慮されなかった同じ情報を持つ語句に対して有効であることが分かった．

表 4 カテゴリーごとの名詞抽出数

	提案手法			比較手法		
	国	ペット	スポーツ	国	ペット	スポーツ
芸能人・有名人	308	67	26	208	12	20
社長・実業家	146	57	13	104	22	13
クリエイター	593	63	52	456	20	48
スポーツ	667	72	145	483	43	88
政治家・議員	1058	91	26	792	26	18

6 むすび

今回の実験では，名詞に「国」,「ペット」,「スポーツ」のカテゴリーとカテゴリーをさらに細かく分類した属性を Mecab の辞書に追加した．それを用いて特定のカテゴリーに絞った特徴語抽出を行った．また，略語など同じ意味を持つ名詞を正式名称に統一させて同じ情報として認識させることによって，Mecab の辞書に含まれていなかった名詞数や特徴語の抽出数を増加させることを可能とした．これらにより推薦システムにおいてより正確な嗜好情報をユーザに提供できるようになると考えられる．

今後の課題として，特徴語の抽出の結果が得られなかった箇所が存在するカテゴリーがあったので，カテゴリーの数を増やすことが挙げられる．また，カテゴリーをより細かく分類した属性のバリエーションを増やすこと，1 名ごとのツイートの取得数を増やすことが挙げられる．また，1 回の実験の被験者数を増やすことでより TF-IDF の精度が上がるだろうと考えられる．今後は，Twitter のユーザ数は圧倒的に多いため，より正確なデータを得るために実験の被験者数を拡大していきたい．

参考文献

[1] Twitter,INC，“Twitter Q2 2016 Shareholder Letter,”https://www.sec.gov/Archives/edgar/data/1418091/000156459016021507/twtr-ex991_6.htm, (Dec.6, 2016, Access).

[2] J. James，“Domosphere Musings, Insights, and Creative Solutions from Our Very Own Domosapiens,”https://www.domo.com/blog/data-never-sleeps-3-0/, (Dec.6, 2016, Access).

[3] 渡邊 恵太，加藤 昇平，“Twitter における語の関連性に着目したユーザ興味語抽出手法の提案,” 人工知能学会全国大会論文集，pp.1-4，2012.

[4] 山本 達也，芋野 美紗子，土屋 誠司，渡部 広一，“嗜好に基づく時事情報推薦システムの構築,” 情報処理学会研究報告，Vol.2013-ICS-170，No.1，pp.1-6，2013.

[5] tamu515@Twitter，“Twitter 日本 フォロワー数 総合ランキング 1-50 位,” http://meyou.jp/ranking/follower_allcat, (Dec.1, 2016, Access).