Twitter を用いた各地域の紅葉状態を表す紅葉マップの作成

江村　歩†　苅米志帆乃‡

†‡長野工業高等専門学校 電気電子工学科 〒381–8550 長野県長野市徳間 716
E-mail: †15206@g.nagano-nct.ac.jp, ††s_karikome@nagano-nct.ac.jp

あらまし インターネットには多種多様な情報があり，紅葉情報も簡単に収集することができる．しかし得られる情報は予測情報や過去の紅葉状態であることが多い．そのため，実際に紅葉を観に行く場合，その時点での紅葉状態を正確に知ることは難しい．SNS の一つである Twitter は発言した時点での状況や状態に関する情報を得ることができる．そこで本研究では Twitter を用いて，地域ごとに紅葉状態を自動的に抽出する手法を提案する．紅葉状態は，落葉，落ち始め，見頃，色づき始め，青葉の 5 つに分類する．さらにそれを分かりやすく可視化するために紅葉マップを作成する．

キーワード 紅葉マップ，紅葉状態，Twitter，可視化

1. はじめに

秋の季節になれば紅葉が日本のいたるところで観られる．その美しさに惹かれ，紅葉を観ることを目的として出かける人もある程度いる．そのような人達が観光として紅葉を観に行く際，紅葉についての情報を把握するため，事前に調べていく人も多い．インターネットが世に普及した昨今，そのような情報は比較的新しく，且つ容易に調べることができる．紅葉状態に関する情報だけでも，インターネット上で調べることができるサイトは多数存在している．例えば，Walker+[1]というサイトでは紅葉が最も見頃になる日付の予測を公開している．また，weather news[2]というサイトでは約 1 週間前の日付の紅葉状態を表している場所もある．このようにネット上で紅葉名所を検索し，紅葉状態を調べることができるが，事前に上記のようなサイトや方法で調べ，実際に紅葉を観に行くと，これらから得た情報と違うことが多くある．予測情報は外れることもあり，実際の紅葉状態と異なることがある．また，紅葉は移り変わりが激しいので過去の情報では現在の紅葉状態と異なることが多い．このように事前に調べた紅葉情報と実際の紅葉状態が違うことが上記のサイトでは起きてしまう．SNS はリアルタイムで更新されるため，投稿時点での状況を把握できる．今回は SNS の一つである Twitter を用いて地域ごとに紅葉状態を分類し可視化する手法を提案する．紅葉状態は落葉，落ち始め，見頃，色づき始め，青葉の 5 つに分類する．

先行研究として遠藤らの研究[3]では Twitter を用いて紅葉のピーク期の推定をしている．この研究は Twitter を用いてリアルタイムの情報を反映しているため実際の紅葉状態との乖離は非常に少ない．他にも，永井らの研究[4]では人工衛星とウェブサイト上の紅葉情報を比較し，その有用性を評価している．この研究では画像処理の技術を用いて色を判断している．

2. 手法

2.1 概要

本研究では Twitter を用いて，紅葉状態を自動的に抽出し，分類する手法を提案する．手法の流れを図 1 に示す．

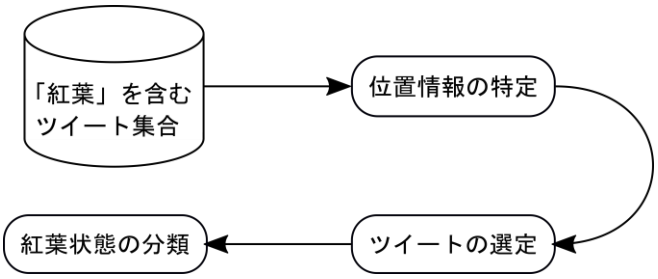

図 1 手法の流れ

まず，「紅葉」という単語を含むツイートを収集する．次に，そのツイートの位置情報の有無を確認する．そして，紅葉状態を表すツイートかどうかを選定し，紅葉状態を表すツイートの中から各地域の紅葉状態の分類を行う．

2.2 位置情報特定手法

本研究での位置情報とは一般的に用いられる，GPS によって自動的に表示される地域の情報だけではなく，位置が特定できるような表現も含む．ここでは GPS によって表示される地域に加えて，ツイート本文中や，ハッシュタグ内に書かれている地域を位置情報とする．また，ツイートの本文に位置情報が含まれているかどうかを判断するために，作成した「地域辞書」と照合する．

2.3 紅葉状態を表すツイートの選定

紅葉状態を分類するにあたって，そのツイートが紅葉状態を表すツイートか，紅葉状態を表していないツイートかを選定する．紅葉状態を表すツイートは過去形で記述されていることが多い．例えば，過去形である例では，

「綺麗な紅葉が見られました.」

がある．これは実際に現地へ観に行き，紅葉状態を表すツイートである．一方で，過去形でない例では，

「紅葉を観に行きたいなぁ...」

紅葉状態を表しておらず，個人の願望を表現したツイートである．したがって，過去形の有無で判断することでツイートを選定する．具体的には，形態素解析を行い,助動詞「た」があれば過去形であると判断する．

2.4 紅葉状態の分類

本研究では，紅葉状態を落葉，落ち始め，見頃，色づき始め，青葉の 5 つに分類する．各ツイートはこれらのいずれかの紅葉状態を表しているものとして分類を行う．分類するにあたって,「予測情報」,「特有の形容詞や，形容動詞の有無」,「ネガティブ語とポジティブ語の有無および数」の 3 つを判断材料として用いる．

2.4.1 予測情報による分類

予測情報を確認する. 1 章の序章でも述べたように,インターネット上には紅葉の見ごろ予測を提示している.これだけでは実際の紅葉状態を表していないため,実際の紅葉状態と違うといった問題が生じていたが,この予測情報を判断材料の 1 つとして使うことで，より精度をあげていくと期待できる．ツイートの位置情報と予測情報の場所を参照し，ツイートされた日付が見頃予測の前後 3 日間の場合「見頃」, 1 週間後の前後 3 日間の場合「落ち始め」, 2 週間後の前後 3 日間の場合「落葉」,1 週間前の前後 3 日間の場合「色づき始め」, 2 週間前の前後 3 日間の場合「青葉」と出力する．

2.4.2 形容詞，形容動詞による分類

判断材料の 2 つ目として，特有の形容詞や形容動詞の有無を確認する．各紅葉状態には，それぞれ似たような形容詞，形容動詞が用いられることが多い．例えば,「落葉」特有の形容詞や，形容動詞がツイート文中に含まれていた場合，紅葉状態が「落葉」である可能性が高い．つまり，特有の形容詞や，形容動詞を調査することで紅葉を分類できる手がかりが得られる．その形容詞を探すために,係り受け解析器を行い,「紅葉」に係っている形容詞および形容動詞を，紅葉状態ごとに収集する．

2.4.3 ネガティブ/ポジティブ語による分類

判断材料の 3 つ目として，ネガティブ語，ポジティブ語の数を確認する．ポジティブ語が含まれる場合，見頃に近く，ネガティブ語が含まれる場合その青葉もしくは落葉に近いと仮定する．ネガポジ語の数を調べるにために日本語評価極性辞書を用いる．辞書のネガティブ語の単語がツイート内の本文に含まれる場合，ネガティブ語の数をカウントする．また，ポジティブ語の単語がツイート内の本文に含まれる場合，ポジティブ語の数をカウントする．ネガティブ語の場合は予測結果と合わせて用いる．予測が青葉の場合，ネガティブ語の数が多ければ多いほど青葉に近い値になるようにする．従って，予測結果が青葉の場合ネガティブ語の数に青葉の数値を掛け合わせる．予測結果が色づき始めの場合ネガティブ語の数に色づき始めの数値を掛け合わせる．同様に，予測結果が落ち始めの場合ネガティブ語の数に落ち始めの数値を掛け合わせ，予測結果が落葉の場合ネガティブ語の数に落葉を掛け合わせる．予測結果が見頃の時，予測日より前だった場合はネガティブ語の数に色づき始めの数値を掛け合わせ,予測日より後だった場合はネガティブ語の数に落ち始めの数値を掛けた合わせた値とする．

2.4.4 総合分類

上記で述べた 3 つの判断材料と表 1 を用いて，総合的に判断した結果を分類結果として出力する．分類の際，紅葉状態を数値化する．各紅葉状態に対する数値は表 1 の通りとする．

表 1 各紅葉状態に対する数値

落葉	落ち始め	見頃	色づき始め	青葉
+2	+1	0	-1	-2

総合的に判断する際に以下の式を用いる．

$$P = (P_1 + P_2 + P_3)/3 \quad (1)$$

ただし，P は分類結果，P_1 は予測情報による結果，P_2 は形容詞による結果，P_3 はネガポジによる結果である．例えば，下記のようなツイートがあったとする．

「鎌倉の紅葉綺麗だった」

ツイート日が 11 月 18 日，予測情報が 11 月 20 日だとすると，$P_1 = 0$，$P_2 = 0$，$P_3 = 0$ となる．結果，この場合 $P = 0$ で,「見頃」に分類される．

3. 評価実験

3.1 評価手法

「紅葉」を含むツイートを収集し，人手で正解データを作成した． 100 件のツイートから位置情報が含まれたツイートは 76 件あった.逆に位置情報が含まれていなかったツイートは 24 件あった.位置情報が含まれたツイートの中から，紅葉状態を表しているツイートは 63 件ありその内，落葉が 7 件，落ち始めが 16 件，見頃が 15 件,色づき始めが 15 件,青葉が 10 件あった．2 章で述べた分類手法を用いて自動で分類した結果と,先ほど述べた手動で分類した結果を比較し，その精度を評価する．

3.2 位置情報特定手法

今回，位置情報は総務省ホームページより都道府県コード及び市区町村コード[5]を用いて対応した語がツイート内本文，もしくはハッシュタグ，位置情報欄に含まれる場合は位置情報があると判断した．地域辞

書には総務省ホームページ全国地方公共団体コードより，「都道府県コード及び市区町村コード」，他にもWikipedia より日本紅葉の名所 100 選[6]や，日本の観光地一覧[7]を参照した．従って今回の辞書は市町村名を含め，有名な観光地名まで検索することができた．

今回収集したツイートの位置情報はこれらの地域辞書で全てを参照することができた．収集してきたツイートが今回はたまたま有名な地名ばかりが書かれていたが，マイナーな観光地名や，区名までは参照できないため，辞書の追加を検討する余地がある．

3.3 紅葉状態を表すツイートの選定評価

先ほど述べたように，位置情報が含まれていて尚且つ,紅葉状態を表しているツイートは 100 件中 63 件あった．この 63 件のツイートを，今回の手法である過去形を調べた結果 43 件取得できた.取得できなかった数は 20 件だった．取得率は約 68.3%である．取得できなかったツイートを以下に示す．

「嵐山の紅葉は今が見頃ですよ.」

例では，実際の紅葉状態を表していると推測できるが，過去形の表現がされていない．今回の手法を用いた場合このようなツイートがノイズとして残ってしまう．精度を向上させるためには，このようなツイートも取得できるような手法が必要である．過去形の有無の他に傾向がないか検討する必要がある．

3.4 紅葉状態の分類結果

分類結果を表 2 に，分類材料ごとの正答数を表 3 に示す．

表 2 紅葉状態の分類結果

分類材料	分類数	正しく分類した数	誤って分類した数	精度［%］
予測情報	63	22	41	34.9
形容詞	40	18	22	45.0
ネガポジ	38	6	32	15.8
総合	63	28	35	44.4

表 3 正しく分類できたツイート数

分類材料	落葉	落ち始め	見頃	色づき始め	青葉
予測情報	5	0	4	4	9
形容詞	0	7	6	3	2
ネガポジ	0	0	6	0	0
総合	2	10	7	7	2

分類材料の 1 つ目である予測情報による分類の結果は，63 ツイートの中から正しく分類できたツイートは 22 件であった．精度は 34.9%であった．予測結果に出力された分類配分は落葉が 25 件，色づき始めが 1 件，見頃が 6 件,色づき始めが 11 件,青葉が 20 件あった．今回は予測情報として，日本気象協会が運営しているtenki.jp[8]の紅葉見ごろ情報を用いた．以上の分類配分より，「落葉」や「青葉」と予測した結果が多かった．つまり，「落ち始め」,「見頃」,「色づき始め」の予測精度が極端に低かったということになる．これは，予測の間隔が 1 週間だけであったので期間の余裕が少なかったことが原因であると考えられる．

分類材料の 2 つ目である形容詞による分類の結果は，63 ツイートの中から 40 ツイートが何かしらの形容詞が含まれていた．その中から正しく分類できたツイートは 18 件あった．精度は 45.0%である．

今回，163 件のツイートを手法で述べたとおりに調査した．特有の形容詞を調べた結果を表 4 に示す．

表 4 各紅葉状態の特有の形容詞

落葉	落ち始め	見頃	色づき始め	青葉
残った	まだ	綺麗な 美しい 真っ赤な	少し	早い

調査対象のツイート数が少なかったことが原因でそこまで多くの形容詞や，形容動詞を得られなかった．この問題は数を増やすことで解決できると思われる．また，今回は「紅葉」に係っている形容詞や，形容動詞が少なかったため,取得した特有の形容詞が「紅葉」に係っているかどうかは関係なく，文中に含まれているかどうかを調べた．その原因も加えて今回のような精度になったと考えられる．

分類材料の 3 つ目であるネガポジによる分類の結果は，63 ツイートの中から 38 ツイートがネガポジの辞書に書かれた単語が含まれていた．その中から正しく分類できたツイートは 6 件あった．精度は 15.8%である．今回はネガポジの辞書として，インターネット上で無料配布されている日本語評価極性辞書[9]を用いる．非常に，低い値となった．明らかに紅葉の見頃と比べたらネガティブな状況にも関わらず，たとえ青葉でも「素晴らしい」といった単語が多く見られた．全体的に今回集めてきたツイートは非常にポジティブな言葉を選んで使っている様な印象があった．従って，ポジティブの判定が多くされた一方，ネガティブな単語が少なかった．

3 つの判断材料を用いて，式 1 にあてはめて，総合的に分類した結果は 63 件のツイートの中から 28 件だけであった．精度は 44.4%である．今回用いた手法ではあまり望ましい精度を得ることは出来なかった．しかし，今回評価実験で用いたツイートの数は非常に少なかったため，ツイートの数を増やして，さらなる評価を検討する必要がある．

4. 紅葉マップの作成

予備実験で収集し，分類した紅葉のツイートを用いて，最終的に紅葉マップを作成する．作成する紅葉マップのイメージを図 2 に示す．地域ごとに紅葉状態が示される．

図 2 紅葉マップイメージ

5. おわりに

本研究では，Twitter を用いて実際の紅葉状態を分類する手法を提案した．ツイートの選定は過去形の有無によって評価した．精度は 68.3%であった．紅葉状態の分類は特有の形容詞の有無，予測情報，ネガポジの数を調査することによって評価した．総合的な精度は 44.4%であった．全体的に結果の精度は実用的なものには至らなかった．現状では予測情報を用いた方がまだ有用性があるため，分類の精度をあげる必要がある．そのために，今後の課題として，より多くのツイートを分析することで新しい傾向が見え，分類材料の改変や増加することによって精度をあげることができると考えられる．更に，正確な統計上の計算を行うことによっても，良い精度が得られると考えられる．

参　考　文　献

[1] 紅葉情報 2019 -色づき情報を毎日更新– ウォーカープラス.https://koyo.walkerplus.com/. (2020 年 2 月 5 日参照).

[2] 紅葉情報 2019 名所の見頃予想なら｜紅葉 ch.ウェザース.https://weathernews.jp/s/koyo/. (2020 年 2 月 5 日参照)

[3] 遠藤雅樹,廣田雅春,大野成義,石川博.マイクロブログを用いた生物季節観測によるピーク期推定手法の検討, 2015.情報科学技術フォーラム講演論文集， No.14, Vol.2, pp.97-102

[4] 永井信,井上智晴,鈴木力英.秋の衛星季節学におけるウェブサイト上で公開されている紅葉情報の有用性, 2015.日本生気象学会雑誌 52 巻 2 号.

[5] 総務省｜電子自治体｜全国地方公共団体コード, 都道府県コード及び市区町村コード. https://www.soumu.go.jp/denshijiti/code.html. (2020 年 2 月 5 日参照).

[6] 日本紅葉の名所 100 選- wikipedia.https://ja.wikipedia.org/wiki/AC

[7] 日本の観光地一覧- wikipedia.https://ja.wikipedia.org/wiki/E3

[8] 紅葉見ごろ情報 2019 -日本気象協会 tenki.jp. https://tenki.jp/kouyou/.(2020 年 2 月 5 日参照)

[9] 小林のぞみ，乾健太郎，松本裕治，立石健二，福島俊一．意見抽出のための評価表現の収集．自然言語処理， Vol.12, No.3, pp.203-222, 2005.