選択式天気情報を用いた ソーシャルメディアからの有用投稿抽出

萩行 正嗣
株式会社ウェザーニューズ

hangyo@wni.com

1 はじめに

近年，マイクロブログなどのソーシャルメディアを活用することで，実世界の動向を掴むことが広く行われている．人手で全ての投稿を確認することは困難なこと，比較可能な指標が求められることなどから，キーワードの出現頻度など動向の性質を示す何らかの数値的指標を用いて定量的に実世界の動向を把握することがおこなわれている．一方，投稿を数値量として表現することは，自然言語などで表現された個々の投稿の情報を捨象することである．ソーシャルメディアを活用した意思決定や動向の詳細な分析の際には，これらの捨象された情報が重要な役割を果たす場合があり，実際の投稿を人の目によって確認したいニーズが存在する．しかし，上述のように全ての投稿を人間が確認することは困難であるため，有益な情報が得られそうな投稿を優先して確認するための指標が必要となる．

ソーシャルメディアから得ることができる情報はソーシャルメディアの性質に大きく依存し，また何を有益とするかは情報の利用者によって異なる．本研究では，ウェザーリポートという気象情報を中心としたソーシャルメディアを対象として研究を行う．ウェザーリポートは話題が気象を中心という性質がはっきりとしており，そこから情報を得ようとする利用者も気象に関連する情報を求めていると考えられる．本研究では，ウェザーリポートに含まれる自然文で記述されたコメントから気象状況を知るにあたって有用な情報を得ることを目的とする．

一般に気象状況を知る手段としては，アメダスなどによる定点観測，気象レーダーや気象衛星による面的な観測がある．また，ウェザーリポートの各投稿 (リポート) にはユーザーが入力した選択式の天気情報 (例: パラパラ，ザーザー，暑い，寒い) が含まれており，これも気象状況を知るために活用できる．これらの観測手段およびウェザーリポートのコメントの長所と短所をまとめたものを表 1 に示す．

ウェザーリポートによるユーザーからの投稿では，観測機による観測に比べると以下の 2 つの長所がある．1 つ目は，観測機による観測が困難な現象を捉えることができることである．このような現象には，道路冠水や積雪など意志判断を行うにあたって重要な現象も多く含まれる．例えば下記の (1) では，ユーザーの周辺の地点で冠水の被害があることが分かる．

(1) 所々冠水してます

2 つ目は，定性的ではあるものの，人間の感覚にあった表現が使用されており，情報の受け手に状況が伝わりやすいことである．数値による雨量表現よりも，下記の (2) の方が，多くの人にとって具体的な降雨の状況が想像しやすいと考えられる．

(2) バケツをひっくり返したような豪雨です

また，ウェザーリポート内の要素で比較すると，コメントによる記述は選択式の情報に比べて詳細な情報が得られるという長所がある．下記の (3) では，ただ寒いというだけでなく，ストーブをつけるほどの寒さであるという程度が分かる．

(3) 朝は肌寒くて、ストーブをつけた（≧∇≦）

コメントにはユーザーの感情が表現されることもあり，これも意思決定においては重要な手掛りとなる場合がある．下記の (4) は，大雨の際のリポートの例だが，防災上の対策等を行う意思決定において非常に有用といえる．

(4) 危険を感じます

これらをふまえ，本研究では，気象状況，被害状況への言及や感情表現が含まれるコメントを有用な情報とし，これらを含むリポートを抽出の対象とする．

*1 例えばアメダスの観測では雨と雪の判別が困難である．

表 1　観測手段の長所と短所

観測手法	長所	短所
アメダス	定量的な観測が可能	観測点が限られる
気象レーダー	面的な観測が可能	降雨を直接観測しているわけではない
		観測不能な場所 (ブラインドエリア) が存在する
選択式天気情報	観測機で観測困難な現象も観測できる[*1]	報告内容が限られる
コメント	報告の内容が多岐に渡る	定量的解析が困難

本研究では，選択式の天気情報を学習に用いることで，半教師あり学習的にコメントにスコア付けを行い，このスコアをもとに有用なリポートを抽出する手法を提案する．本研究の目的は，選択式の天気情報を予測することではなく，有用なリポートを抽出することである．しかし，コメントから選択式の天気情報を予測する枠組みでリポートに与えられるスコアは，有用なリポート抽出において有用であると考えられる．例えば，下記の (5) のリポートでは，選択式の天気は「ザーザー」であった．しかし，このコメントから雨に関する意志決定に有用な情報は得られるとは言い難い．

(5)　おはようございます。寒い朝でしたね

一方，コメントに対し機械学習的手法でスコアを付与する場合には，雨に関する情報が少ないため，低いスコアが与えられ，有用な情報はないと判断される．また，下記の (6) では，選択式の天気は「影うっすら[*2]」であったが，コメントからは雨について有用な重要が得られると言える．

(6)　南に凄い雨雲

このような場合でも，コメントに対して機械学習的にスコアを付与することで，有用なコメントと判定できる．

ソーシャルメディアからの情報抽出では，素性を抽出するための前処理として形態素解析や表現の正規化が行うことが多い．しかし，崩れた表現は著者の感情を表現している場合が多く，正規化を行ってしまうと，この情報を失ってしまうことになる．また，顔文字も著者の感情を表現するが，顔文字全体ではなく部分ごとの意味が重要となる場合がある．

(7)　今朝は寒い ((((；Д))))

((((；Д)))) は本来は恐怖などで震えることを表現する顔文字であるが，(1) では寒くて震えることを表現している．これは．顔文字両端の括弧の並びが震えを表現していることから判別できる．しかし，顔文字全体を一まとめにして扱ってしまうと，顔文字の部分ごとの情報を扱うことができない．そこで，本研究ではこれらの問題に対して頑健な文字 n-gram を単位として扱う．

表 2　ウェザーリポートの選択式の項目の例

天気情報名	選択肢の例
天気	影はっきり， 影なし， ポツポツ，パラパラ，ザーザー， フワフワ， シンシン， ドカドカ
体感	ムシっと暑い，ジリジリ暑い， 暖かい，ちょうどいい， 涼しい，肌寒い，寒い

本研究では，ウェザーリポートに含まれる選択式の天気情報を活用し，文字 n-gram と気象状況 (「雨」「雪」「暑さ」「寒さ」) との関連度を算出する．その関連度をもとにリポートのコメントと気象状況との関連度を算出し，有用なリポートを抽出する手法を提案する．

2　ウェザーリポート

ウェザーリポートは (株) ウェザーニューズの運営する気象情報に特化したソーシャルメディアである．1 日平均約 2 万通のリポートが投稿される．投稿されるリポートには，写真やコメント，GPS の位置情報に加え，表 2 に示したような選択式の天気情報が付与されている．リポートの話題は天気のことが中心だが，紅葉など季節に関する話題や体感などに関連する生活の話題なども多く含まれる．

3　提案手法

各リポート r は，$(wx, comment)$ で構成される．wx は対象とする気象状況 (「雨」「雪」「暑さ」「寒さ」) に対応する天気情報であり，「雨」の場合は表 3 の天気情報のいずれかの値を取る．$comment$ はリポートの本文である．

提案手法では，各リポートのコメントと気象状況との関連度を半教師あり学習的に算出し，この関連度が高いものを有用なリポートとする．具体的な関連度の算出は以下で説明する．

[*2] 薄曇りを示す．

表 3　天気情報と雨の関連度

天気情報 (wx)	w_{wx}
ポツポツ	1
パラパラ	2
サー	4
ザーザー	8
ゴォー	16
その他	0

表 4　天気情報と雪の関連度

天気情報 (wx)	w_{wx}
ベチャ (みぞれ)	1
チラ	2
フワフワ	4
バチバチ (あられ)	4
シンシン	8
ドカドカ	16
ブワァー (吹雪)	24
その他	0

表 5　天気情報と暑さの関連度

天気情報 (wx)	w_{wx}
少し暑い	2
ムシっと暑い	3
カラっと暑い	3
暑くなった	4
蒸し暑くなった	4
暑い	5
ジリジリ暑い	5
耐えられない暑さ	10
その他	0

表 6　天気情報と寒さの関連度

天気情報 (wx)	w_{wx}
涼しい	1
涼しくなった	1
涼しい気分になった	1
肌寒い	3
しっとり肌寒い	3
寒い	5
極寒	10
その他	0

3.1　n-gram 気象状況辞書の作成

提案手法では各文字 n-gram(cn) と気象状況の関連度 ($R(cn)$) を算出する．本研究では，文字 n-gram として，1-gram から 4-gram を利用する．まず，cn と天気情報 (wx) との $pmi(cn, wx)$ を以下の式で計算する．

$$pmi(cn, wx) = log\left(\frac{\mathrm{freq}(cn, wx) \times N}{\mathrm{freq}(cn) \times \mathrm{freq}(wx)}\right)$$

ここで freq(x) は x を含むリポートの総数であり，N は全リポート数である．

次に，$pmi(cn, wx)$ を重み付き和を以下の式で計算することで n-gram と気象状況の関連度 $R(cn)$ とする．

$$R(cn) = \sum_{\{wx|pmi(cn,wx)>0\}} pmi(cn, wx) \times w_{wx}$$

ここで w_{wx} は wx との気象状況との関連度に対応する値だが，本研究では人手により設定することとし，「雨」の場合には表 3 のように設定した．

3.2　リポートと気象状況の関連度算出

リポート r に対する気象状況との関連度 $R(r)$ は，$comment$ に含まれる n-gram に対する $R(cn)$ の合計をリポートの長さで正規化することで計算する．具体的には以下の式となる．

$$R(r) = \frac{\sum_{cn \in N\text{-}gram(comment)} r(cn)}{\mathrm{length}(comment)}$$

ここで，$N\text{-}gram(comment)$ は $comment$ に含まれる n-gram の集合である．

4　実験

4.1　実験設定

2014 年のウェザーリポート 6,986,600 通を利用して n-gram 気象状況辞書の作成とリポートと気象状況の関連度算出を行なった．「雨」「雪」「暑さ」「寒さ」に対応する w_{wx} は表 3，表 4，表 5，表 6 のように設定した．

表 7　$R(cn)$ の例 (雨)

cn	$R(cn)$
雨風です	105.1
雨が窓	104.3
で冠水	97.0
停電しま	93.1
学校は休	92.9
水たまり	51.0
今はパラ	11.0
足湯	1.0

表 8　$R(cn)$ の例 (雪)

cn	$R(cn)$
開かなく	237.0
強く吹雪	233.3
0cm 以	223.6
雪まみれ	206.4
雪降ろし	188.3
積雪は無	51.45
路はツル	15.6
今夜も暖	2.0

表 9　$R(cn)$ の例 (暑さ)

cn	$R(cn)$
の暑さで	88.3
あぢぃ	74.3
あつーい	73.0
ι（´Д	70.0
熱中症注	63.0
射しは強	34.3
一雨来る	13.0
また曇り	7.4

表 10　$R(cn)$ の例 (寒さ)

cn	$R(cn)$
して寒い	45.3
((+_	41.1
ストーブ	38.2
)}}	35.2
さぶっ	34.1
路凍結	28.9
月が綺麗	13.6
爽やかに	1.2

4.2　n-gram 気象状況辞書

作成された辞書の例を表 7，表 8，表 9，表 10 に示す．「雨:雨風です」[*3]，「雪:強く吹雪」など気象状況を直接表現する n-gram だけでなく，「雨:停電しま」「雨:学校は休」「雪:開かなく」といった気象状況の影響を表現する n-gram に対しても高い関連度を与えていることが分かる．また，「暑い:あぢぃ」「寒い:さぶっ」のような崩れた表現や，「暑い:ι（´Д」や「寒い :((+_」「寒い:)}}」のような顔文字にも関連度も与えることができている．「)}}」は顔文字の中で震えを示す文字列であり，顔文字の構成性を扱うことができるといえる．

[*3] 気象状況「雨」と n-gram「雨風です」の関連度を示す．

表 11　リポート関連度 (雨)

関連度	コメント
185.9	突然の豪雨です
177.2	横殴りの雨
154.0	土砂降り道路冠水
152.9	バケツをひっくり返したような豪雨です
118.5	激しい雷雨と落雷停電しました
61.9	危険を感じます
45.4	今、少しの間だけど停電しました（－＿－；）
35.5	空は霞んでいる。雨降りはじめた
20.4	散歩の途中東根温泉の足湯に浸かって来ました
10.6	おはようございます。今日はいい天気になっています。

表 12　リポート関連度 (雪)

関連度	コメント
509.2	積雪 10cm 以上
495.4	現在積雪 30cm 以上
279.3	道路上はまだ積雪は無し
250.5	通路が雪で埋まりドアが開かなくなりました (゜ロ゜;
199.9	駐車場の大型トラックのタイヤが埋まってます
100.1	雪が降ってきました (。＞д＜) サブっ (--;)
70.3	今日は時々雪がチラチラです
33.4	足元から冷たい
12.3	今夜も冷え込んでいる。
3.5	青空～。

4.3　リポートの気象状況関連度

リポートと各気象状況の関連度の例を表 11，表 12，表 13，表 14 に示す．これらの結果から，算出された関連度がコメントの示す気象状況を捉えられていることが分かる．一方，「積雪 10cm 以上」の関連度が「現在積雪 30cm 以上」より雪との関連度が高くなっている．これは，後者は「現在」という言葉が入っているため，長さで正規化した結果，関連度が小さくなってしまうためである．このように関連度のみで細かい状況の推定は困難であるが，人が確認する際に有益なリポートを抽出するという目的は達成することができている．

5　関連研究

気象状況と自然言語文の対応付けを行った研究としては，Liang ら [2] による天気予報の数値データと概況文の対応付けの研究がある．彼らの研究では，降水などの数値や時間を自然文と対応付けいている．数値データと概況文という性質上，体感などは扱われていない．

崩れた表現を扱った研究としては，Brody ら [1] の極性判定の研究がある．彼らは，文字の繰り返しによる表現を正規化する手法を提案しており，そのような表現が極性判断において重要であるとしている．この研究では，繰り返し以外の崩れた表現は扱っていない．

表 13　リポート関連度 (暑さ)

関連度	コメント
161.1	耐え難い暑さ。
143.4	真夏の暑さです
108.1	あつーい！水分補給忘れずに
106.5	汗が流れる暑さです (; ´д｀)
93.9	蒸し暑いのでエアコン ON!
73.6	蒸し暑い (>_<)
60.7	水分補給しま～す≧∀≦／
40.4	積乱雲発達中!!!
10.1	激しい雨は通り過ぎたようです。
0.3	冷え冷えの寒い朝

表 14　リポート関連度 (寒さ)

関連度	コメント
98.2	放射冷却で寒い朝
93.5	今朝も寒い ((+_+))
68.8	道が凍結して寒いです。
64.5	今朝は寒い ((((；゜Д゜))))
51.0	朝は肌寒くて、ストーブをつけた（≧∇≦）
36.9	昨日とくらべて、肌寒く感じます。
30.0	風が非常に強いです。
25.9	雨、涼しい朝
10.2	風もなく、暖かい
1.3	爽やか

6　まとめと今後の課題

本論文では，メタ情報を活用することで，ソーシャルメディアから意思決定に活用できる投稿を抽出する手法を提案し，気象情報においてその有効性を確認した．今後はユーザーからの報告だけでなく，気象観測のデータなどとの関連性を算出していく．

謝辞

本研究に利用したウェザーリポートを投稿してくださった，ウェザーリポーターの方々に深く感謝します．

参考文献

[1] Samuel Brody and Nicholas Diakopoulos. Cooooooooooooooollllllllllllll!!!!!!!!!!!!!! using word lengthening to detect sentiment in microblogs. In *Proc of EMNLP 2011*, pp. 562–570, July 2011.

[2] Percy Liang, Michael Jordan, and Dan Klein. Learning semantic correspondences with less supervision. In *Proc. of ACL 2009 and IJCNLP 2009*, pp. 91–99, August 2009.