

Twitter データを用いた天候の主観的評価

渡辺 優樹[†] 廣田 雅春^{††} 石川 博^{†††} 横山 昌平^{†††}

[†] 静岡大学情報学部情報社会学科 〒432-8011 静岡県浜松市中区城北 3-5-1

^{††} 大分工業高等専門学校 〒870-0152 大分県大分市大字牧 1666

^{†††} 首都大学東京システムデザイン学部 〒191-0065 東京都日野市旭ヶ丘 6-6

^{††††} 静岡大学情報学部情報学部 〒432-8011 静岡県浜松市中区城北 3-5-1

E-mail: ^{††††} tia13100@s.inf.shizuoka.ac.jp, ^{††} m-hirota@oita-ct-.ac.jp, ^{†††} †ishikawa-hiroshi@tmu.ac.jp,

^{††††} ††††yokoyama@inf.shizuoka.ac.jp

あらまし 本研究では、Twitter に投稿された天気に関連した用語を含むジオタグ付きのツイートを用い、気温や天候などの客観的な評価ではなく、体感的な風の強さや寒暖などの主観的な天候の評価手法を提案する。現在用いられている気温と湿度から計算する不快指数などの客観的な評価では、気流や熱放射などが考慮されておらず、屋外における体感的な寒暖等を表すことは難しい。本研究では、ある地域の天気における特徴的な語の発見を行い、作成した天気関連用語辞書に基づいて快・不快指数を算出し、主観的な天候の評価を行うことで新たな指標を提案する。また、主観的な天候の評価を行うことによって、風が強いなどといった地域独自の特徴的な天候の発見が可能になる。

キーワード Twitter, 天候, 位置情報

1. はじめに

天気予報を見る際、晴れや雨といった天気や気温といった情報によって、私たちはその日の過ごしやすさを考える。天気や気温といった情報は客観的な指標であるため、実際に外出してみると、雨がそんなに降っていなかったり、思っていたより暑かったり寒かったりすることがある。このような場合において、天気や気温といった情報はあてにならない。そこで、天気や気温といった客観的な天候の評価だけでなく、実際に感じる体感での寒暖といった主観的な天候の評価を行うことができればよいのではないかと考えた。

天候に対する体感を表す現在用いられている指標として、不快指数が知られている。不快指数は人間が生活する上で不快と感じる体感を、気温と湿度を用いて表した指標である。ただし、不快指数は気流と熱放射の影響を含んでおらず、屋外環境における実際の体感とは異なる。[1] 気流や熱放射を考慮した指標としては SET (標準新有効温度) と呼ばれるものがあるが、屋内に特化した指標であり、これも屋外環境における実際の体感とは異なる。[2] 実際に一日の生活を行う上では、屋内環境・屋外環境の両方で行動する機会が多いため、両者に対応した総合的な評価指標が必要であると考えられる。

また、ヒートアイランド現象といった都市部ではコンクリートによって熱の逃げ場が無くなることで熱中症になる患者が増えているといったことの要因として、都市部での天候の観測所とビル街での体感とで実際の温冷感が異なることが考えられる。熱中症や凍傷といった症状を回避するために、実際の体感を知ることによって対策を講じることができると考える。

本研究では、SNS 上の天気・気候に関する個々人の感想を集めて分析する事により、不快指数等に代わるような、屋内環境・屋外環境に対応した主観的な指標を実現する。天候の主観的な

評価を行うために、本研究ではマイクロブログ Twitter^(注1) の Tweet のうち、ジオタグと呼ばれる、発言位置の緯度経度情報が付与されたものを用いる。Twitter はリアルタイム性の高いマイクロブログとして利用されており、ユーザは日々の出来事を気軽に発信している。そのため、天気や気候に関する個々人の感想も多い。

類似したものとして weathernews^(注2) が挙げられるが、天気を伝えるものとして主に利用されており、晴れ・雨といった情報と暑い・寒いだけの情報がほとんどであり、それに付随する寒い日はお酒で暖まるなどといった副次的な情報を得るのが難しいため、主観的な天候を表すのに不十分であると考えられる。

天候の主観的な評価の手法については、天気関連用語辞書を作成し、用語ごとに数値を割り当て用語の頻度、ツイート数によって快・不快指数を算出する。そして、天候に関する眩きに対して寒いという単語が含まれる文章には雪という単語が含まれやすいといったような共起度を計算することで、その地域独自の特徴的な天候の発見や実際の体感に近い指標の提案を行う。

本論文の構成は次のとおりである。2 章では、本研究と関連研究の差分について述べる。3 章では、提案手法の概要について述べる。4 章では提案手法の妥当性について実験を行い、それに対する考察を述べる。5 章では、本研究で得られた成果のまとめについて述べる。

2. 関連研究

屋外空間における温冷感の指標を得るための研究として木内 [3] の研究がある。木内は人体熱収支解析と呼ばれる気象条件、着衣状態、作業状態、発汗作用を考慮し、人体がどれだけ熱を

(注1) : <http://twitter.com>

(注2) : <http://weathernews.jp/index.html>

帯び熱が発散していくかという解析に基づく温度負荷量 TL と、風速と作用温度（気温が同じ室内でも、壁などの物体が放射する熱などによって変わる体感温度を加味した温度）の関数としてあらわされる温冷感指数 TSI の二つの指標を提案した。しかし、TSI はグローブ温度を用いており、グローブ温度は周囲からの熱放射による影響を観測するために仮想黒体の球を用いて測られる温度であるが、変動が激しく、地点によってばらつきがある。また、一地点あたりの被験者数が少なく、温冷指標として確立するまでに至っていない。

本研究では、その地域における実際の天候に対する眩きを集めることによって実際の体感に近いデータを用い、十分なデータ数を確保できる。そして、寒いという言葉とともに風が強いという言葉が用いられやすいといったような共起度を計算することによって、天候に対して実際の体感により近い指標の提案を行う。

Twitter からの天候データを収集することで天候のリアルタイムな状況を観察するという手法としては Jeff ら [4] の研究が挙げられる。この研究では、Twitter の StreamAPI から現在眩かかれているツイートを集め、そのツイートをジオタグがついているものとそうでないものに分類する。ジオタグがついているツイートの中から天候に関するワードを抽出することでリアルタイムでの天候の状況を把握している。しかし、ジオタグがついているツイートの数が少ないという問題があり、限定的な地域しか網羅できないという欠点がある。

本研究では、SearchAPI をともに用いることによって、この問題を解決し、天候の算出だけでなく、天候の評価も行う。

また、天候に関するテキストを本研究は収集するが、阿部 [5] は、Twitter での特徴的な語彙の収集において、Twitter のハッシュタグに着目し、基礎的な語彙などのノイズを排除していくことで、特徴的な語彙の収集に成功している。しかし、ハッシュタグを用いない場合での話題抽出に関しては、行えておらず、本研究では、ハッシュタグだけでなくハッシュタグが用いられていないツイートに対しても特徴的な語彙の抽出を行う。

青木ら [6] は、発信したツイートをどのような感情でつぶやいたかを推定する仕組みを作成している。本研究では、ネガティブやポジティブといった評価だけでなく、具体的な形容詞を抽出することによってより詳しくどのような思いで眩きがなされたのかを推定する。そして、個々のツイートではなく、全体的な評価としての良否を推定する仕組みを作成する。

3. 提案手法

本研究では、Twitter から収集した位置情報が付与された天候に関する眩きを用いて、天候の主観的な評価を行う。

3.1 概要

提案手法では、全国の都市を対象に、天候に関する眩きを集め、天気関連用語辞書をもとに快・不快指数を算出することで都市ごとの主観的な天候の評価をする。

まず、情報共有サイトである Twitter から、投稿日時、投稿されたテキスト、ジオタグなどの情報が付与されたメタデータを収集し、データベースに格納する。次に、収集した眩きを独自に

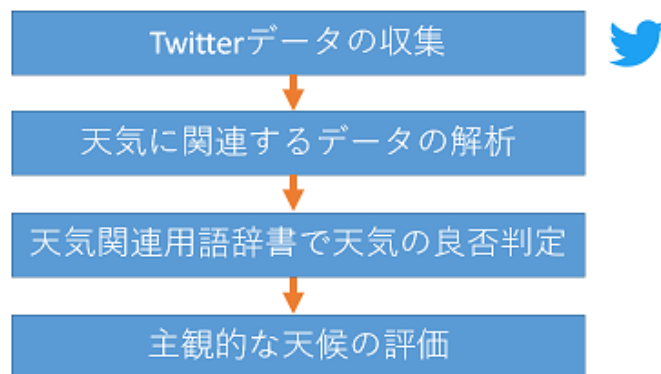


図1 提案手法

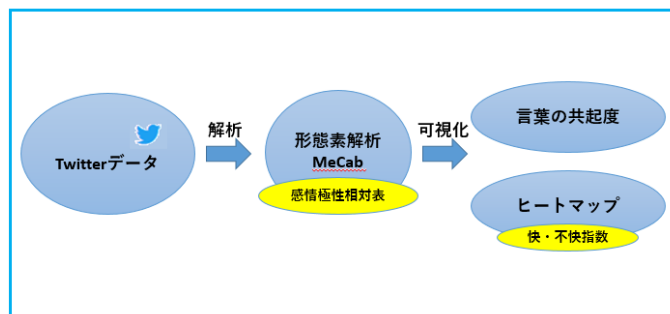


図2 システム概要図

表1 天気関連用語辞書

単語	読み	品詞	値
寒い	さむい	形容詞	-0.9926
暑い	あつい	形容詞	-0.988786
涼しい	すずしい	形容詞	0.97513
花粉	かふん	名詞	-0.445794
大風	おおかぜ	名詞	-0.994602

作成した天気関連用語辞書を元に形態素解析し、天候に関連する用語の頻度や共起度の計算を行う。そして、天気関連用語辞書で割り振られた指数を用いて、快・不快指数を算出する。そして共起度と指数の値をもとにその日の天候の主観的な評価を行う。提案手法の概要を図1に示す。システム概要図を図2に示す。

3.2 データセット

天気関連用語辞書は暑い、寒いなどの主観的な感情を表現した用語と天候や花粉などの客観的な事実を表した用語を登録し、それぞれに-1から+1までの指数を付与する。この指数の付与は高村らが作成した単語感情極性表 [7] に基づいて行う。単語感情極性表とはその語が一般的に良い印象 (+1) を持つか、悪い印象を持つか (-1) を二値属性で表した表であり、語彙ネットワークを利用して計算されている。

表1で天気関連用語辞書の例を挙げる。今回使用する天気関連用語辞書の単語数は48単語である。

3.3 データの収集

Twitter Streaming API を用いて、ジオタグが付与されている投稿を収集する。テキスト、投稿日時、位置情報を収集し、データベースに格納する。Streaming API では、瞬時性の高いツイー

トを取得することができ、データ取得制限がないため特定のワードにこだわらずデータを集めることができる。

ジオタグが付与されている投稿のみでは、十分なデータ数を集めることができないため、SearchAPI を使い天候に関連した単語を含む眩きも収集する。SearchAPI では Twitter でのキーワード検索を行うことができ、位置情報を指定して収集することや期間を指定して収集することができる。

3.4 データの解析

収集した Twitter データを天候に関連する用語の入ったテキストデータと入っていないデータに分類する。収集した Twitter データの集合に形態素解析を行い、天候に関連する用語の入ったデータだけを抽出する。分類には作成した天候関連用語辞書をもとに形態素解析エンジンの MeCab を用いて分類する。また、位置情報をもとに分類した天候に関連する Twitter データを都市ごとに分類する。

そして、天候に関連する用語の入ったデータに対して、天候関連用語辞書に登録されているワードと、そのワードとともに使われるワードを抽出し、共起度の計算を行う。また、共起度が低いと思われるものに対してはノイズとして判断し、天候に関連のあるツイートだけが選ばれる。

3.5 天候の良否判定

天候関連用語辞書に割り振られた指数を用いて、都市ごとに分類されたツイートにそれぞれ指数をかけ、-1 から 1 までの値で都市の天候を評価する。1 に近いほど快適であり、-1 に近いほど不快である。この快・不快指数を用いて都市をヒートマップで表す。ヒートマップでの不快度の表し方には、眩かれた Twitter データによって算出された快・不快指数の値を利用し、値が 0 以下である場合に地図上にマッピングされる。赤く表示されている地域ほど天候の評価を不快とする。

ヒートマップは、マップデータとして Google Maps API を使用し、ヒートマップのレイヤの表示は Google Maps JavaScript API を使用した。

3.6 主観的な天候の評価

天候の評価には、単語同士の共起度を計算することで、雪が降っているから寒い、日差しが強いから暑いといったような要因で現在の天候が構成されているのかということや花粉はどれくらいなのかといったことを算出し、先ほどの快・不快指数によるヒートマップと組み合わせることで主観的な天候の評価として利用する。これにより、その地域が快適な温度で過ごせるのかどうかということと現在の気候がどのような要因で構成されているのかということがわかる。

主観的な意見から算出された快適かどうかと現在の気候の様子がわかることによって、新しい主観的な天候の評価手法として提案する。

4. 実験

本研究では、実際の天候と比較して提案手法が適切に寒暖等を表現できているかについての実験を行った。

4.1 類似した気候条件における比較

類似した気候条件で、人によって天候に対する感じ方に差異

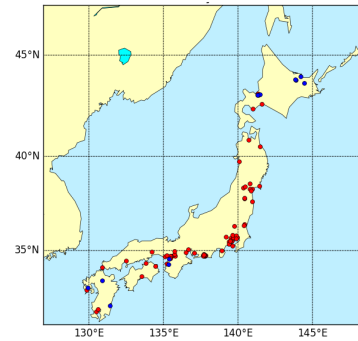


図 3 類似した気候条件における比較

表 2 2014/07/12 当日の天気

都市	天気	最高気温	最低気温
札幌	曇後雨	21.9 °C	17.5 °C
仙台	晴後曇	31.5 °C	20.3 °C
東京	曇	32.8 °C	26.5 °C
名古屋	晴後曇	33.4 °C	22.7 °C
大阪	曇	32.0 °C	24.3 °C
福岡	曇後雨	31.3 °C	23.8 °C

表 3 2017/1/15 不快指数 浜松市

時間	温度	湿度	不快指数	体感
6 時	-1.5 °C	83%	32	寒い
12 時	2.7 °C	52%	42.4	寒い
18 時	1.8 °C	58%	40.5	寒い
24 時	-0.3 °C	85%	33.6	寒い

があるのかを実験した。今回の実験では、Twitter から収集した 2014 年 7 月 12 日のデータから抽出された暑い・寒いという天候に関連する用語が入った 105 件のデータを地図上にマッピングし、可視化することで類似条件での天候に対する感じ方の差異の有無を確認した。

実験結果を図 3 に示す。図 3 の赤丸は暑いと感じた眩きがされたことを表し、青丸は寒いと感じた眩きがされたことを表している。暑いと感じたデータの件数は 93 件、寒いと感じたデータの件数は 12 件であった。また、主要都市における当日の天気については表 2 に示した。

4.2 ヒートマップを用いた実際の不快指数との比較

気温などの数値から算出される不快指数は、現在ではエアコンがかけられていた部屋から出た場合に感じる気温など、実際に感じる不快度と異なっていると考えられる。そこで、ヒートマップを用いて AMeDAS での気温・湿度から算出された不快指数とツイートから算出した不快度を比較する。

AMeDAS での気温・湿度から算出された不快指数は 2017 年 1 月 15 日の浜松市のデータを用いた。表 3 にて当時の不快指数を算出した。また浜松市内における 2017 年 1 月 15 日のデータを利用した。データ件数は 109 件であった。図 4 にヒートマップでの結果を示す。このヒートマップは眩かれたツイート数が多く、不快指数が大きいほど色が赤く表示される。

4.3 共起度を用いた実際の気候との比較

実際の何°C といった気温だけでは、暑いか寒いかを総合的に



図4 2017年1月15日の浜松市内のヒートマップ

表4 寒いという単語と共起した単語

順位	単語	出現頻度
1	雪	24
1	今日	24
3	浜松	15
4	風邪	10
5	路面	9

表5 2017/1/15 浜松市内における実際の天候

天候	平均気温	平均湿度	不快指数	独自の指標
雪時々曇後晴	0.6	68	37.4	-0.87

判断することは難しいと考える。その地域の風の強さや地理的条件によって同じような気温でも天気の違いは異なる。そこで、寒暖や風の強さといったものをツイートから算出した不快指数をもとに共起度でランキング化し、実際の真夏日や真冬日とされる気候とで比較する。表4にて、例として寒いという単語で同時に使われた単語として多かったものを表す。

この実験では、SearchAPIによって「寒い」という単語のツイートを収集した。データは浜松市内における2017年1月15日のデータを利用した。データ件数は109件であった。表5は実際の浜松市内における天候^(注3)である。

4.4 tf-idf法を用いた天候の特徴抽出

tf-idf法を用いて一日の天候に対しての特徴語抽出を行った。一日の天候の特徴語を抽出することによって、その日一日がどのような日であったかだけでなく、地域ごとの天候の評価に対して重みづけを行うことができる。たとえば、北海道では、冬に雪が降るということを当然のように考えているため、雪が降ったかどうかではなく、降雪量などによって天候の評価が変わる可能性がある。浜松では、あまり冬に雪が降らないため、主観的にいつもより寒いと感じてしまう場合が考えられる。このような地域差を、tf-idf法によって特徴語を抽出することで表すことができるのではないかと考え、実験を行った。

tf-idf法とは特徴語を算出するための方法である。tf値とidf値を掛けることで算出することができる。tfとは単語の文書内における出現頻度のことであり、頻出する単語であればあるほど重要とするものである。tf値は次の式によって求められる。文

書d内におけるある単語tのtf値を $tf(t,d)$ とすると、

$$tf(t,d) = \frac{\text{ある単語 } t \text{ の文書 } d \text{ での出現回数}}{\text{文書 } d \text{ での全ての単語の出現回数の和}}$$

で表される。また、idfとはそれぞれの単語が他の文書内においても共通して使われているかを表しており、いくつもの文書内で使われている単語はある文書において重要ではない単語とするものである。idf値は次の式によって求められる。ある単語tのidf値を $idf(t)$ とすると、

$$idf(t) = \log \frac{\text{全文書数}}{\text{ある単語 } t \text{ が出現する文書の数}} + 1$$

と表される。この二つを掛けて算出された値が大きいほど、その文書内における特徴語として考えることができる。

今回の実験では、2017年1月14日から2017年1月20日までの一週間の浜松市内におけるTwitterデータを用い、tf-idf法によって特徴語の抽出を行った。一日に呟かれたTwitterデータを一文書として今回は扱う。表6は浜松市内における一週間の中での一日の特徴的であるとされた単語を5番目まで表している。

表6 浜松市内における一週間の中での一日の特徴語

1月14日	1月15日	1月16日	1月17日	1月18日	1月19日	1月20日
雪	路面	傘	インフル	バリーダンス	コタツ	しょうが
明日	真っ白	降水確率	熱	病院	スイッチ	電車
勉強	凍結	昨夜	受験	インフルエンザ	ミツパチ	そっち
空模様	昨夜	イマソラ	お天気	今夜	春	ヒーター
部活	温度	朝	お出かけ	レッスン	スキー	気温

4.5 不快指数と独自の指標との相関

本手法における独自の指標である快・不快指数について、実際の天候と比較して相関があるかを調べた。相関があるかを調べることによって、本手法における指標の妥当性を検証した。

今回使用したデータとして2016年1月1日から2017年1月1日における一年間の浜松市内における天候に関して呟かれたデータを利用した。データ件数は、32301件であった。

表7は浜松市内における不快指数と独自の指標との相関を表している。

5. 結果・考察

5.1 類似した気候条件における結果・考察

季節では夏であるため、暑いと感じる人が多いが、北海道を見ると、寒いと感じている人が多い。当時の天気と見比べてみると、ほかの都市に比べて10℃ほど最高気温が低かったためであると考えられる。

当時の天気では、東京と大阪を比べると共に曇であり、最高気温が32℃前後で最低気温が24℃前後であるため、似たような気候条件であるが、大阪では寒いと感じている人のツイートが見られた。実際のツイートの内容の例を表8に挙げる。表8のツイートの内容で、1のテキストではその日の天候について寒いと語っていると推測できる。2のテキストでは、室内における空調の温度に対してであり、実際の天気とは関係ないことがわかる。

今回の実験では、少ないデータ数ではあるが、同じような条

(注3) : <http://www.jma.go.jp/jma/menu/menureport.html>

表 7 浜松市内における不快指数と独自の指標との相関

期間	値
1月	0.504088
4月	0.007724
8月	-0.61147
10月	0.364894
年間	-0.01298

表 8 実際のツイートの内容の例

	テキスト
1	今日寒い絶対笑
2	むり休憩室クーラー効きすぎ寒い鳥肌立ってるから!!!

件でも寒いと感じる人もいるということがわかった。このことから、地域によって寒暖を感じる度合いが異なる場合があるとわかった。しかし、実際の天気と関係のないツイートを取得してしまう可能性があるため、これは寒いという単語に対してクーラーという言葉が用いられている場合にデータを省くといった共起度を用いることによって排除する必要がある。

5.2 実際の不快指数との比較結果・考察

実際の不快指数を全体を通してみると、寒いという結果になっている。ヒートマップでも赤に染まっており、不快であるということがわかる。このことから、天気関連用語辞書から算出された快・不快指数における値と実際の不快指数とで、関係があるということがわかる。

課題として、ヒートマップの表示方法として、データ数と独自の指数を掛けて図に示しているため、データ件数が少なく、一つのデータによる快・不快指数の影響が大きいことである。また、該当地域でのツイートが少ないと、ヒートマップの見た目としても貧相なものになってしまう。

また、データ数が多ければ多いほど不快として表示してしまう可能性が高いため、都市圏になればなるほど不快な地域ということになってしまう可能性もある。

5.3 実際の気候との比較結果・考察

浜松市内での2017年1月14日の実際の天気は雪であり、当時のツイートの内容を見ると寒いという言葉とともに雪という言葉が一番多く使われていた。理由としては初雪であったということが挙げられる。普段あまり雪が降らない地域であり、雪が降っているということを共有したかったためであると考えられる。

また路面といった情報から、道路が雪によって凍結されているといった可能性があるということがわかる。この実験から、どのような天気を表しているかということと実際にその天気をどのように感じているかということをついートから理解することはできるといったことがわかった。また、道路の路面状況といった副次的な情報も判別できると考えられる。

5.4 tf-idf法を用いた天候の特徴抽出 考察

浜松市内における特徴語を算出した場合、浜松では雪が降ることが珍しいため、降雪があった1月14日と1月15日では、雪に関連するワードが多く見られた。この結果から、雪があまり降らない地域においては、雪が降ることによる影響が大きいと

考えられ、地域によって雪が降ることによる影響の差異を考慮した指標を作成することができるのではないかと考えられる。

1月17日や1月18日では、インフルエンザや病院といった単語が特徴語として挙げられている。当時、浜松市内でインフルエンザが流行していたという可能性があると考えられ、この情報から該当地域での病気に対する対策が必要といったことを想定して、一日を過ごす必要があると考えることができる。

1月19日と1月20日では、コタツやヒーターという暖房器具が特徴語として挙げられており、とても寒い日であったということが想定できる。しかし、実際の気温を見ると1月14日から1月20日の中では、1月19日、1月20日の順に平均気温が高いという結果になっていた。最高気温が高く、暖かい気温から寒い気温に朝から夜にかけて寒くなっていったため、このような結果になったと考えられる。

1月18日のベリーダンスやレッスンといった単語は、天候に関する特徴語としては関係があまりあるとは考えられず、tf-idf法の特徴から1週間の中でその日しか使われることがなかったためにスコアが高くなってしまったと考えられる。このような特徴語を省くために、天気関連用語辞書を強化し、天気関連用語辞書に存在する単語のみのスコアを参照するといった方法が考えられる。

5.5 不快指数と独自の指標との相関 考察

1月という括りや8月という括りでは、暑い・寒いという片方の不快を表すことができるため、相関関係があった。

しかし、年間や4月や10月では、暑い場合と寒い場合とで快・不快指数の値が変わらないため、相関が見られないということになった。暑い場合と寒い場合とでの不快な思いの違いを、今回の感情極性相対表ではネガティブとして同じ括りで扱っているからだといえる。

また、不快指数との相関が一部見られたことで、天候に対してなんらかの相関関係があるということが考えられ、本手法における指標を用いて、天候の主観的な評価を表すことができるとわかった。

6. おわりに

本研究では、Twitterから取得したテキストデータとそれに付与されているジオタグを用いて、地域ごとに天気の主観的な評価手法を提案した。実験では、ツイートを分析することによって主観的な天候を表すことができた。

しかし、今回の実験では、天気関連用語辞書での取り扱いワードが少ないため、抽出できたデータ数が少ない。そのため、今後は天気関連用語辞書での取り扱い用語をさらに増やしていく必要がある。

また、天気の良否判定における指数による地図の分類方法が、天候の主観的な評価に対して適切かどうかということ、共起度におけるノイズ削除について考えていく必要がある。また、今回の実験では、不快指数の相関について求めたが、不快指数だけでなく、独自に算出した指標がどれくらいの値になると熱中症や凍傷といった危険性が高まるのかといった実際の天候と比較して考えていく必要がある。

文 献

- [1] 井幕 知伸, 堀越 哲美, "日本における夏季の不快指数の経年変化と分布に関する研究", 人間と生活環境, 18(2), pp.67-82, 2011
- [2] 都市緑地を活用した地域の熱環境改善構想検討委員会, "都市緑地を活用した地域の熱環境改善構想の検討", 平成 16 年度 ヒートアイランド現象による環境影響に関する調査検討業務報告書, pp.8-16, 2005
- [3] 木内 豪, "屋内空間における温冷感指標に関する研究", 社団法人日本気象学会 天気 48(9) pp.661-671, 2001
- [4] Jeff Cox, Beth Plale, "Improving Automatic Weather Observations with the Public Twitter Stream", IU School of Informatics and Computing, 2011
- [5] 阿部 一哉, "Twitter を利用した, 特定の話題に特徴的な語彙の収集", 跡見学園女子大学人文学フォーラム vol14, pp.32-40, 2016
- [6] 青木 誠, 勝山 亮, 木村 敦, 冬瓜 成人, 島田 尊正, 宮保 憲治, 新津 善弘, "Twitter システムにおけるツイートによる感情推定法とその評価", 電子情報通信学会総合大会講演論文集 pp.49, 2016
- [7] 高村大也, 乾孝司, 奥村学 "スピンモデルによる単語の感情極性抽出", 情報処理学会論文誌ジャーナル, Vol.47 No.02 pp. 627-637, 2006.