Vol.2014-MBL-73 No.10
Vol.2014-ITS-59 No.10
2014/11/20

Tweet 内容に影響を与える気象条件と特徴語の抽出

伊藤拓[†1]　深澤佑介[†2]　朱丹丹[†1]　太田順[†1]

気象条件はユーザの行動とツイート内容に影響を与える．我々は，晴れか雨かといった天気概況よりも，気温と湿度の方がユーザの行動に影響を与えることを発見した. 詳細に言えば, ツイート内容が変化する閾値としては, 気温 9℃と湿度 42%が最適であることを発見した.

Climate Condition that Mostly Affects the Change of Tweet Content

TAKU ITO[†1]　YUSUKE FUKAZAWA[†2]
DANDAN ZHU[†1]　JUN OTA[†1]

Climate condition affects users' action and tweet content. We discover that temperature and humidity affects users' action more than the general weather category such as sunny or rainy. In detail, we discover that 9 degree of temperature and 42% of humidity are the best thresholds to affects the change of tweet content.

1. 序論

近年，ビッグデータの普及により，膨大な情報源の中からユーザの嗜好に合った情報を見つけ出し，情報推薦に利用する推薦技術が注目されている[1].ユーザの嗜好を推定する上の情報源として，ユーザが簡単に自身に関する情報を投稿可能な Twitter を利用した研究が行われている[2].
また，ユーザの置かれている状況（コンテキスト）が嗜好に影響を与える重要な要素として考えられており，コンテキストを用いてユーザの嗜好を推定する研究が多くなされている.例えばコンテキストとしては「場所」「時間」が挙げられ，場所や時間は入手可能性および影響度の高さからコンテキストの一部として非常に重要視されてきた[3]．一方で, ユーザの嗜好に影響を及ぼすコンテキストとして「天気」が挙げられる．Yahoo によれば，天気によってその日の予定を変更したことがある人は全体の 54%いるという報告がある[4]．また，観光者が行先を決定する重要な外的要因として天気が挙げられている[5]．天気というコンテキストによってユーザの嗜好が変わることは明らかであり，嗜好推定において天気は重要なコンテキストであると言える．コンテキストを用いた嗜好推定に関する従来の研究においては, 位置情報を利用した行動予測[6][7]や時間による行動予測[8], 同行者に応じた行動予測[3]といった手法が提案されている.しかし，天気は嗜好を左右する重要な要素であるにも関わらず，どの天気が具体的にどのように人の嗜好に影響を及ぼすかについて，定量的に分析している研究は未だない.

そこで本研究では，天気コンテキストという観点から，Tweet 内容の変化に着目することによって，天気がユーザの嗜好にどのように影響を与えるかについて検証する．また，コンテキストごとによく使われる名詞を抽出する．

天気は一般的に「晴れ」「曇り」などの言葉で定義されることが多いが，ユーザの行動に影響を与える要因としてそのほかのさまざまな気象条件（気温や湿度，風速など）も考慮する必要がある.本研究では，気象庁が発表している天気情報[9]のうち，「気温」「湿度」「風速」「降水量」「雲量」「日照時間」「天気概況」を気象条件として扱う.「天気概況」とは，「晴れ」や「雨時々曇り」などその日の天気の概況を文字で表したものである．本研究は，上に挙げた複数の気象条件のうち，どの気象条件がユーザの嗜好に影響を与えるコンテキストとして適切かを検証することを目的とする．

本研究のチャレンジングポイントについて説明する．上で述べたように，天気コンテキストは複数の気象条件から構成されている．また，各気象条件は独立にコンテキストとなっているのではなく，その組み合わせによって天気コンテキストが成り立っているという点が，場所や時間と異なる困難な点である．

上記チャレンジングポイントに対するアプローチを述べる．各気象条件について，複数の閾値を設けてクラス分類することによって，気象条件ごとに適切な閾値を求める．次に気象条件間での比較を行い，嗜好により影響を与えるだろうと推定される気象条件を組み合わせることで，上記問題を解決する．

†1 東京大学
University of Tokyo
†2 （株）NTT Docomo
NTT Docomo

Vol.2014-MBL-73 No.10
Vol.2014-ITS-59 No.10
2014/11/20

2. 問題設定

本研究では，位置情報を含んでいる Tweet のうち，2011 年 5 月から 12 月までの土日に東京都で呟かれた日本語の Tweet 約 100 万件を分析対象とする．土日限定にした理由は，平日は仕事や学校など天気に関係なく行動する人が多いため，平日の Tweet よりも土日の Tweet の方がより天気に影響された特徴をもつだろうとの仮説に基づくものである．

前節で述べた内容を踏まえ，Tweet 内容に注目し，天気コンテキストにおける各気象条件のうち，どの気象条件がユーザの Tweet 内容に影響を与えるかを求める．また，各気象条件において，どの閾値を境界として Tweet 内容が変化しやすいかについても考える．嗜好に影響を与えているかの評価方法については 4 節で述べる．

3. 提案手法

3.1 手法概要

手法のフローチャートを図 1 に示す．

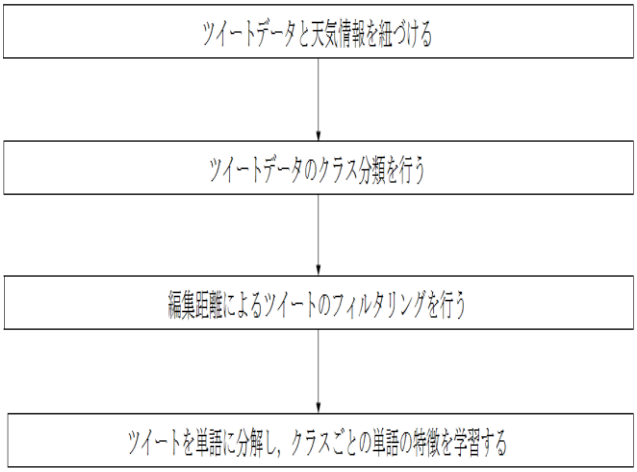

図 1　手法フローチャート
Figure 1　Flowchart of Method

はじめに，ツイートデータと天気情報とを位置情報をもとに紐づける.紐づける天気情報は，気象庁が発表しているもので，「気温」「湿度」「風速」「降水量」「雲量」「日照時間」「天気概況」の 7 つである[9].

次に，ツイートデータのクラス分類を行う.気象条件ごとに複数の閾値を設定し，2 クラス分類を行う.例えば，気温であれば 17° C， 18° C，……，27° C と閾値と設定し，その気温よりも高いときに呟かれたツイート，低いときに呟かれたツイートの 2 つにクラス分類する．

クラスによる単語の特徴を学習する際，bot とよばれる機械による自動投稿の宣伝は天気に関係なくツイートされるため，ノイズになると考えられる．そのため，2 つのツイートがどれだけ近いかを表す指標である編集距離を用いて，直近のツイートで編集距離が近いツイートをデータから排除することでノイズ除去処理を行う．

ツイートデータのうち 8 割を学習用データ，残りの 2 割を訓練用データに分ける．単語の特徴を学習するため，すべてのツイートデータを名詞に分解する．分類器には線形分類器である Stanfordclassifier[10]を用いる．学習用データを分類器に入力することにより，分類器はツイートに含まれる名詞がどちらのクラスで多く使われているかを学習し，重み係数とともにどちらのクラスの名詞かという結果を出力する．重み係数が大きい名詞は，どちらかのクラスに偏って多く呟かれている名詞であるので，そのクラスでの特徴語とする．

3.2 手法詳細

(1) ツイートデータと天気情報との紐づけ

ツイートデータには位置情報として，緯度，経度が含まれている．緯度経度の情報を都道府県情報に変換する逆ジオコーディングソフト[11]を用いて，ツイートデータにどの都道府県で呟かれたかという情報を付与し，東京都で呟かれたツイートを抽出する．

気象庁から東京都の東京観測所での天気情報をダウンロードする．天気情報は，1 日ごとに「気温」「湿度」「風速」「降水量」「雲量」「日照時間」「天気概況」が記録されている[9]ため，ツイートデータに含まれている日付情報と比較して，各ツイートに上記天気情報を紐づける．

(2) ツイートのクラス分類

天気情報が紐づいた各ツイートについて，気象条件ごとに閾値を設定して 2 クラス分類を行う．閾値の詳細は表 1 に載せる．

表 1　各気象条件と閾値の設定方法
Table 1　Thresholds on Each Climate Condition

気象条件	単位	刻み幅	最小値	最大値
気温	℃	1	8	28
湿度	%	3	42	78
風速	m/s	1	0	12
降水量	mm	3	0	15
雲量		1	0	10
日照時間	h	1	0	12

天気概況については，その日のメインとなる天気について，雨が降っているかいないかの 2 値化を行う．例えば，「雨一時晴れ」であれば「雨」のクラスであると分類する．設定した閾値について，ツイートの天気情報の値と閾値を比較し，全ツイートを閾値よりも高いクラスと低いクラスに 2 クラス分類する．

2 要素の組み合わせについては，1 要素でのクラスタリングを行ったのち，次節で述べる評価実験を行い，評価が高かった要素同士を組み合わせる．2 要素を組み合わせることで 4 つのクラスに分類することができるので，そのうち

の 1 クラスとそれ以外の 3 クラスを分ける組み合わせを 4 通り行う．

(3) 編集距離によるフィルタリング

分類器を用いて，クラスごとの特徴を学習させる．Bot などの宣伝は天気に関係なくツイートされるため，特徴語を抽出する際のノイズとなる．したがって，概要で述べたように 2 つの文章がどれだけ似ているかを示す編集距離を用いて，フィルタリングを行う．

編集距離は，2 つの文字列が与えられたときに，文字の挿入，削除，置換を行うことで一方の文字列を他方の文字列に変換する最小回数として定義される．

本研究では，あるツイートを取り出したときに，そのツイートを取り出す前に取り出した 1000 件のツイートとの編集距離を比較し，20 以下であった場合にそのツイートは前と似たようなツイートであり，bot である可能性が高いことからデータとして採用しない，というフィルタリングを行った．

(4) 分類機を用いた特徴学習

日本語の形態素解析ソフト Sen[12]を用いて，全ツイートを単語に分解し，ツイートごとに名詞を抽出する．名詞のみを抽出したツイートのうち 8 割を学習用データ，残り 2 割を試験用データとする．

学習用データを用いて，ツイートに登場する名詞のクラスごとの特徴を学習させる．分類器には線形分類器 Stanfordclassifier[10]を用いる．学習用データを分類器に入力すると，分類器は各名詞がどのクラスに多く登場するかを学習し，名詞ごとにどちらのクラスに属するかという結果と，その名詞がクラスを特徴づける重み係数を出力する．重み係数は，クラスを線形分離したときの，分離平面を表す各名詞の係数を表している．よって，重み係数が大きい名詞はそのクラスをより特徴づける特徴語ということができる．

4. 評価実験

今回得られた各クラスの特徴語について，正しくクラスの特徴を表しているかを評価するために評価実験を行う．評価実験には，前節で説明した試験用データを用いる．

本研究の目的は，Tweet 内容に影響を与える気象条件とその閾値を求めることであるから，天気によってクラスタリングした結果，クラスによってツイートに現れる名詞が分離されることが望ましい．分類器が，学習用データで学習した各単語の重み係数を用いて，試験用データに含まれるツイートのクラスを推定し，その推定の精度が高ければ，分類器は正しくクラスの特徴を学習していたと言える．

そこで，本研究では評価関数としてクラスタリングの性能を示す F 値を用いる．F 値は次式により計算される．

$$F = \frac{2}{\frac{1}{P} + \frac{1}{R}} \tag{1}$$

ここで，P は適合率を表している．あるクラスだと判定されたツイートのうち，実際にそのクラスに属しているツイートの割合が適合率である．R は再現率を表している．実際にあるクラスに属しているツイートのうち，そのクラスであると判定されたツイートの割合が再現率である．P と R の調和平均である F 値は，分類器がどれだけ正しくツイートをクラス分類できたかという性能を示すものであり，この値が高いほど分類器による特徴の学習精度が高かったと評価することができる．

次に，出力された特徴語が，本当に天気コンテキストに関係のある単語なのかを検証するために，出力された各単語についての主観的な評価のためのアンケートを行った．アンケートの内容としては，出力された単語をクラスごとに提示し，その単語がそのクラスのときに話題になりそうか，という質問に対して（1:全く関係していない，2:あまり関係していない，3:どちらともいえない，4:関係している，5:強く関係している）の 5 段階で評価してもらう．アンケート調査は 8 人に行った．

5. 結果

表 2 に，それぞれの天気要素のクラスタリングにおける F 値の最大値を載せる．

表 2　各天気要素と *F* 値の最大値との関係

Table 2　Max F-value for Each Climate Condition

天気要素	*F* 値（最大値）	条件
平均気温	0.64	9°C
平均湿度	0.62	42%
最大風速	0.58	8m/s
降水量	0.56	15mm
平均雲量	0.58	3
日照時間	0.55	8 時間
天気概況	0.56	

この結果より，F 値の最大値をみると F 値が高い天気要素は気温，湿度，風速の 3 つであると言うことができる．

この 3 つの条件について，分類器により出力された名詞のうち，重み係数上位 10 件を表 3-5 に載せる．

Vol.2014-MBL-73 No.10
Vol.2014-ITS-59 No.10
2014/11/20

表 3　気温 9°C を閾値としたクラス分類の出力結果

Table 3　Output of Classification on Temperature Threshold

9°C 未満	9°C 以上
月食	梅雨明け
東京モーターショー	網戸
クリスマスイブ	
酉の市	
年末年始	
年の瀬	
ハイデン	
銀世界	

表 4 湿度 42%を閾値としたクラス分類の出力結果

Table 4　Output of Classification on Humidity Threshold

42%未満	42%以上
古戸	ラテン
忘年会	梅雨明け
年越し	
クリスマス	
矢崎	
受け売り	
全編	
大晦日	

表 5 風速 8m/s を閾値としたクラス分類の出力結果

Table 5　Output of Classification on Wind Velocity Threshold

8m/s 未満	8m/s 以上
散会	おろか
木立	正明
チョウ	ガラス張り
野川	春日部
遠方	
なつい	

次に 2 条件の組み合わせの結果を載せる．F 値をより高くするために，天気要素の組み合わせとして，F 値が最も高い気温と，それに準じる湿度，風速との組み合わせ 2 通りについて実験した．

2 条件を組み合わせると，2.3.2 で述べたように 1 つの閾値に対して 4 回クラスタリングを行うことができるので，今回はそのうち F 値が最大となった結果を図 2, 3 に載せる．

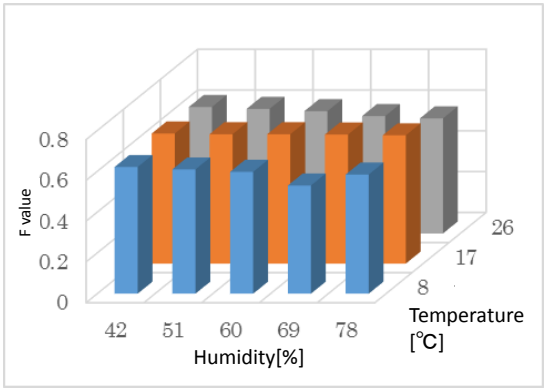

図 2　平均気温と平均湿度の組み合わせにおける各閾値での F 値の最大値

Figure 2　The change of F-value according to the combination of temperature and humidity thresholds.

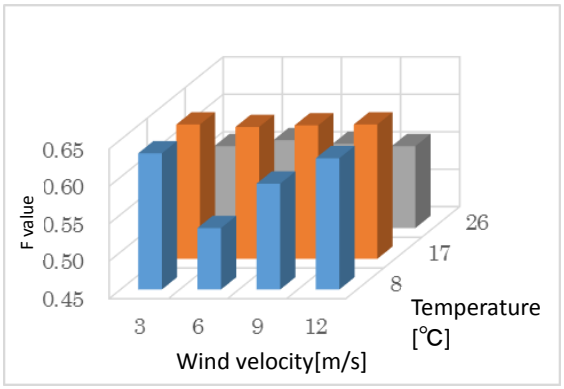

図 3　平均気温と最大風速の組み合わせにおける各閾値での F 値の最大値

Figure 3　The change of F-value according to the combination of temperature and humidity thresholds.

図 2, 3 ともに縦軸が平均気温，高さ軸が F 値となっており，図 2 の横軸は平均湿度，図 3 の横軸は最大風速を表している．

図 2 より，気温と湿度の組み合わせで F 値が最大となるのは，気温 17° C と湿度 42%の組み合わせ，図 3 より，気温と風速の組み合わせで F 値が最大となるのは，気温 17° C と風速 6m/s の組み合わせであると分かる．

F 値が最大となるこの 2 つの組み合わせにおける分類器からの単語の出力結果のうち，重み係数上位 10 件を表 6, 7 に載せる．

表 6　気温と湿度の組み合わせにおける出力結果
Table 6　Output of Combination of Temperature and Humidity

気温 17°C 未満 かつ 湿度 42%未満	気温 17°C 以上 または 湿度 42%以上
忘年会	ローラン
古戸	山道
酉の市	軽自動車
クリスマス	
焼きうどん	
イルミネーション	
年末年始	

表 7 気温と風速の組み合わせにおける出力結果
Table 7　Output of Combination of Temperature and Wind Velocity

気温 17°C 未満 かつ 風速 6m/s 未満	気温 17°C 以上 または 風速 6m/s 以上
酉の市	ファンファーレ
ハロウィン	網戸
イルミネーション	エスペラント
年越し	
古戸	
年末年始	
抗体	

次に，2 条件の組み合わせにおける，F 値とアンケート結果を表 8 に載せる．アンケートは，2 条件の組み合わせ 2 通りについて行った．

表 8　各天気要素と *F* 値，アンケート結果の関係
Table 8　Climate Condition, F-value,and results of Questionia

クラス	*F* 値	アンケート結果	
		平均	標準偏差
気温 17°C 未満 かつ 湿度 42%未満	0.622	3.5	1.5
気温 17°C 以上 または 湿度 42%以上	0.654	2.0	1.2
気温 17°C 未満 かつ 風速 6m/s 未満	0.609	3.0	1.6
気温 17°C 以上 または 風速 6m/s 以上	0.643	1.9	1.5

6. 考察

6.1 気象条件間での F 値の差異に関する考察

4 節でも述べたように，F 値が高いクラス分類は，よりツイートの単語を明確に分離していると言える．よって，気象条件間での F 値の差異を比較することによって，どの気象条件がよりツイートの単語を切り分けられるかを比較することができ，ツイート内容へ与える影響の大きさを論じることができる．表 2 を見ると，雨が降っているかいないかを示す「天気概況」によるクラス分類よりも，「気温」「湿度」によるクラス分類の F 値の方が高いことが分かる．このことから，雨が降っているかいないかというコンテキストよりも，気温が高いか低いか，湿度が高いか低いか，というコンテキストの方がツイートの分類に適していると言える．

次に，2 要素を組み合わせた結果について考察する．最高値で比較すると，気温よりは F 値が低下している．これは，2 要素を組み合わせることによってクラス内のツイート数が少なくなってしまい，スパースになっているからであると推測する．

Vol.2014-MBL-73 No.10
Vol.2014-ITS-59 No.10
2014/11/20

6.2 出力された単語の考察

実際に出力された単語を見てみると，表 3 と 4 では「銀世界」「クリスマスイブ」など寒い時期に関係のある単語が出力されている．また，表 4 においては「梅雨明け」など，湿度に関係がある単語が出力されており，コンテキストの特徴語が出力されていることが分かる．また，気温，湿度ともに閾値よりも低いクラスの特徴語が多く抽出されていることから，2 クラス分類したときに，天気に関係のある特徴語が出やすいクラスと出にくいクラスに偏りがあることが分かる．

1 条件と 2 条件の組み合わせとの比較を行う．表 6, 7 で出力された単語を見てみると，1 要素同様，天気に関係のある単語が出力されており，1 要素のときには上位 10 件に入っていなかった「ハロウィン」「イルミネーション」「焼きうどん」といった天気に関係のある単語が上位に出力されている．このことから，2 要素を組み合わせることによって出力される特徴語が存在することが分かる．また、表 6 を見てみると，気温が低く湿度の低いクラスの単語が多く，1 要素のときと同様にクラスによって特徴の出やすさが違うことが分かる．

6.3 F 値とアンケート結果の関係の考察

考察のために，以下のようにクラスを定義する．

クラス 1: 気温 17°C 未満かつ湿度 42%未満
クラス 2: 気温 17°C 以上または湿度 42%以上
クラス 3: 気温 17°C 未満かつ風速 6m/s 未満
クラス 4: 気温 17°C 以上または風速 6m/s 以上

表 8 を見てみると，クラス 1 と 3 はアンケート評価が高く，F 値が低い．クラス 2 と 4 はアンケート評価が低く，F 値が高い．この理由として，まず上段で述べたように，特徴語が出やすいクラスと出にくいクラスがあることが挙げられ，クラス 2 と 4 はどちらも気温が高いクラスであり，1 要素で考察したように気温が高いクラスは天気と関係のある特徴語が出にくいクラスであったために，アンケート評価が低くなったのだと推測する．

クラス 1 と 3，クラス 2 と 4 を比較すると，F 値が高いほどアンケート評価も高くなっているため，F 値は分類器が正しくクラスの特徴語を出力できているかという性能を示す評価関数として妥当であると言える．

F 値が高いにも関わらずアンケート評価が低い単語について考察すると，クラス 2 において「山道」「ミステリー」などが特徴語として出力されている．これらの単語が実際に使われているツイートを調べると，山道を散策したり，ミステリーに関するイベントに参加したりするなど，そのクラスでの特徴を示している単語であると言える．このように，明らかに天気とは関係あるとは言えないが，天気コンテキストと関係のある特徴語を取り出すこともできた．

7. 結論と今後の展望

天気には晴れているか雨が降っているかだけではなく，気温や湿度，風速といった複数の要素が存在するが，本研究では，どの気象条件によって Tweet 内容が変わるかを検証することを目的として設定した．気象条件ごとに複数の閾値を設定してツイートを分類し，分類器に特徴を学習させることによって, 天気ごとの特徴を示す単語を抽出した．F 値による分類器の評価を行うことにより，気温，湿度，風速の 3 条件がユーザの嗜好に影響を与えるコンテキストであることを発見した．また，実際に出力された特徴語を見ることで，気象条件に応じた特徴を表す単語が出力されていることを確認した．F 値と評価が一致していない，アンケート評価が低かった単語についても，その単語が使われているツイートを見ることによって，一見天気と関係なさそうに思える単語でも，天気コンテキストの特徴を表している可能性を示した．

今後の展望としては，上で述べたような自明でない天気の特徴も含めてクラス分類の性能を評価できるように，現在の評価関数である F 値を改良した客観的な評価関数の構築を進めていくこと，および天気によるコンテキストによる独立な影響であると評価できるような設計手法を考案することを行いたい．

参考文献

1) 奥健太，中島伸介，宮崎純，植村俊亮: “Context-Aware SVM に基づく状況依存型情報推薦方式”，日本データベース学会 letters 5(1), 5-8, 2006
2) 佐々木健太，長野伸一，長健太，川村隆浩: “Web 上のライフストリームからのユーザ行動情報の抽出”， 人工知能学会, 3F3-4in, 2011
3) 深澤佑介，太田順: “同行者に応じたトピックモデル”，情報処理学会論文誌 55(1), 413-424, 2014
4) Yahoo JAPAN Corporation, ”「天気」がインターネットユーザーの情報取得行動に与える影響”, http://promotionalads.yahoo.co.jp/online/blog/market/whitepaper_weather.html
5) Matthias Braunhofer, Mehdi Elahi, Francesco Ricci, Thomas Schievenin: Context-Aware Points of Interest Suggestion with Dynamic Weather Data Management, Information and Communication Technologies in Tourism 2014, 87-100, 2013
6) 斉藤祐樹，高山翼，山上慶，戸辺義人，鉄谷信二: “マイクロブログのジオタグと発言コンテキスト解析による行動予測手法”, 情報処理学会論文誌 55(2), 773-781, 2014
7) Eisenstein, J., O'Connor, B., Smith, N.A. and Xing, E.P.: A latent variable model for geographic lexical variation, Proc. EMNLP, 1277-1287, 2010
8) Kawamae, N.: Trend analysis model: trend consists of temporal words, topics, and timestamps, Proc. WSDM, 317-326, 2011
9) 気象庁，http://www.data.jma.go.jp/obd/stats/etrn/
10) Stanfordclassifier http://nlp.stanford.edu/software/classifier.shtml
11) 逆ジオコーダー, http://www53.atpages.jp/usoinfo2/urgeocoding.html
12) 形態素解析ソフト「Sen」, https://www.mlab.im.dendai.ac.jp/~yamada/ir/MorphologicalAnalyzer/Sen.html