ツイートデータに対するテキストマイニングによる持続可能な開発目標の社会的視座に関する調査

A survey on the social perspective of sustainable development goals by text mining on tweet data

北海道大学工学部環境社会工学科　○学生員　塘安奈(Anna Tomo)
北海道大学工学院工学研究院　正員　ヘンリー・マイケル(Henry Michael)

1. はじめに

持続可能な開発目標(SDG s)とは 2015 年の国連サミットで採択された「持続可能な開発のための 2030 アジェンダ」に記載された、環境、社会、経済に関する 17 の目標からなる 2016 年から 2030 年までの国際目標である。日本でも企業などを中心に積極的に取り組まれている。しかしより効率的に、具体的な活動をするためには現状を把握し市民の意見を明確にする必要がある。従来ならアンケート調査などを行いデータを収集してきたが、その方法では多くの時間と費用を必要する上にデータ数も限られてしまう。そこで近年急速に発達しているソーシャルネットワークサービス(SNS)に注目する。SNS により誰でも情報を容易に発信できるようになり、そこから得た大量なデータは「ビッグデータ」と呼ばれ多様な分野での活用方法が研究されている。SNS のなかでも日本でのツイッターの利用者は約 4500 万人と言われている。ツイッターとはツイートと呼ばれる 140 字以内のメッセージや動画、画像を投稿し、投稿を他人と共有するサービスである。その１日のツイート数は約 5000 万件にものぼるため、SDGs に関する投稿も多くあると考えられる。ゆえに本論文ではツイッターのビッグデータを用いてテキストマイニングを行い SDGs の社会的視座を明らかにする。本論文では 17 のゴールのうちの 11「住み続けられるまちづくりを」から 6 のターゲット「2030 年までに、大気の質及び一般並びにその他の廃棄物の管理に特別な注意を払うことによるものを含め、都市の一人当たりの環境上の悪影響を軽減する」に絞って分析した。11.6 は日本でも問題になっている大気とごみがテーマのターゲットであり、身近なテーマのため市民の関心が高く、データも多く取れることが期待される。

2. 調査概要

はじめにツイートを検索するためのキーワードを設定した。キーワードはターゲットの文章からメインワードとサブワードを抽出し、さらにその言葉を具体化したものや、反義語もキーワードとした。決定したキーワードは表１に示す。

次に決定したキーワードを用いてデータを収集した。Twitter developers(1)のサイトで API を取得し、その API を用いて R の rtweet(2)パッケージでキーワードに一致するツイートデータを取得した。データの取得条件は表２に示す。

次に集めたデータを整理した。整理は Excel と、正規表現と呼ばれる記号で文字列を表す表現方法を用いて行った。データ整理の詳細は表３に、関係のないツイートを削除するときに用いたキーワードは図１に示す。「空気」は事前に同義語として「大気」に変換した。

最後に分析を行う。分析は KHcoder(3)というソフトを使って行った。KHcoder とは立命館大学の准教授樋口耕一が作った日本語のテキスト計量分析のためのフリーソフトである。R と連携しているので、難しい操作なしに容易にテキストマイニングができるようになっている。

表１　11.6 キーワード表

メインワード	サブワード
大気/空気/排出ガス/ばい煙/粉塵/揮発性有機化合物/微粒子物質	管理/把握/汚/きれい/悪化/改善/対策
ゴミ/ごみ/廃棄物/びん/かん/ペットボトル	管理/取り締まり/分別/処理/増/減/多/少

表２　データ取得条件

取得期間	2019/9/26～2019/10/26
言語	日本語
データ種類	テキスト/日時/ユーザーID
その他	リツイートは含まない

表３　データ整理詳細

整理項目	手段	方法
重複	Excel	テキストデータの重複を削除
URL	正規表現	空白に置換
記号	正規表現	空白に置換
改行コード	正規表現	空白に置換
アカウント名	正規表現	空白に置換
関係のないツイート	正規表現	正規表現で検索し削除 検索ワードは（表５）

bot|お知らせ|ご案内|^【.*】|^◆|^■|^[.*]|空気.*凍る|空気.*読|空気.*ピリピリ|^ペットと住まう |場の空気|空気に流|取得できる資格と業務|^【.*】|^[.*]|Q:|イーブライト|ご..ください|分別奮闘記|ご..ください|クソ|やつ|みたいな|マスゴミ|カス|ゴミ過ぎ|ゴミすぎ|クズ|ゴミ情報|ゴミ人間|性処理|性欲処理|ゴミ度|ゴミレベル|ゴミデータ|スケジュール管理ゴミ|ゴミデータ|体調管理ゴミ|管理能力ゴミ|ゴミ野郎|ゴミ管理人|ゴミみたい|ゴミみて|ゴミ男|ゴミ女|ゴミ教師|本当にゴミ|ほんとゴミ|まじゴミ|まじでゴミ|人ゴミ|楽天|amazon|送料無料

図１　関係のないツイートの検索ワード

３．結果と考察

KHcoder で 11.6 を共起ネットワーク分析した結果を図２に示す。

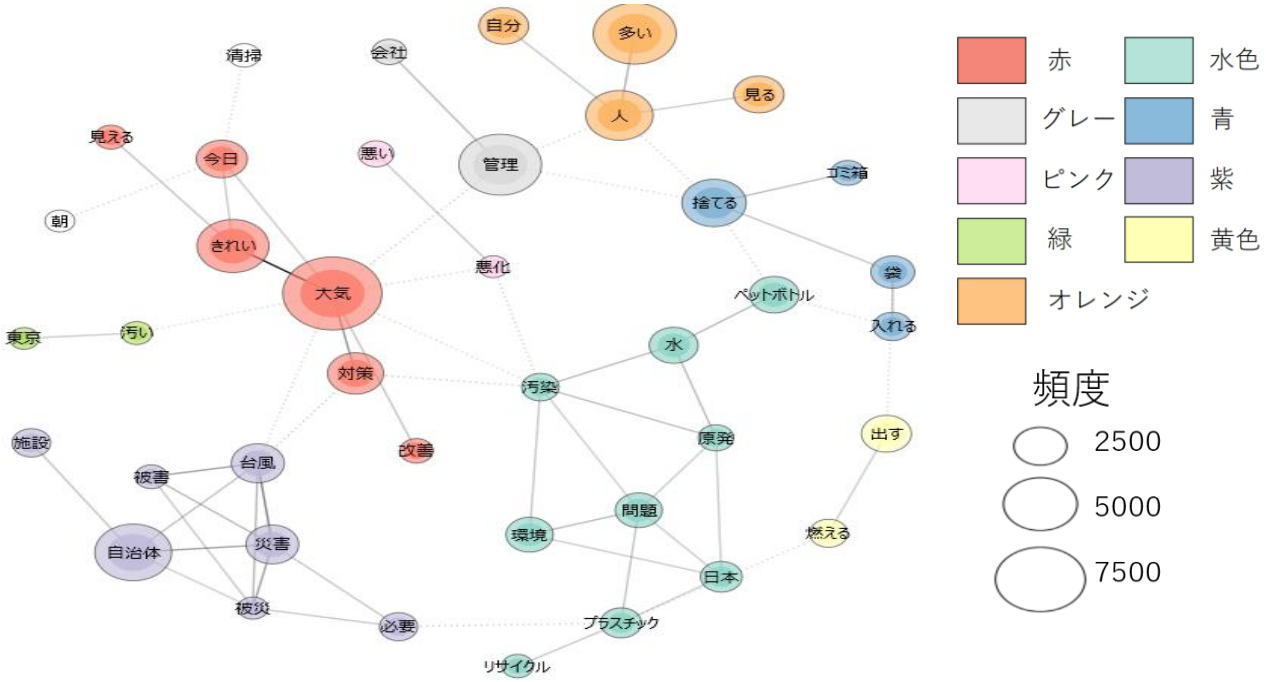

図２　11.6 共起ネットワーク分析結果

今回はどのような意見があるかおおまかに把握するためにツイート単位で共起ネットワーク分析を行った。これは線でつながっているバブル同士が関係性の強い語である。線の色が濃いほど強い関係性を示している。またバブルの大きさは語の出現頻度が多いほど大きくなっている。

赤のバブルに注目すると、「大気」「きれい」が非常に関連性の高い語であることがわかる。大気とつながっているピンクのバブルでは「悪化」や「悪い」、緑のバブルでは「汚い」とあることから、完全にきれいなわけではないが、バブルの大きさや、線の濃さからも大気がきれいだと思っている人が多いと推測される。さらに緑のバブルの「汚い」が「東京」とつながっているのを見ると、特に東京の空気が汚いと思っている人が多いことがわかる。このことから大気汚染については東京を中心に改善について考える必要があるだろう。

グレーのバブルを見ると「きれい」と同様に「管理」という語も「大気」との関係性が高い。しかし「大気」「管理」が一体何を意味するのかこれだけだとイメージしづらい。「管理」が「会社」とも関連があることを考えたとしても、大気に対しての意識が高い会社位しか推測できないので、ここはどうしてこの関係が生まれたのか文章検索などでもっと詳しく分析する必要がある。

またオレンジのバブルも同様で、「管理」「人」「多い」と関係かあることから大気の管理をしている人が多いのかと推測されるが、これもさらに詳しく分析する必要があるだろう。

水色のバブルに注目すると、「プラスチック」「問題」「環境」から、プラスチックが日本で環境問題になっていることがわかる。同時に「リサイクル」「必要」とも関係があることから環境問題はあるが、プラスチックは必要不可欠なものでもあり、リサイクルの意識もあることも推測される。また同じく環境問題で「原発」「汚染」の問題がある。東日本大震災で話題になり今も問題視されていることがわかる。この図から多くある環境問題のなかでも「プラスチック」と「原発」の問題が一番関心が高いということが推測される。

青のバブルは物を入れるための袋、また袋がゴミ箱に捨てられているということがわかる。黄色のバブルの「燃える」「出す」とも関係があることからゴミ袋やレジ袋の有料化などの問題が背景にあるのではないかと考えられる。

紫のバブルは９月にあった台風関係である。このデータはすべてゴミや大気に関係していることから、災害ゴミや台風後の空気に関するツイートがあったことが推測される。

４．まとめ

本研究の成果から、ツイートをテキストマイニングすることで市民が何に注目し、どのように思っているのかある程度推測できることが示された。これらの結果を踏まえて、今後は不鮮明な部分は語句を絞り調べたい言葉を中心にして分析して、より詳しく明確で正確な結果を出し、社会的視座を明らかにすることを目指す。

５．参考文献

(1) Twitter Developers
https://developer.twitter.com/content/developer-twitter/ja.html
(2) rtweet
https://cran.r-project.org/web/packages/rtweet/rtweet.pdf
(3) KHcoder　https://khcoder.net/