Twitter利用者の雪に係わる発言のテキストマイニングに関する研究

竹内　祥一[1]・萩原　亨[2]・高野　伸栄[3]

[1]学生員　北海道大学大学院工学院 北方圏環境政策工学専攻（〒060-8628 北海道札幌市北区北13条西8丁目）
E-mail:takeichi0415@ec. hokudai. ac. jp

[2]正会員　北海道大学工学研究科 北方圏環境政策工学部門（〒060-8628 北海道札幌市北区北13条西8丁目）
E-mail:hagiwara@eng. hokudai. ac. jp

[3]正会員　北海道大学工学研究院 北方圏環境政策工学部門（〒060-8628 北海道札幌市北区北13条西8丁目）
Email:shey@eng. hokudai. ac. jp

近年のブログやソーシャルネットワーキングサービスの発達は顕著である. 実際にそれらに投稿された内容というのは非常に価値のあるものだという見方も出てきている. 本研究では, SNSの一つであるTwitter上に投稿されるtweetの投稿内容に着目し, 「札幌」, 「雪」という単語を含むtweetから, 札幌における降雪がTwitter上での発言にどう影響を与えているかを可視化することを目的としている. 収集したtweetの投稿内容をKHcorderから形態素解析し, 共起ネットワークを構築, 投稿内容に降雪がどのように影響するかについて分析を試みた.

Key Words : *twitter, social networking service, text mining, content analysis, collective intelligence*

1.　はじめに

近年, ブログやソーシャルネットワークサービス（以下SNS）が注目されている. 近年の SNS に共通の特徴として, 「即時性」, 「双方向性」, 「波及性」が挙げられる. それらの特徴を強く持つ Twitter（http://twitter. com/）のアクティブユーザー数（Twitter 社に記録されている現在のユーザーの数）は 2012 年 4 月に世界で 1 億 4000 万人, 日本国内でも 2000 万人いると見られている. Twitter 上での投稿数は世界で 3 億 4000 万件／日と非常に利用者数の多いサービスとなっている. Twitter とは, 140 文字までの文章を投稿するシンプルなサービスである. 2006 年 7 月 Twitter 社（当時 Obvious 社）が米国でサービスを開始し, 2008 年 4 月, 日本国内でのサービスを開始した. また, 国内での Twitter 利用者の端末別利用割合は, 半数以上が外出先で利用できる携帯やスマートフォンからの投稿である. そのため, 投稿する際の時間や場所を選ばないという特性を持ち, リアルタイムでの投稿が可能となる. 実際に Twitter 上に書き込まれるものは, その時々の投稿者の目の前で起こったことや, 投稿者の気持ちを表すものが多く見られる. そのような特徴を生かし, 企業の中には Twitter 上に投稿される利用者の書き込みから, 商品に対するニーズを読み取るマーケティングや顧客の不満や疑問を早期に発見することに利用されたりしつつある.

Twitter 上の投稿は強制された発言とは違い, 利用者の独り言に近い. そのため「即時性」, 「双方向性」, 「波及性」を持った Twitter 上の発言内容は, 独立性が高く情報として価値が高い. そのため, Twitter 上の発言内容を集積し現実の世界で起きていることを検知するセンサーとしての可能性がある. 高橋らは, 実世界のセンサーとして Twitter の可能性について検証した. 実験の結果から, 花粉のような特定の分野においては tweet の数と実センサーから得られた情報の間に正の相関があることを示した [1]. 本研究では, 利用者の数も比較的多い札幌に焦点を当て, 「札幌」と「雪」という単語を含む Twitter への投稿を収集し, それを分析することでその時の札幌における降雪状況を知りうる, センサーとしての可能性について議論する. Twitter に投稿された発言内容に, 札幌における降雪状況やそれに付随する問題に関するものが多数あり, 「雪」と人々の反応の関係を読み取れる tweet が収集されることを期待した.

2. 分析手法

2.1 Tweetの収集

(1)日時

Tweetの収集には, Togetter（http://togetter.com/）を用いた. Togetterは, Tweetを取捨選択しまとめて保存することができるウエブサービスである. 本研究では, 「札幌」・「雪」を含むTweetを時間帯別に収集した. 「札幌」を入れた理由は, Tweetが発信された場所に関する情報を付加するためである. Tweetを収集し期間は平成23年12月1日から平成24年1月31日までの2ヶ月間とした. 時間帯は8:00から9:00, 17:00から18:00, 23:00から24:00とした. 各々の時間帯でTweetを収集し, 保存した. 8:00から9:00は通勤, 通学時間である. 17:00から18:00は帰宅時間である. 23:00から24:00は, 一日のうち最もTweet数が多い時間帯である. 朝と夕方の2つの時間帯は, 利用者移動しており, 降雪状況, 路面のすべり, それらによる交通機関の障害などをTweetするユーザーが多いのではないかと考えた. 深夜の23:00から24:00は, Twitterの利用数が多くなり, 雪に関する話題が含まれたTweetが多くなることを期待した.

(2)Tweetの種類

収集したtweetは「一般的なtweet」, 「@から始まるtweet」, 「文中にRTを含むtweet」の3つである.

・一般的なtweet：@から始まらない, 文中にRTを含まないもので, その時の本人が一番伝えたいことを示している.

・@から始まるtweet：関連tweetと呼ばれ, 特定のユーザーに宛てたtweetであるため, 自分の状況とは関係のない内容が含まれている. このtweetで考慮しなくてはならないのは, 自分の状況を特定のユーザーに伝えるものもあるという点であり, そのtweetには雪に関する発言が含まれている可能性がある.

・文中にRTを含むtweet：非公式retweetと呼ばれるもので, 特定のユーザーの投稿を引用し自分のコメント等を付けて投稿し直すものであり, 自分の状況とは関係のない内容が含まれている可能性が高い.

Twitterでは利用者の個人設定で, 自分の投稿を他人に非公開にすることもできる. そのため, 一般的に公開されていない投稿の取得は除外し, 誰でも閲覧できるtweetのみ収集した. また, 投稿されたtweetがRTから始まるtweetは公式retweetと呼ばれ, あるユーザーの投稿を, 自分のフォロワーに向けて再発信する機能によるもので, 本研究では, 利用者の独り言に近い投稿に注目している. そのためRTから始まるtweetは除外した.

2.2 KHcorderによる投稿内容の分析

収集したtweetに書かれた投稿内容は多岐に渡っており, 情報集約や関連性の検討が容易ではない. 投稿内容を自動的に解析するためフリーソフトウェアであるKHcoder（http://khc.sourceforge.net/）を用いた[2]. KHcoderを利用し, 収集したtweetの投稿内容を形態素解析後, 出現頻度の高い単語を用いて共起ネットワークを作成することで, 収集した投稿内容を定量的に可視化した. 形態素解析とは言語で意味を持つ最小単位の列に分割し, 語, 品詞, 活用形を判別するものである. 例えば, ”札幌は雪が降っている”という投稿があれば, 「札幌/は/雪/が/降って/いる」となる. この文章中の名詞や動詞を集めて集計し, その単語の頻出頻度による分析を行った. 共起ネットワークとは, 文書からその文書を特徴づける語の抽出を行い, 頻出パターンの似通った語同士の共起関係を図にしたもので, 出現数が多い単語ほど大きく, また共起の程度が強いほど太い線で描かれる. 共起ネットワークでは, 同じ文脈で用いられる語と語の関係を抽出できるので, 文章全体の文脈を捉えるときに有用である[3]. 語Xと語Yの共起の強さを測る指標には, ジャッカード係数J(X;Y)を用いる.

$$J = (|X \cap Y|)/(|X| + |Y|) \qquad (1)$$

ここで, |X|, |Y|, |X∩Y|, はそれぞれ語X, 語Y, 語Xかつ語Yの出現件数

ジャッカード係数が大きい単語同士は, 太線で結ばれる. ジャッカード係数Jが小さいと, 単語同士には線が引かれない.

2.3 tweetを収集した時期の降雪量

降雪状況による影響を明らかにするために, 平成23年12月1日から平成24年1月31日までの降雪量のデータを取得した. データの取得には, 札幌管区気象台気象統計情報の過去のデータより, 札幌における1時間ごとの降雪量データを使用した（表1）.

表1　札幌気象台観測降雪状況

札幌　2012年1月1日

時	気圧(hPa)		降水量	気温	風向・風速(m/s)		日照時間	全天日射量	雪(cm)	
	現地	海面	(mm)	(℃)	風速	風向	(h)	(MJ/m²)	降雪	積雪
1	1017.4	1020.8	—	-1.7	1.6	南南東			—	42
2	1017.3	1020.6	—	-1.5	2.3	南東			—	42
3	1017.2	1020.6	—	-2.1	1.9	南東			—	42
4	1017.1	1020.5	—	-3.5	1.1	東南東			—	42
5	1017.2	1020.6	—	-3.0	1.6	南東			—	42
6	1017.2	1020.6	—	-1.8	2.8	南			—	42
7	1016.7	1020.0	—	-1.6	2.7	南		0.00	—	42
8	1017.4	1020.8	—	-2.0	2.6	南	0.5	0.17	—	41
9	1017.5	1020.9	—	-1.8	2.4	南南東	0.1	0.33	—	41
10	1017.3	1020.6	—	0.5	1.5	南南東	0.4	0.70	—	41
11	1016.2	1019.5	—	2.2	1.0	南東	0.9	1.07	—	41
12	1015.2	1018.5	—	1.9	4.3	東南東	0.0	0.82	—	41
13	1014.5	1017.8	—	1.9	3.0	南東	0.0	0.67	—	41

3. 分析結果

3.1 収集 tweet 数と降雪量の関係

平成 23 年 12 月と平成 24 年 1 月に収集した tweet 数はほぼ同じとなった. 3 つの時間帯別の tweet 数には差違があった. 通勤通学時間帯である 8:00 から 9:00 の間に取得した tweet 数が最多となった. 各 tweet 収集時間帯から遡り, 最大 10 時間前から取得終了までの, 全 11 時間帯の降雪量の合計値とその時間帯の収集 tweet 数との関係性を調べた.

(1) 8:00-9:00 の結果

3 つの時間帯の中で, 最も tweet 数が多くなった. 「札幌は今日から雪！吹雪いてます!!」のような発言が多くみられた. 図 1(a)に降雪量と tweet 数の関係を示す. 降雪量が増えると tweet 数は増える傾向が見られた. 一方, 降雪量が 2cm で tweet 数が 319 件となり, 降雪量とは関係なく tweet 数が多くなった日（平成 24 年 1 月 24 日）もあった. 平成 24 年 1 月 24 日は, 関東で降雪が記録され, 「東京」という単語が多く含まれていた. 関東での降雪が取得ツイート数に影響を与えたことがわかる.

(2) 17:00-18:00 分析結果

tweet 数は少なく, 2 か月間の調査で, 1 日当たり 100 件以上の tweet 数は記録されなかった. 図 1(b)のように降雪があった日でも tweet 数は伸びなかった. 降雪量が 11cm を記録した日もあったが, tweet 数は 18 件と 32 件であった.

(3) 23:00-24:00 分析結果

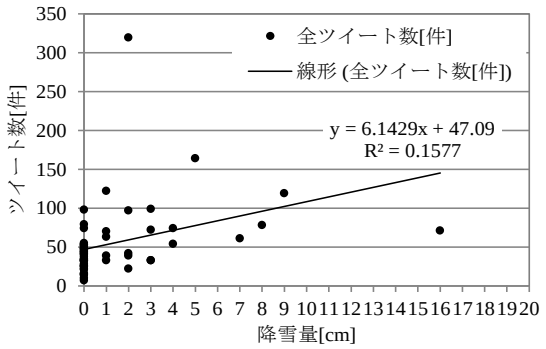
tweet 数は朝の通勤時間帯より少なく, 夕方の帰宅時間帯より多くなった. 図 1(c)に降雪量と tweet 数の関係を示す. 降雪が 0cm でも tweet 数が増えることが多くみられた. たとえば, 平成 24 年 1 月 23 日は最多 tweet 数となったが, 札幌での降雪量が 0cm であった. 通気時間帯と同様, 「東京」という単語が多く含まれており, 関東での降雪が影響していた. また, この時間帯には, 「札幌なまら雪降ってるやん！」「明日札幌雪かなぁ…」のような投稿があり, 明日の予報や過去のことが多く, 雪に関するリアルタイム性に欠く内容が見られた.

3.2 収集 tweet の投稿内容分析

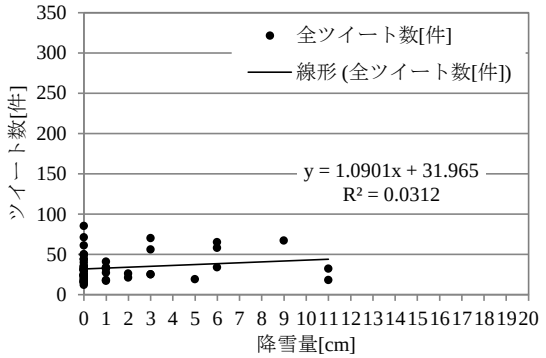
収集した全 tweet を, それぞれの時間帯で使われた形態素に分類し, 単語「名詞・地名・形容詞」のみを共起ネットワーク上に表示した. 「札幌」, 「雪」という語と関係性を共起ネットワークとして表し, 時間帯毎にどのような話題が多く投稿されているのかを明らかにする. 共起ネットワークの単語と単語が線で結ばれていれば, 同じ文章上で使われていた回数が多い. 例えば「札幌」, 「雪」, 「降る」という単語と線で結ばれていれば, その単語は同じタイミング, 同じ文章で使われる頻度が高いことになる. また, 単語の頻度が高くなると共起ネットワークの語を示す円が大きくなる.

(1) 8:00-9:00 の結果

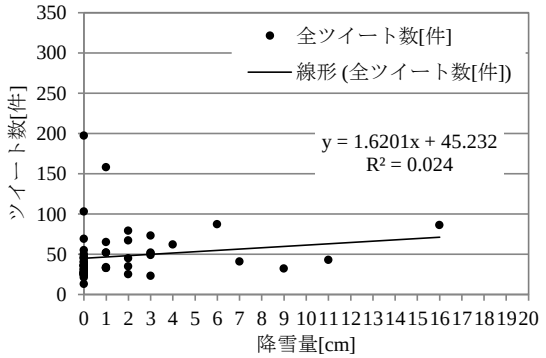
収集した tweet の投稿内容を形態素解析した(表 2(a)). 図 2(a)は, 選択した単語による共起ネットワークを示している. 「札幌」, 「雪」という単語を中心として, 「降る」, 「寒い」, 「おはよう」, 「今日」といった単語が現れた. 「今日」という単語が含まれていることからも, tweet 時の状況を表すものも多く投稿されていたことがわかる. また, 「札幌」, 「雪」と共起関係になく出現数が多い単語として, 「北海道」, 「ミク」といった単語が現れていた. これらの単語は, 「札幌」, 「雪」と線で結



(a) 8:00－9:00



(b)17:00－18:00



(c) 23:00－24:00

図 1 収集 tweet 数と収集開始 10 時間前から終了までの降雪量

ばれていない. この時間帯では当時の降雪の状況を表した「札幌」,「雪」を含む, その時目の前で起きている事を描写する tweet が多く投稿されていた. 一方, 利用者の興味のある時事的な事象に関する tweet も多く投稿されていた.

(2) 17:00-18:00 分析結果

この時間帯の頻出語を表 2(b)に示す. 図 2(b)は, 選択した単語による共起ネットワークを示している. この時間帯に取得された tweet ツイートには, 「雪多いな…さすが札幌」などの発言が多く見られた. 図 2(b)の共起ネットワークでは, 「札幌」,「雪」と関係の深い単語も「降る」や「ミク」,「電車」という単語が現れていることから, 時事的状況を表す投稿が多かったことがわかる. また, 他の単語を見ると, 札幌雪祭り, 時事的な内容, 話題となったニュースなどに関する tweet が多く投稿されていたことがわかる.

(3) 23:00-24:00 分析結果

この時間帯の tweet 頻出語を表 2(c)に示す. 図 2(b)は, それらの頻出語の関係を示す共起ネットワークとなっている. 「札幌」,「雪」と関連の強い語として, 「降る」,「明日」,「電車」,「ミク」,「電車」が多く, 降雪に関しては, 当日のことよりも翌日の話題と関連する傾向が見られた. この時間帯の tweet の特徴としては, そのとき目の前で起きていることを描写するというよりも, その時の利用者の心情や興味のある時事的事象に関する投稿が多くなった. 「札幌」,「雪」がたくさんの単語と関連があることから, 多様な話題が投稿されていたことがわかる.

表 2　単語（出現数）

(a)　8:00-9:00

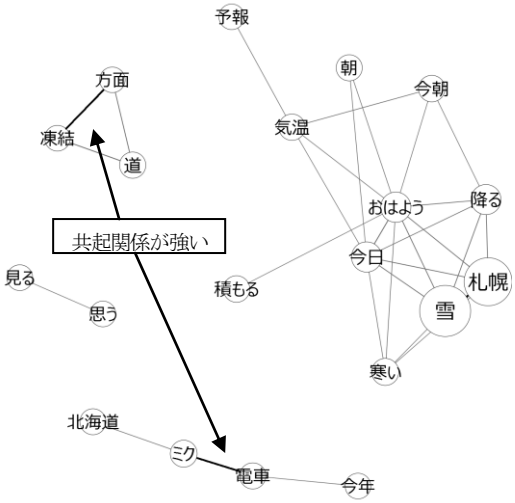
雪	(1421)	札幌	(1421)	今日	(333)	降る	(326)
寒い	(138)	ミク	(131)	今朝	(107)	積もる	(106)
気温	(93)	道	(92)	電車	(85)	朝	(81)
北海道	(74)	昨日	(67)	見る	(58)	思う	(58)
今年	(53)	凍結	(52)	方面	(47)	予報	(44)
おはよう	(43)						

(b)　17:00-18:00

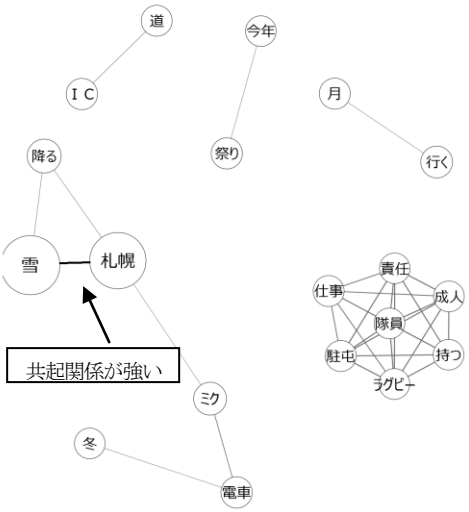
雪	(1130)	札幌	(1130)	降る	(186)	ミク	(142)
行く	(97)	電車	(79)	ＩＣ	(78)	祭り	(70)
月	(65)	駐屯	(64)	仕事	(62)	道	(61)
ラグビー	(57)	成人	(54)	隊員	(52)	冬	(50)
持つ	(49)	今年	(47)	責任	(46)		

(c)　23:00-24:00

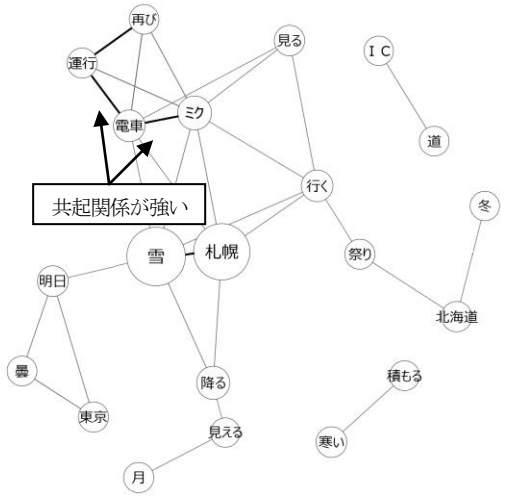
雪	(1406)	札幌	(1406)	ミク	(274)	降る	(233)
電車	(160)	行く	(159)	寒い	(111)	明日	(111)
見る	(98)	北海道	(91)	月	(87)	東京	(83)
祭り	(79)	運行	(76)	曇	(68)	積もる	(67)
道	(61)	再び	(55)	ＩＣ	(50)	見える	(43)
冬	(43)						



(a)　8:00-9:00



(b)　18:00-19:00



(c)　23:00-24:00

図 2　収集 tweet による共起ネットワーク

4. まとめ

本研究では, SNS の一つである Twitter 上に投稿される tweet の投稿内容に着目し,「札幌」,「雪」という単語を含む tweet から, 札幌における降雪が Twitter 上での発言にどう影響を与えているかを可視化することを目的としている. 収集した tweet の投稿内容を KHcorder から形態素解析し, 共起ネットワークを構築, 投稿内容に降雪がどのように影響するかについて分析を試みた.

tweet の投稿時間によって, 雪に係わる発言の内容は異なることがわかった.「札幌」,「雪」を含む tweet 数は設定時間帯の中で 8:00-9:00 が最も多く, 降雪があると tweet 数が増えた. 降雪の状況と投稿された tweet 数には降雪があると投稿数が増えるという正の相関関係があることがわかった. 8:00-9:00 の共起ネットワークの広がりが小さいことから, 他の時間帯と比べその時の雪に関連する話題が分散していないという特徴もあった. 朝の時間帯は特に雪による生活への不満が多く投稿されていたと言える.

一方, 夕方は降雪があってもそれほど tweet の発言に現れることがなかった. 8:00-9:00, 17:00-18:00 は共に通勤・通学, 帰宅時間であり, 利用者が外出していることが予想されるが, 取得された tweet 数には大きな差があり, 共起ネットワークによる頻出語も二つの時間帯では異なる様子を示した. 朝の方が雪による影響を重要視することが, 発言に現れるのかもしれない.

8:00-9:00, 17:00-18:00 は「今日」,「今」といった単語が含まれ, 当時の状況に関する内容のツイートが多く取得されたが, 23:00-24:00 のみ,「明日」といった翌日と関連する発言を含んでいた. ユーザーにとっての関心が降雪に直接反応せず, 雪に関わる発言が多様な内容を持つようになった. この時間帯に取得されたツイートの特徴として, 札幌で行われるイベントであるさっぽろ雪祭りに関係する語が, 共起ネットワーク上で「雪」と関連が強い「語」として現れた.

以上から, 雪に着目して tweet の投稿内容を分析した結果, 利用者の関心事が tweet の投稿内容に大きく影響していることがわかった. 降雪が通勤や通学に影響する朝の時間帯は, 降雪量が多いと投稿された tweet 数が増加する傾向が見られた. 一方で, 深夜時間帯は, ユーザーにとっての関心が降雪に直接反応せず, 雪に関わる発言が多様な内容を持つようになった. Twitter は一般的に即時性が強いと言われているが, 投稿時の状況を表す場合もあれば, その時の投稿者の関心, 心情を反映させる場合もあった.

本研究の課題として, (1)取得されたツイートは「札幌」というキーワードが含まれているが, 札幌以外の話題が含まれた点, (2)取得された tweet の話題が当時の状況とは関係のないものが含まれている点, が挙げられる. 地域の限定, 検索単語の拡大, 収集時間帯の拡大などを実施できる WEB サービスを作成することから, 冬期の雪が市民に与える影響を Twitter から鮮明に抽出できる可能性はあり, データを別途収集する努力が不要なメリットを考慮すると, 社会的な情報収集装置として tweet 利用を今後も検討する意義はある.

参考文献

1) 高橋哲朗・野田雄也, 実世界のセンサーとしての Twitter の可能性, 信学技報, ITEICE Technical Report NLC2010-38(2011-1)
2) 樋口耕一, KH Coder 2. x チュートリアル, 平成 24 年 7 月 22 日
3) Higuchi K, Komoda N, Tamura S, Ikkai Y: A Support Tool for Composing Social Survey Questionnaires by Automatically Summarizing Questionnaires Stored in Data Archives. WSEAS Transactions on Information Science & Applications 2007, 4:280-287.
4) 拡大を続ける Twitter の震災における活躍と今後の展 ADSTUDIES Vol.36 2011
5) 東日本大震災に見る大災害時のソーシャルメディアの役割 佐々木 放送研究と調査 JULY2011量
6) ソーシャルブックマークとしてのTwitterリスト機能の応用 榊等 人工知能学会 2010年度全国大会
7) 樋口耕一, テキスト型データの計量的分析, 理論と方法 (Sociological Theory and Methods), 2004, Vol. 19, No. 1: 101-115
8) 齋藤朗宏, 日本におけるテキストマイニングの応用, Workinng Paper Series No. 2011-2012

(2012. 8. 3 受付)