マイクロブログユーザの行動傾向の可視化
Visualization of micro-blog users' action tendencies

原口　義経
Yoshitsune Haraguchi
法政大学情報科学部コンピュータ科学科
E-mail: 07k0143@stu.hosei.ac.jp

Abstract

With the spread of micro-blog services such as Twitter, more people daily record their actions in micro-blogs. If we can obtain their action tendencies from the micro-blogs, we will be able to utilize such information for businesses in service, food, and other industries. For example, electronic newsletter services will be able to send appropriate messages in timely manners. For this purpose, researchers have been developing methods for visualizing micro-blogs. Previous work on micro-blog visualization typically used times of postings, frequencies of actions, and relations between users. This paper proposes a method for visualizing action tendencies of micro-blog users by using information on the locations and times of their postings. The method also adopts information on weather that is determined by locations and times of postings. It lays out texts expressing users' actions by using location information, and exploits the colors and sizes of the texts to represent weather and action frequencies. It also allows zooming of the visualization to solve the problem of overlapping texts. This paper shows the results of the experiments that used data obtained from Twitter for a month.

1. はじめに

近年，Twitter 等のマイクロブログを利用する人たちが増加しており，ユーザは行動を頻繁に記録するようになってきた．つまり，マイクロブログから読み取ることのできる情報量が増加しているとも言える．このことに着目してマイクロブログから行動傾向を探ることができればサービス業などで有効に利用することができる．しかし，マイクロブログの情報量は膨大で，漠然とデータを眺めただけでは行動傾向を読み取ることは難しい．

Shiroi らは Twitter のデータを俯瞰的に読み取るために ChronoView を開発した[1]．ChronoView では時刻を中心に考え，円状の時計をイメージした可視化を行う．またこの時，時刻が曖昧になること回避するために独自の「金平糖表現」を利用している．ChoroView で可視化するものは基本的に行動，行動頻度，時刻の関係である．

しかし，それら以外にも行動傾向を読み取る上で必要な情報があると考えられる．例えば天候及び位置である．日常的な行動や仕事に関する行動は天候に左右されるとは考えづらいが，プライベートでは影響していると考えられる．また位置は住宅街やオフィス街などといった様相の違う地域でも同じような行動傾向となるとは限らない．これら 2 つの情報も読み取れるようになれば，例えば企業はメールサービスなどでより適切な情報を配信できるようになるはずである．

本研究ではマイクロブログユーザの投稿した行動を位置情報，時刻情報に基づき可視化をする手法を提案する．位置情報や時刻情報がわかると天候に関する情報が導き出されるので，天候も利用する．行動，位置，時刻に関する情報は Twitter から 1 ヶ月間取得し，天候は天気予報サイトから取得した．これらの情報から必要なデータを抽出した上で，その可視化を行った．可視化では行動を表すテキストを位置情報に基づき配置し，テキストの色で天候を表現している．行動の頻度はテキストのサイズを利用することで表現している．さらに可視化したときに文字が重なって見づらくなる問題に対処するために可視化のズーミング機能を実装した．抽出したデータを用いて実験を行ったので，その結果を報告する．

2. 関連研究

本研究の先行研究の中でも特に参考にしたものが 2 つある．1 つは Misue による Anchored Map であり，もう 1 つは Shiroi らによる ChronoView である．

Anchored Map [2]はアンカーと呼ばれる特定のノードを多角形の頂点に配置し，フリーノードと呼ばれる他のノードをノード間の接続関係に基づいて配置する手法である．描画ではまずアンカーを円周上に等間隔で配置する．アンカーは最初にランダムに配置され，いくつかのルールに従ってアンカーを最適な並びに入れ替える．その後，アンカーを固定して力学的な手法により，アンカーとフリーノードの関係を表せる位置にフリーノードを配置する．フリーノードは特定のアンカーに近いほどそのアンカーと深い関連性があり，図の中心にくるほど関連性は均等である．

ChoroView [1]は時刻と行動を可視化するものである．これは時刻によりどの行動が起こりやすいかを俯瞰的に見ることができるようになっている．この手法の特徴は「金平糖表現」という独自の可視化手法を使用していることである．可視化ではまず円状に時刻を等間隔に配置し，行動との時刻の関係によりテキストを描画する．これは先述の Anchored Map を利用している．しかし，テキストの配置をする際に曖昧な状況が生まれる．ある時刻の集合 A とある時刻の集合 B が同じような頻度で行動を行った場合，それぞれの行動が同じ位置に来るようになる．それではどの時刻と関係があるのかがわからなくなってしまう．これを解決するための手法として金平糖表

Supervisor: Prof. Hiroshi Hosobe

現を使っている．金平糖表現は行動を表すテキストの中心部から放射線状に線を延ばしていく．延ばし方は，テキストの中心から真上の位置を時刻 0 時として 24 時間の円状の時計として見立てる．描画される線は時刻の方向へ伸びていく．この時，線分の長さはテキストの時刻と発生頻度の関係から決められる．ある時刻で発生頻度が多ければ長くなり，少なければ短くなる．この表現により大まかな発生時刻が読み取れるようになり，テキストの位置による曖昧さの回避につながる．

Twitter の可視化に関する他の先行研究に，竹内らによるツイートの拡散状況の可視化[3]や，太田によるリツイートの関係の可視化[4]の研究がある．

3. 提案手法

3.1 データの取得

本研究はマイクロブログの中でも特に有名な Twitter [5]を可視化の対象とした．Twitter ではユーザによる投稿はツイートと呼ばれる．公開されている Twitter の API を利用し，ツイートの内容，時刻，位置の情報を取得した．データは関東全域を覆うような範囲で取得し，ここからさらに必要な地域のデータを抽出した．範囲を広めにとったのは，位置情報付きのツイートが少ない問題を回避するためである．

本研究では取得したデータを 3 つの地域に分ける．具体的には，東京 23 区，東京都の 23 区以外，埼玉県の 3 つと，東京 23 区を 3 つに分けた 2 パターンである．

3.2 キーワードの抽出と辞書の作成

ユーザの行動を表すためにキーワードを導入し，キーワードからなる辞書を作成した．キーワードには，行動の内容を読み取れるもので，なおかつ，一般的であるものを選択した．具体的には，行動が読み取れると思われる動詞，及び名詞をキーワードと定義した．名詞をキーワードに含めたのは，行動を読み取ることのできるものがあるためである．例えば，「昼食」といった名詞によって，昼食を取っていることがわかる．また，「ラーメン」などのより具体的な名詞も利用できる．このように，抽象的な名詞に加えて，より具体的な名詞も行動として抽出するようにした．また，一部のキーワードについては，「らーめん」のようにひらがなで記述された場合にも判定できるようにした．

3.3 可視化手法

3.3.1 配置

本研究では先述の通り，天候や位置を可視化する．Anchored Map のように頂点を固定し，それ以外の要素を頂点との関係によって配置する．先行研究では時刻を頂点に配置していたが，天候や位置によっても行動傾向が変化すると考えられるので，少なくとも天候や位置に関する情報も直感的にわかりやすく可視化する必要がある．本研究では位置をアンカーとして，頂点に配置することにした．位置が直感的にわかりやすく可視化されると考えたからである．地域を 3 つに分けたため，基準となる位置を 3 角形の頂点に配置する．

キーワードは頂点との関係によって配置する．しかし単純な方法では， 3 つの地域のうちでツイートの総数が大きく違い，地域ごとの数で比較した場合，ほとんど全てのキーワードが 1 箇所に集まってしまうという問題が生じる．本研究の以下の方法で 1 種の正規化を行った．まず対象のデータからキーワードごとのツイート数を抽出する．各キーワードは地域ごとのツイート数を持っているので，そこから地域ごとのツイート総数を求める．地域ごとのツイート総数を各キーワードのツイート総数で除算する．ここで得られた 3 つの結果を w_1，w_2，w_3 とし，重みとする．また，各キーワードにおける地域ごとのツイート数を c_1，c_2，c_3 とする．ただし，添え字は地域を表現している．これらの値を使い，それぞれの地域における割合を求める．つまり，

$$p_n = (w_n c_n)/(w_1 c_1 + w_2 c_2 + w_3 c_3)$$

となる(n = 1，2，3)．この p_n をそれぞれ対応する頂点の座標に乗算することでキーワードを配置する座標を求める．これは ChronoView の手法と同様である． キーワードの x, y 座標はそれぞれ

$$x = p_1 x_1 + p_2 x_2 + p_3 x_3$$

$$y = p_1 y_1 + p_2 y_2 + p_3 y_3$$

である．ただし，x_n y_nはそれぞれアンカーn の x, y 座標を示す．こうすることでそれぞれの地域との関係を壊すことなく配置することができる．

3.3.2 色合い

次に重要な要素である天候は，配置されるキーワードの色を使い表現する．基本的には晴れの日が多ければ赤く表し，雨の日が多ければ青く表すようにする．複数の天候が関係しあう状況を扱うため，それらの比率を使い RGB の 3 原色を混ぜるようにした．このようにすれば，晴れの日にしかツイートされなかったキーワードは真っ赤になり，雨の日にしかツイートされなかったキーワードは真っ青になる．天候も地域と同様に偏りがあり，晴れの日が多く，雨の日が少ない．このため，3.3.1 節と同様に正規化を行った．まず，キーワードごとのツイート数を抽出する．次に天候ごとのツイート総数をキーワードごとのツイート総数で除算する．これを重みとして使い，重みから各天候の割合を算出する．その算出した値に 255 を乗算し，RGB の値を求めた．

3.3.3 文字サイズ

キーワードの表示のときに文字のサイズを利用することで行動頻度を表す．全体の総数と各キーワードの数の比率を求める．当初，配置や色合いの場合と同様に正規化をしたが，ある 1 つのキーワードの行動頻度が極端に大きな値になる場合には，例えば図 1 のような結果が得られた．

図 1. サイズ調整の失敗

そこで正規化のみを利用せずに文字のサイズに制限をつけることにした．ただし，文字のサイズを求めるときには平方根を使い，文字のサイズを面積として見て大きくなりすぎるのを防ぐようにした．また，逆に小さすぎる場合にも制限を設け，読み取ることができる大きさまで拡大した．

3.4 ズーミング

Anchored Map では配置の際に文字が重なることを回避しているが，本研究の手法では意図的にそれを行っていない．重なることを避けるために，本来の配置とは違う場所に配置することが必要になるためである．しかし，そのために文字が密集しやすくなってしまう問題がある．その対策として可視化結果のズーミングを実装した．ズーミングするとき，文字自体は大きくしたり，小さくしたりはせずに，座標変換によって文字と文字の隙間を変化させる．ズーミングはクリックによって実行され，拡大の場合にはクリックした位置が中心に来るように計算する．拡大の倍率は一律で 2 倍としている．

さらにズーミングによる拡大時にクリックした位置，つまり中心にしていた位置を保持しておくようにすることで，縮小機能も実装している．この機能を利用することで読み取りたい位置の情報をより正確に読み取ることができる．

3.5 金平糖表現

ChronoView で時刻の曖昧さを回避するために用いられた金平糖表現を，本研究では曖昧さの回避のためではなく，時刻の可視化表現として実装した．また，ChronoView のように 1 時間ごとに 1 本の直線を描画するのではなく，本研究では 3 時間ごとに 1 本とした．直線の本数が多くなると，直線でキーワードのテキストがつぶれることがあるからである．線分の長さは他の可視化情報と同様に正規化し，さらに定数を乗算して決定した．線分はある座標を中心にして，その位置から外側に伸びている．真上に伸びた線分を 0 時から 2 時とし，そこから 45 度刻みで 3 時間ごとの時刻を描画している．これをキーワードの近くに配置し時刻情報を可視化した．

3.6 詳細表示

可視化によって位置，時刻，発生頻度，天候，キーワードは一目でわかるようになるが，キーワードの詳細が必要になる場合もある．キーワードの詳細とはそのキーワードを含むツイートの内容，時刻，地域のことをいう．たとえば，「酒」というキーワードを見たときに，どのような種類の酒であるのか，酒を飲んでいるのか，買っているのかまではわからない．そこでそのような曖昧さを回避するために，キーワードを右クリックしたときにテキストエリアでキーワードの詳細を表示するようにしている．

4. 実装

4.1 元データの取得

本研究ではプログラムの実装に Java を使用した．Twitter からデータを取得する際には Twitter4j [6]という Twitter API [7]の Java 用ライブラリを使用した．データの取得期間は 2013 年 11 月 1 日から 11 月 30 日までの 30 日間である．

データ取得のためのプログラムは Twitter にアクセスした後にデータの取得を開始する．データを取得し，関東の範囲にあるかどうかをチェックし，関東外であるならば次のデータをチェックし，関東内であれば次の作業に進む．次に取得したデータの時刻を読み取り，その時刻のファイルがあるかどうかを確認する．ファイルがあれば加筆し，なければ新規のファイルを作成する．ファイルは 1 時間単位で作成した．ファイルの文字コードは UTF-8 で統一されている．ファイルにはツイートした人の ID，ツイート内容，緯度，経度，時刻を記録した．

天候に関するデータはプログラムを使わずに，天気予報サイト goo 天気[8]から取得したものを手でファイルに入力して作成した．天候は午前，午後で分けている．

4.2 可視化用データの作成

取得した元データをさらに加工することで，可視化に適したデータを作成した．必要となるデータはツイート内容，時刻，位置である．最初に Twitter から取得したデータを CSV 形式に変換するプログラムを作成して，CSV 形式のデータを得た．

次に CSV 形式になったファイルを地域ごとに分けるプログラムを実装した．これによってファイルは ID，ツイート，日付，時刻の 4 つの最低限の情報からなる．

さらに上記のファイルに天候の情報を付加して出力するプログラムを実装した．天候は行の終端に付け加えるようにした．

次にキーワードの辞書と比較して，辞書にあるキーワードがあったならば，時刻や天候情報を更新し，最終的に出力するプログラムを作成した．このプログラムは Java の HashMap を利用し，辞書のキーワードから時刻や天候に関する情報が得られるようにしている．時刻に関しては 1 時間ごとのツイート数，天候に関しては天候ごとのツイート数がそれぞれ格納されている．ファイルの出力は，キーワード，キーワードの総ツイート数，天候ごとのツイート数，時刻ごとのツイート数である．これが最終的に可視化に使うファイルである．

これらのプログラムは個別のプログラムとして実装している．理由の 1 つは基本的なものから徐々にデータの加工を進めたためである．もう 1 つの理由は工程ごとのファイルを残しておきたかったためである．工程ごとのファイルを残しておくことで違う組み合わせの情報が欲しくなった時に役に立つ．

また，作成したファイルは 2 種類あり，1 つは 23 区と 23 区外，埼玉県で地域を分けたものであり，もう 1 つは 23 区を 3 つに分けたものである．

4.3 可視化プログラム

可視化のプログラムにはいくつかのクラスを使った．具体的には，配置を決めるのに必要な情報のクラス，色を決めるのに必要な情報のクラス，文字のサイズを決めるのに必要な情報のクラス，可視化のためのファイルを定義するクラス，特定の値を計算するためのクラス，プログラムを実行し，描画するクラスを実装した．ここでも HashMap を利用して，上記と同様にキーワードからそれ以外の情報が得られるようにしている．このプログラムの主な流れは HashMap 内の要素を計算クラスのメソッドに引渡して得られた結果を，描画クラスの変数に格納

するというものである．計算クラスには 3.3 節で述べた計算式が実装されている．

5. 実験

作成したデータとプログラムを用いて可視化の実験を行った．可視化結果の例を図 2 に示す．

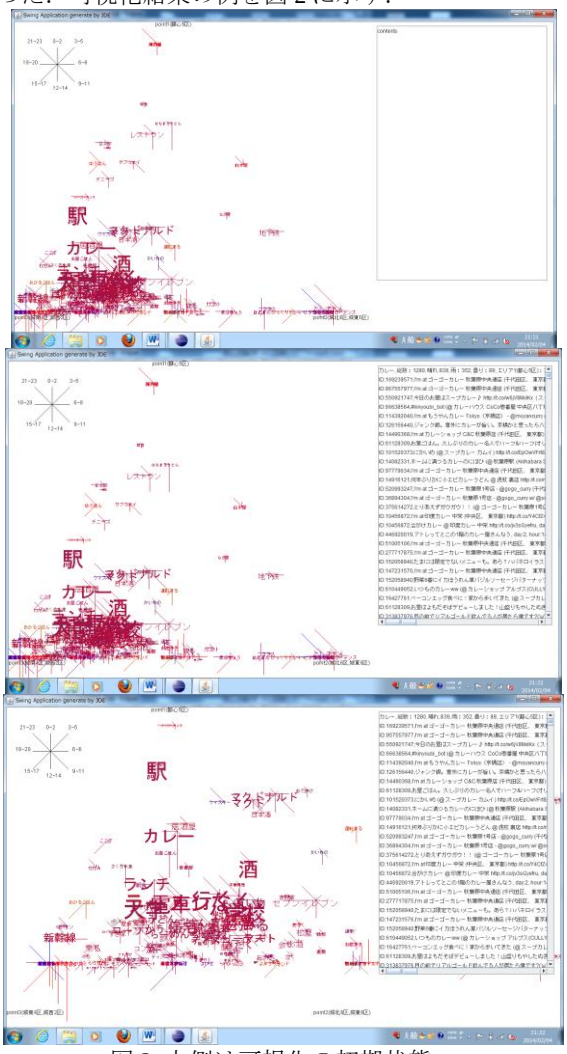

図 2. 上側は可視化の初期状態，
中央はカレーの詳細表示，
下側はカレーの周辺を拡大したもの

実行結果で 3 角形の領域の頂点に配置されているものがアンカーである．アンカーとキーワードは近ければ近いほど関係がある．また，キーワードが赤に近いほど晴れの日が多く，青に近いほど雨の日が多い．線分はキーワードの頻出時刻を表しており，各キーワードの時刻ごとの頻度が多いほど長い．また，画面右上に線分に対応した時刻を描画している．図 2 の下側はカレーの周辺をクリックしたことによって拡大された結果を示す．

可視化結果から特定の地域にツイートが偏っていることが読み取れる．つまりそれだけ人口密度が高いとも言える．また，例えばシーパラダイスという文字がほぼ赤だけで書かれている．このことから外のレジャー施設は晴れの日に行く可能性が高いことがわかる．また，駅，ランチなどといった日常的に利用するものは多くの人がツイートすることがわかる．特にラーメンに人気があることがわかる．また駅というツイートが多いことから，電車の中で携帯電話やスマートフォンを利用する人が多いであろうこともわかる．また，時刻情報から飲食店の利用は朝が少なく，昼夜は多いことがわかる．昼休みの時間や帰宅途中もしくは帰宅後のツイートが多いことも読み取れる．

6. 議論

1 箇所にツイートが集中する傾向があったために，正規化を使う手法を導入したものの，それだけでは限界があった．また，データの取得期間が 1 ヶ月であったため，天候による変化がうまく現れない問題も生じた．

さらにキーワードの辞書の内容も十分でなかった可能性がある．現状ではキーワードの抽出が自動化されていないため，随時追加することで対処する必要がある．

ツイートから抽出したキーワードが行動でないことがある．これに対処するためには，辞書とキーワードのマッチングだけで抽出を行うのではなく，別の抽出方法も組み込む必要性がある．

7. まとめ

本研究では，地域や天候によってマイクロブログのユーザの行動がどのように変わるかを読み取れるようにすることを目的として，可視化の手法を提案して実装し，Twitter から取得したデータを用いて実験を行った．データの取得期間，辞書の強化，可視化するときの地域の選定は今後の解決すべき課題である．

文　献

[1] Satoko Shiroi, Kazuo Misue, and Jiro Tanaka, ChronoView: Visualization Technique for Many Temporal Data, Proceedings of 16th International Conference Information Visualization (IV2012), pp.112-117, 2012.

[2] Kazuo Misue, Graph Drawing Technique to Support Network Mining. IEICE Trans. Inf. & Syst., Vol. E91-D, No. 11, pp. 2599-2606, 2008.

[3] 竹内俊貴, 谷川智洋, 西村邦裕, 廣瀬通孝, Twitter における情報の広がりの可視化と分類, 第 17 回サイバーワールド(CW)研究会, 2010.

[4] 太田侑介, Twitter におけるリツイート経路の可視化とユーザ発見支援(卒業論文), 電気通信大学情報通信工学科, 2011.

[5] Twitter, https://twitter.com/

[6] Twitter4j, http://twitter4j.org/ja/index.html

[7] Twitter API, https://dev.twitter.com/

[8] goo 天気, http://weather.goo.ne.jp/