人工知能学会 インタラクティブ
情報アクセスと可視化マイニング研究会(第9回)
SIG-AM-09-04

タグマッピングによるTwitter特性と話題の関係解析

Relationship Analysis between Twitter's Parameter and Topic using Tag Mapping

清政 貴文[1*]　六井 淳[1]
Takahumi Seimasa[1], Jun Rokui[1]

[1] 島根大学総合理工学研究科
[1] Graduate School of Synthesis Sience and Enginearing, Shimane University

Abstract: We propose the new analysis method of Twitter, which uses 3 types parameters in single Tweet. We show relations of Topics-User-Trend visually based on Hashtag. System contains 3 types Self-Organizing Map used for visualize. We operate maps and analyze interactively.

1　はじめに

近年，世間の思想や流行を解析するための情報源としてTwitter[1]やFacebook[2]などのソーシャルメディアサービスが注目されている．本研究ではTwitter上の話題について興味をもつユーザ層や盛り上がりへの寄与度を解析し，影響力の強い話題の特徴を抽出する手法を提案する．解析にはハッシュタグが付けられているツイートを利用し，1個のツイートからツイート主のフォロワー数などユーザ属性，流行を示す被リツイート数，ツイートに含まれる単語群の3種の情報を抽出する．ハッシュタグを含むツイートはタグが示す話題に関するものであると考え，3種の情報から生成したマップをタグで繋げて各話題を視覚的に解析する．

Twitterユーザの影響力を解析する手法は数多く研究されている．フォロー関係をリンク，発言内容の類似度を重みとしてPageRankを応用する手法 [1] では，より多くのユーザに影響を与えそうな人物の特定に成功している．また，特定の話題において影響力をもつユーザを特定する手法 [2] では，キーワード検索で集めたツイートのリツイートやお気に入り情報を解析している．

ユーザの影響力ではなくタイプを分類する研究も盛んに行われている．ユーザがＢｏｔなのか人間なのかを判定する研究 [3] では，ユーザのフォロー関係やツイートの内容・時刻のパターンを利用して分類している．その他に，ユーザの主なツイートがどのような範囲の集団に向けたものであるかを解析して分類する手法 [4][5] なども存在する．

*連絡先：島根大学総合理工学研究科
〒 690-8504　島根県松江市西川津町 1060
E-mail: s149505@matsu.shimane-u.ac.jp

[1]https://twitter.com/
[2]https://www.facebook.com/

SOM(Self-Organizing Map, 自己組織化マップ)[6] を利用してTwitter上の豪雪トピックを解析する先行研究 [7] では，1種のタグツイートについて単語頻度ベクトルのSOMを構成し，内在する話題同士の類似関係を視覚化している．

SOMをインタフェースに用いてユーザ適応的なWeb検索を行う研究 [8] では，SOMとして表示した検索結果を操作して主観を反映した情報統合を行っている．SOMを用いた対話的推薦システムの研究 [9] では，ユーザのコンテンツ利用履歴から構成したSOMの操作により推薦を行っている．

本稿の2章では収集したデータ対する予備解析の結果を解説し，3章で構築したシステムの機能と仕組みについて解説，検証と考察を行う．4章ではまとめと今後の展望を述べる．

2　対象データ

解析対象としてハッシュタグがつけられたツイートを収集した．対象のタグは2014/11/8〜11/15の予備期間にTwitterStreamingＡＰＩのパブリックストリームから収集した出現頻度上位10%(112233種) から，日本語を含むタグ5839種を抜き出し，そこからランダムに100種類を選択した．その後，Twitter REST APIのsearch/tweetsにより2014/12/1〜12/25の間に出現した各タグツイートを収集した．なお，データの取得にはtwitter4j[10] を利用した．本調査期間で一日平均10回以上利用された89種のタグを解析対象とする．89種で合計1,834,923個のタグツイートを収集し，各タグの平均では20,617個，最も多く集まったタグで168,846個，最少のタグで509個のツイートが集まった．

各タグツイートからは表1に示す情報を取得して解

析に利用している．ツイート主の情報は，タグツイートをしたユーザのプロファイルから取得する情報である．オリジナルツイート主の情報は，タグツイートがリツイートだった場合にツイート主の情報とは別に元のツイートをしたユーザから取得する情報である．リツイート以外の場合，オリジナル情報の各値は全て 0 として扱っている．被リツイート数は各タグが付与されたリツイートについて，収集時点での被リツイート数である．被リツイート数に関して，クロールは期間中 5 分おきに行ったため，各リツイートの生成時点ではなく，クローラが最初に各リツイートを収集した時点での大元の被リツイート数を参照している．なお，被お気に入り数は評価タイミングが難しいことに加え，タイミングを合わせる場合に調査期間初期のツイートの修正に手間が掛かることから除外している．単語の抽出には形態素解析器 kuromoji[11] を利用し，2 文字以上の数でない名詞，または動詞であると分類されたものを単語として扱っている．また，本文から全てのハッシュタグ部分を除いたテキストの単語頻度を求めている．解析対象として選ばれた 89 種以外のハッシュタグも含有タグとして収集しており，単語解析時には本文から除いている．

ツイート主の情報	フレンド数
	フォロワー数
	お気に入り数
オリジナルツイート主の情報	フレンド数
	フォロワー数
	お気に入り数
流行情報	被リツイート数
単語情報	本文単語ヒストグラム
タグ情報	含有ハッシュタグリスト

表 1: タグツイートからの取得データ

解析対象中でツイート数の多い上位 3 種のタグの統計を表 2 に示す．ツイート数は期間中のタグツイート数，利用者数はタグを利用したユーザ数である．ツイート数に比べユーザ数が少ないほど，個々のユーザの呟きが多いことを示す．リツイート数はタグツイートの内リツイートの数であり，R ユーザ数はリツイート元のユーザ数を表す．リツイート数に比べて R ユーザ数が少ないほど，少数のユーザに注目が集まっていることを示す．フレンド平均はタグ利用ユーザのフレンド数の平均を表す．同様にフォロワー平均はフォロワー数の平均を表し，お気に入り平均は利用者のお気に入り数の平均である．ツイート主の統計は，各タグツイートのユーザ数として数えられているユーザを対象に求めている．頭に R がついているパラメータは，それぞれリツイート元のユーザについての統計である．リツイート元の統計は，各タグツイートの R ユーザ数を構成するユーザを対象にしている．各パラメータの偏差は，それぞれ利用者の多様性，リツイート元の多様性を示している．被リツイート平均は全てのタグツイートについて被リツイート数の平均を求めたものである．単語数平均，分散は各タグツイートに含まれる単語数の平均と分散を表す．総単語種は期間中の各タグツイート全体で出現した単語の種類数を表す．

表 2 の結果からは，次のようなことが推測できる．ツイート数とユーザ数の比率から，「#2ch まとめ」は広報用のアカウントに多く利用されている．フレンド平均，フォロワー平均から，「#進撃の巨人」は他 2 種よりも閉じたコミュニティに属すユーザに利用されている．お気に入り平均と R お気に入り平均から，リツイートされる側よりも，する側の方がツイートの収集意欲が高い．ツイート数と総単語数の比から，「#進撃の巨人」は他 2 種よりもツイートの内容が偏っている．

ハッシュタグ名	89 種全体	# アニメ	#2ch まとめ	#進撃の巨人
ツイート数	1,834,923	168,864	142,890	122,592
ユーザ数	180,268	12,504	555	24,475
リツイート数	505,930	23,329	937	66,024
R ユーザ数	12,808	1,235	43	1,357
フレンド平均	599	1,011	1,465	790
フォロワー平均	696	1,263	1,708	886
お気に入り平均	2,898	4,248	4,820	2,441
フレンド偏差	2,550	4,217	8,064	2,894
フォロワー偏差	4,495	7,974	8,795	5,825
お気に入り偏差	12,775	18,814	20,164	9,323
R フレンド平均	1,019	2,020	1,759	1,939
R フォロワー平均	2,355	3,195	1,695	4,568
R お気に入り平均	2,296	2,212	1,290	1,789
R フレンド偏差	6,227	9,333	3,675	8,636
R フォロワー偏差	48,594	19,737	3,743	26,420
R お気に入り偏差	14,272	12,319	7,955	8,605
被リツイート平均	33	4.1	0.015	68
被リツイート偏差	181	28	0.43	274
単語数平均	7.2	7.2	5.0	7.0
単語数偏差	5.2	4.8	4.0	5.0
総単語種	47,740	21,630	21,568	8,377

表 2: 抜粋したタグツイートの統計

また，89 種のタグについて，標本が 89 個あるとして統計パラメータ同士の相関係数を求めた結果，表 3 に示す傾向が得られた．表 3 から，フレンド数が多いユーザに利用されるタグはフォロワー数が多いユーザにも利用されている．これは，先行研究におけるフレンド数とフォロワー数の正の相関 [1][3] がタグ毎の統計にまとめられても有効であることを示している．また，フレンド数，フォロワー数の高いユーザに利用される

タグではツイート数に対するユーザ数，リツイート数が少なくなる傾向が現れている．この傾向から，フレンド数，フォロワー数の高いユーザに使われるタグは個人の使用頻度が高く，新規のタグツイートが生まれやすいと推測できる．次に，個々のユーザが均等に少なくツイートしている話題について，被リツイート数が高くなる傾向が現れている．これは，一過性に近い話題ほどリツイートされやすいと考えられる．お気に入り数が多いユーザに利用されるタグでは，出現する単語の種類が多くなる傾向がある．このことから，収集意欲の高いユーザはタグが示す話題の多様性に貢献している，または，多様な話題を好むと考えられる．

パラメータ組	係数
(フレンド平均, フォロワー平均)	0.886
(フレンド平均, ユーザ数/ツイート数)	-0.358
(フレンド平均, リツイート数/ツイート数)	-0.376
(フォロワー平均, ユーザ数/ツイート数)	-0.349
(フォロワー平均, リツイート数/ツイート数)	-0.426
(被リツイート平均, ユーザ数/ツイート数)	0.417
(お気に入り平均, 総単語種/総単語数)	0.479

表 3: 統計パラメータ間の相関

本章の解析結果から，タグツイートの統計をもとに各話題の傾向を読み取ることが可能であると示唆された．

3 提案システム

3 章では収集したデータをタグ毎に分類して傾向を見たが，フレンド数別や単語別など，様々な観点から解析を行うことでより詳細な傾向を観察できると考えられる．そこで，SOM[6] を用いてデータの特徴を圧縮・抽出し，マップの操作により対話的に解析を行うシステムを提案する．システムの構成を図 1 に示す．提案システムは学習ユニットと解析ユニットからなり，下記の 3 ステップで解析を行う．

1. 対象データを学習する
2. 各 SOM のセルにツイートをマッピングする
3. マップを操作して解析を行う

3.1 SOM の仕様

SOM とは与えられた高次元入力の特徴を低次元のマップに写像する手法であり，提案システムでは各ツイートの特徴を 3 種類の 2 次元マップに写像する．各 SOM への入力を表 4 に示す．ユーザ SOM，リツイート SOM への入力は全てツイートから取得した値をそのまま与える．単語 SOM について, ツイート本文に各

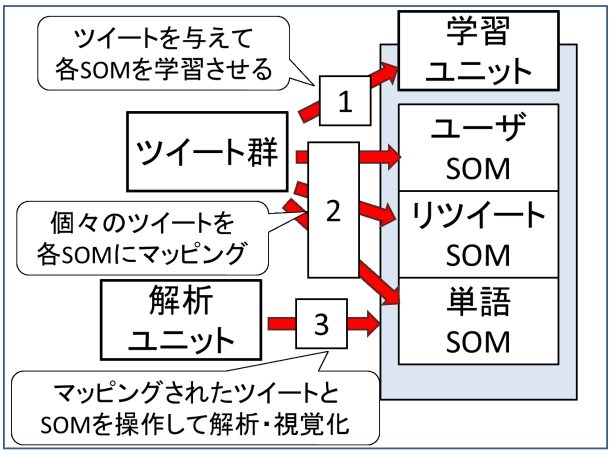

図 1: システムの概要図

タグの代表語が含まれる個数を与える．代表語は各タグツイートの頻出上位 5%の単語について，他のタグの上位 5%に含まれていないものを抽出する．タグ 89 種の内，6 種について代表語が存在しない結果となったが，単語 SOM の入力は 89 次元で行っている．各 SOM のセルはそれぞれに対する入力を表現できる重みをもつ．ユーザ SOM，リツイート SOM のセルの初期化には，表 2 に示す 89 種の全タグツイートについて求めた統計データを基に式 (1) で発生させた正規乱数を利用する．負の値で初期化される場合もそのまま変更していない．単語 SOM の初期化では，各セルに対して 0〜4 の一様な整数乱数を発生させ，その後 89 個の次元から 0〜4 個をランダムに選択して 1 とし，他の次元を 0 とする．学習ではバッチ学習でマップを均した後，逐次学習を行う．勝者セルの探索に用いる距離計算は式 (2) により行う．各パラメータ毎に値の大きさが異なるため，式 (2) ではパラメータ毎の分散を用いて各パラメータの距離の正規化を狙っている．重みの更新について，勝者セルには学習データのパラメータをコピーし，近傍セルの重みはパラメータの差分と勝者ユニットからの距離に応じて近づける．

ユーザ SOM	リツイート SOM	単語 SOM
フレンド数	R フレンド数	タグ 1 の代表語数
フォロワー数	R フォロワー数	タグ 2 の代表語数
お気に入り数	R お気に入り数	…
	被リツイート数	タグ 89 の代表語数

表 4: 各 SOM への入力

$$\begin{cases} p = p\text{の平均} + \text{標準正規乱数} \times\ p\text{の標準偏差} \\ p : \text{ユーザ } SOM,\ \text{リツイート } SOM \text{ の各入力} \end{cases} \tag{1}$$

$$\begin{cases} D_{ij} = \sum_{p \in\ P} d_{ijp} \\ d_{ijp} = (Tjp - Cip)^2 / VARp \\ P : \text{セルのパラメータ集合} \\ Tjp : \text{学習データ } j \text{ のパラメータ } p \\ Cip : \text{セル } i \text{ のパラメータ } p \\ VARp : \text{パラメータ } p \text{ の分散} \end{cases} \tag{2}$$

3.2 システムの機能

システムのメイン画面を図 2 に示す．メイン画面は上部のマップパネルと下部の情報パネルからなる．上部で選択したセル，セルにマッピングされているツイートの情報を下部に表示し，セルに集約された傾向を確認できる．下部は 3 領域に分かれており，それぞれ各 SOM で選択された 1 個のセルに対応した情報を表示する．各領域では対応パラメータに加え，マッピングされたハッシュタグの頻度情報を表示している．また，ある SOM で指定した 1 個以上のセルにマッピングされたツイートが他の SOM でどのように分布しているかを確認することができる．

より詳細な傾向を見る場合，1 個以上指定したセルのマッピングツイートを図 3 に示すサブ画面に表示する．サブ画面では SOM の操作で得た集合を平面に配置し，各領域ごとの傾向を掘り下げて見ることができる．平面に対応するパラメータを変更することで視点を変えた解析が可能である．

図 4 のタグ画面から 3 種までのタグを選択することで，メイン画面・サブ画面におけるタグの分布状況を比較することができる．比較の際は各セル・領域の内，マッピングの最大個数との比を RGB の強さに対応させ，マップの色として重ねて表示している．人気タグと似た傾向をもつ有望なタグの探索，特定のタグに注目した解析が可能である．

3.3 検証

タグ 89 種について集めた 1,834,923 個の全対象データを学習したシステムに対して検証を行った．3 種の SOM 全てに関してマップサイズは 10 × 10，即時的な分類性能を測るため，学習回数はバッチ学習 1 回，逐次学習 1 回と少なく設定している．

学習ユニットの結果において，ユーザ SOM のお気に入り数，リツイート SOM の R お気に入り数，被リツイート数の 3 パラメータは学習後の偏りが強かった．

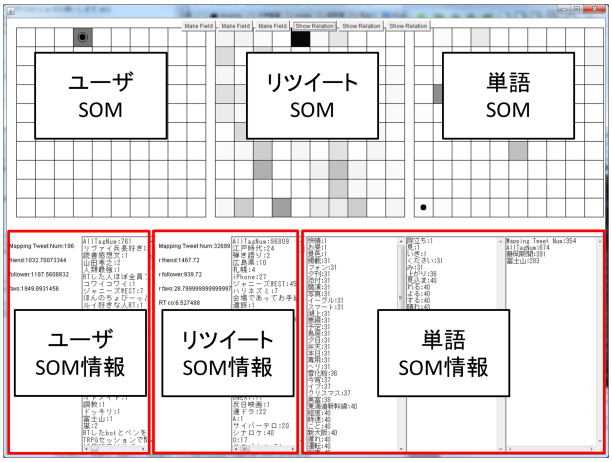

図 2: メイン画面

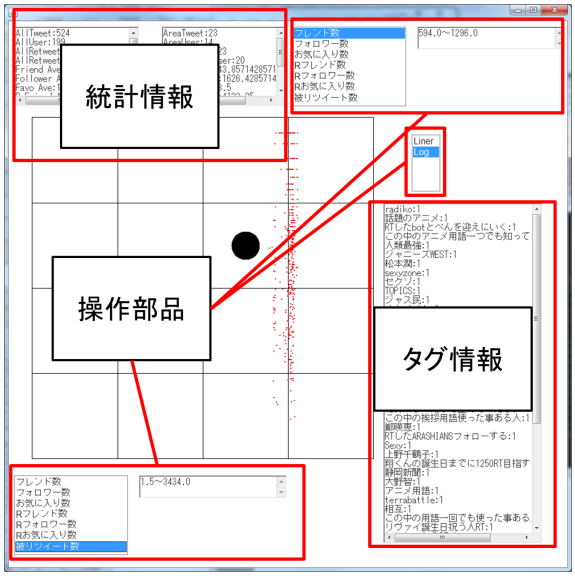

図 3: サブ画面

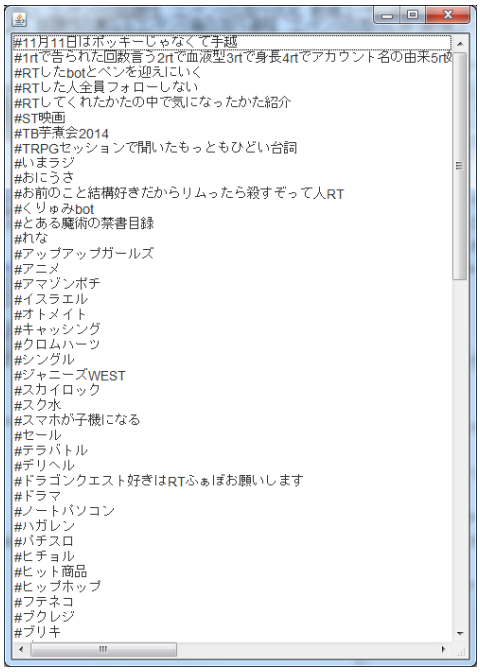

図 4: タグ操作画面

各試行で前述の 3 パラメータが 0 に近いセルが大多数を占め，SOM 上で形成される集団の多様性を狭めていた．そのため，SOM 上でお気に入り数，被リツイート数に着目した解析を行うことができなかった．多様性欠如の原因として，お気に入り数の分散が他に比べて大きく距離計算で軽視されやすいこと，元が同じリツイートを各リツイートで別々に学習させたことが考えられる．話題の盛り上がり要因を解析するためには被リツイート数の多様性が求められるため，リツイート SOM は特に改善が必要である．ユーザ SOM ではマッピング数の偏りや無駄なセル数が最も小さく，良好な学習がなされている．単語 SOM ではマッピング数 0 のセルが目立ち，入力構成や代表語抽出法の改善が必要である．

システムを利用した解析について，タグ操作画面から表 2 に記載した 3 種のタグを選択して分布を比較した画面を図 5，6 に示す．図 5 はユーザ SOM における 3 タグの分布を示しており，赤が「#進撃の巨人」，緑が「#2ch まとめ」，青が「#アニメ」に対応している．各セルの色の強さは全セル中の最大マッピング数との比なので，濃い色のセルが多いほど均等に分布しており，少ないほど集中している．

図 5 では一番右列の中心辺りの赤いセル 1 に「#進撃の巨人」が集中しており，その直上の緑のセル 2 に「#2ch まとめ」，中央よりやや左上の青いセル 3 には「#アニメ」のタグツイートが多くマッピングされている．また，赤と青は緑よりも広くマップ上に点在している．

図 6 は図 5 の画面のユーザ SOM から各タグが最も多くマッピングされている赤，緑，青の 3 セルについて，マッピングツイートを細かく表示したものである．

図 6 では赤点でタグツイートの分布を示しており，右ほどフレンド数が多く，下ほどフォロワー数の多いユーザが呟いたものである．マップ中の線は対数で引いており，一番左上のエリアはフレンド数・フォロワー数が 10 未満，その右隣はフレンド数が 10〜99 でフォロワー数が 10 未満のユーザによるツイートがマッピングされている．各エリアの色の対応は図 5 と同じである．図 6 中のエリア 1〜4 のタグ分布を表 5 に示す.「#進撃の巨人」は主にエリア 1,2，次いでエリア 4 に分布しており,「#2ch まとめ」は大部分がエリア 3,「#アニメ」はエリア 4，次いでエリア 1,2 に存在する．表 2，5 より，各タグツイートの約 1/4〜1/3 が 3 個のセルに集約されており，主利用者層の抽出に成功していると考えられる．

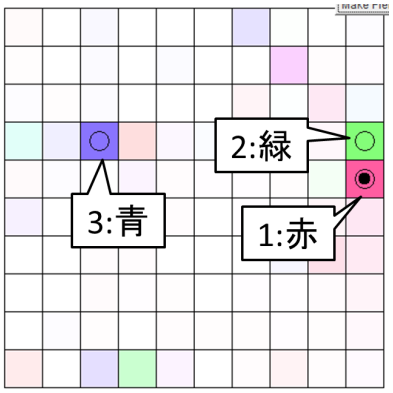

図 5: タグ 3 種のユーザ SOM 上の分布比較

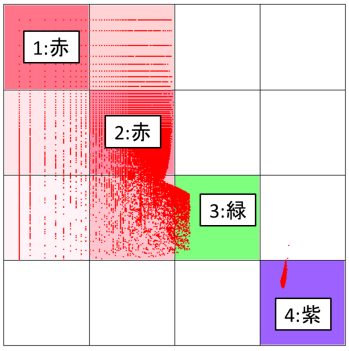

図 6: 注目セル 3 個中のタグ 3 種の分布比較

主利用者層の比較として見ると,「#進撃の巨人」は小規模コミュニティのユーザに多用され,「#2ch まとめ」はフォロワー数の方が多く 100〜1000 人にフォローされているユーザに利用されている.「#アニメ」は両方が 1000 人以上のユーザに多く利用されている.「#進撃の巨人」,「#アニメ」の分布は 2 章のタグ間比較に符合する.「#2ch まとめ」は表 2 に照らし合わせると，主利用者とは別層の人気アカウントにも利用されていると考えられる．セル 3 個分のツイートでは各タグの全

	ツイート数	ユーザ数	#アニメ	#2chまとめ	#進撃の巨人
3 セル合計	442,675	40,197	54,406	57,455	33,613
エリア 1	63,229	5,090	5,256	2	10,198
エリア 2	100,867	26,150	6,097	96	8.265
エリア 3	61,165	1,243	1,196	56,447	375
エリア 4	91,555	1,140	35,852	180	3,888

表 5: 図 6 のタグ分布

体傾向を見れていないが，各利用者層の視覚的解析は成功している．SOM の選択セル数を増やして各タグの解析範囲を広げることで表 2 を踏まえた結果に近付くと考えられる．

3.4 考察

検証結果より，提案システムにより対象データの傾向を発見できる可能性が示唆された．そのため，ユーザと話題の関係を解析する，という観点において提案システムは有効だと言える．

盛り上がりの要因解析についてはリツイート SOM の学習が上手く行えておらず，現状の提案システムでは不十分である．学習手法や入力構成の改善によりリツイート SOM の分類性能が向上すれば有用になると考えられる．

また，本研究では個々のツイートを主体としているため，ユーザ側のフォロー関係や前後の発言内容は考慮していない．そのため，各話題に参加している Bot や広告用アカウントなどを区別できておらず，話題の性質をより詳細に解析するためには個々のユーザを詳しく見る必要がある．また，各ユーザの Twitter 利用期間を考慮せずに各プロファイルを参照しており，各パラメータの高低が充分にユーザの利用傾向を表しているとは限らないため，改善の余地が大きい．

対象データ内において各タグの利用者特性に差異が認められたため，ユーザの影響力やタイプを解析する研究 [1][2][3][4][5] などを包含した，当該話題に対して各系統のユーザがどの程度参加しているか，という指標が Twitter 解析において有効であることが示唆された．

4 まとめ

本稿ではツイート情報を 3 種の SOM に学習させ，視点を変えたマップ表示により対話的に解析する手法を提案した．対象データの静的解析と提案システムを用いた解析の比較から，提案システムは一定の解析性能を有することが示された．今後は考察で述べた改善に取り組むと共に，細かい期間におけるタグツイートの統計情報とタグツイート数の増減の関係を解析して流行を予測する手法を研究する予定である．

参考文献

[1] Weng, Jianshu. , Lim, Ee-Peng. , et al.: Twitterrank: finding topic-sensitive influential twitterers, *Proceedings of the third ACM international conference on Web search and data mining. ACM*, (2010)

[2] Noro, Tomoya. , Ru, Fei. , et al.: Twitter user rank using keyword search, *Information Modelling and Knowledge Bases XXIV. Frontiers in Artificial Intelligence and Applications*, Vol.251, pp.31-48 (2013)

[3] Chu, Zi. Gianvecchio, Steven. , et al.: Who is tweeting on Twitter: human, bot, or cyborg?, *Proceedings of the 26th annual computer security applications conference. ACM*, (2010)

[4] 竹村 光, 田島 敬史: 情報発信の対象範囲に基づく Twitter ユーザの分類, *DEIM Forum B1-6*, (2013)

[5] Yan, Liang. , Ma, Qiang. , et al.: Classifying Twitter Users for Spatio-temporal Entity Retrieval, 電子情報通信学会技術研究報告. *DE,* データ工学, Vol.112, No.346, pp.93-98 (2012)

[6] Kohonen, Teuvo.: The self-organizing map, *Neurocomputing*, Vol.21, No.1, pp.1-6 (1998)

[7] 澤田 義人, 他: 2011 年山陰豪雪に関連する Twitter メッセージ解析法の開発, 生産研究, Vol.64, No.4, pp.467-473 (2012)

[8] 佐野 綾一, 波多野 賢治, 田中 克己: 自己組織化のマップを用いた Web 文書の対話的分類とその視覚化, 情報処理学会研究報告. データベース・システム研究会報告, Vol.98, No.57, pp.33-40 (1998)

[9] 藤森 洋昌, 土方 嘉徳, 西田 正吾: 協調フィルタリングにおける近傍グループの可視化, 情報処理学会研究報告. 情報学基礎研究会報告, pp.59-66 (2004)

[10] twitter4j-3.0.4:http://twitter4j.org/ja/index.html

[11] kuromoji-0.7.7:
http://www.atilika.com/ja/products/kuromoji.html