RO-006

Twitter におけるコミュニティ分析に関する検討
Study on Analysis of Community in Twitter.

佐藤 真弥子† 吉開 範章‡ 栗野 俊一‡
Mayako Sato Noriaki Yoshikai Shun-ichi Kurino

1. はじめに

近年、様々なソーシャルネットワーキングサービス(SNS)の登場を受け、企業や組織で SNS が導入されている。しかし、実際に導入後の利用・運営が成功しているケースは、導入企業・組織の中の 10%程度に留まっているという報告がある[1]。この原因として、社員が SNS 上での活動を通してコミュニティを形成することや、コミュニティに対して知識を提供することに対する動機づけが不十分であるといった、”目的の欠如”が考えられる。つまり、経営者側が、社員の SNS 活用状況や使用頻度といった SNS での組織活動を把握し、それに応じて利用促進をねらいとする SNS 活用の魅力や価値の提示をしていかなければ、企業での安定した SNS の利用・運営は困難であると考えられる。

そこで、我々は、SNS における組織活動の活性化を目標とし、その第一段階として、Twitter を対象にしたコミュニティの抽出と、その特徴分析について検討している。

今回、イベント性および類似性抽出の安易さ等を考慮して、Twitter における「衆議院議員選挙情報」を対象に、ケーススタディを行ったので報告する。

2. 研究の位置づけ

ソーシャルネットワーク分析（以下、SN 分析）は、社会構造をノード(点)とエッジ(辺)の構造として捉える”グラフ理論”を用いて分析する手法である。SN 分析では、個人やグループなどを対象ノードとし、その対象ノード間における関係性に着目して分析することで、ネットワーク上で起こる現象を捉えるため、ノード固有の属性に分析結果が左右されないといった特徴がある。

先行研究として、ネットワーク上のコミュニティ構造を明らかにし、ネットワーク上の活動の活性化や効率化を目的に、SN 分析を用いた研究がなされている[2]。我々もこれまでに、メールデータに基づく、実際の研究機関の活動評価に関する研究[3]を行って来た。

近年、mixi や Facebook といった様々な SNS が登場し、その新しさや手軽さ・話題性などから急速に普及したことで、多くの人々が SNS を通して、ネットワーク上でコミュニケーションを取るようになった。これを受け、特に Twitter を対象とした、SNS 上の活動やコミュニティに関する研究が盛んに行われている。その中の 1 つとして、有向リンク構造のネットワークに対し、隣接ノードの持つ Hub 値や Authority 値の和から、ノードの評価を行う HITS(Hyperlink-Induced Topic Search)[4]を用いて算出される各値を基に、Twitter におけるユーザーの活動を、情報提供・情報収集・情報共有の 3 タイプに分類する方法が提案されている[5]。

さらに、コミュニティ自体を分析するパラメータとして、類似性というキーワードが注目されつつある[6]。ある程度、共通の興味・関心を持つ者同士が集まることによりコミュニティが形成されるという前提を置くと、ある時点では直接的な繋がりを持たない者同士であっても、共通の興味・関心という類似点を持っていることで、その類似点が将来、直接的な繋がりを生み出す要因となる可能性を秘めていると考えることができる。よって、コミュニティ内の類似性を研究することにより、コミュニティの将来予測につながる可能性があると考えられる。

そこで本研究では、SNS における効率的なコミュニティ抽出方法およびノードの属性から得られる類似性を用いた、各コミュニティの特徴づけについて検討を行った。

3. データ収集および分析方法

本研究における、データ収集および分析の全体フローを図 1 に示す。

スクリプト言語 Python を用い、Twitter API を使用することで、Twitter DB から Couch DB(Cluster Of Unreliable Commodity Hardware)へのデータ収集を行い、さらにデータの出力プログラムの作成と実行を経て、収集データの出力も行った。また、Excel および SN 分析ソフト NetMiner[7]を用いてデータの分析を行った。

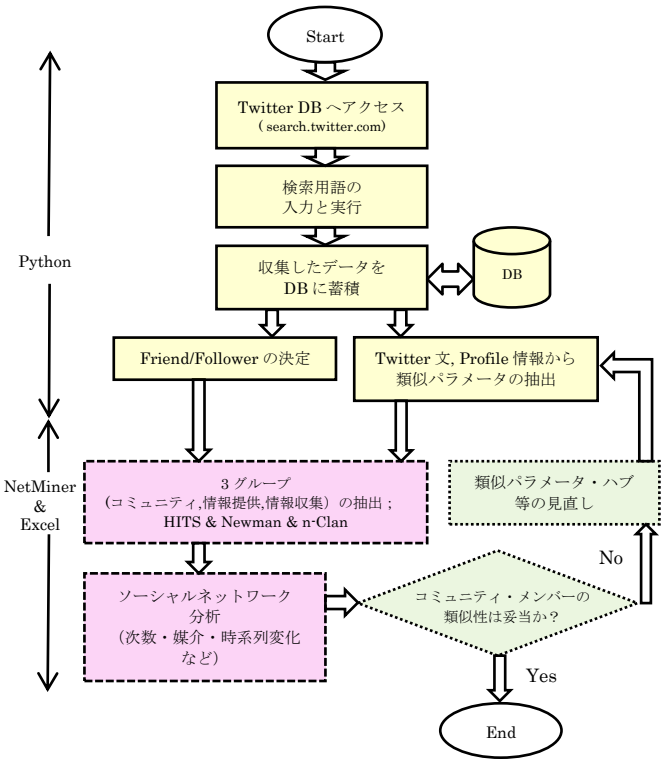

図 1 データ収集と分析フロー

† 日本大学大学院理工学研究科, College of Science and Technology, Nihon University
‡ 日本大学理工学部, College of Science and Technology, Nihon University

3.1 データ収集

3.1.1 データ収集の流れ

Twitter 上の対象アプリケーションを特定するような検索用語とデータの日時等を指定し、データ収集を行った。データ収集のフローを図 2 に示す。

データ収集プログラムに関しては、文献[8]を参考とした。

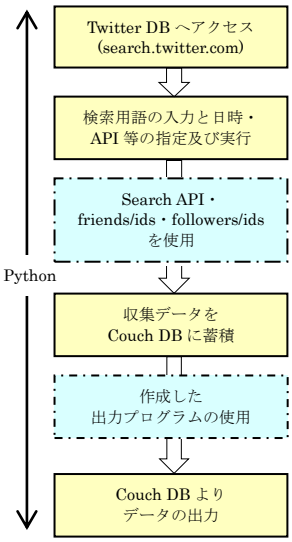

図 2　データ収集フロー

Twitter API の利用には、OAuth 認証が採用されている。OAuth 認証とは、サービス提供者間でユーザーの個人情報を受け渡すことを許可する際の、ユーザー認証である。Twitter API 利用における OAuth 認証の流れには、ユーザーによる情報受け渡し認証と、コンシューマ(API を利用したサービス提供者)とサービスプロバイダ(Twitter)間での、リクエスト・トークンやアクセス・トークンのやり取りによるユーザーの個人情報受け渡しがあり、この作業を経て API の利用が可能となる。

また、Python を用いた Twitter API でのデータ収集およびデータ出力プログラムの作成と実行も行った。データ収集に用いた Twitter API は、Search API・Friends/ids・Followers/ids の 3 つである。

Python で作成した Twitter API の実行プログラム、検索用語、収集データの日時、DB での保存フォルダ名を指定し、コマンドプロンプト上で実行することで、Twitter DB からデータを収集し、Couch DB に蓄積する。蓄積したデータを、ユーザー間のつながりを 0, 1 で表す隣接行列といった、SN 分析ソフト NetMiner でのインタフェースに変換するプログラムを作成し、出力処理を行った後、3.2 に示す分析を行った。データ出力に使用したプログラムの例を、図 3 に示す。

```
# -*- coding: utf-8 -*-
import sys
import couchdb
import httplib
import json

DB = sys.argv[1]

try:
    server = couchdb.Server('http://localhost:xxxx')
    db = server[DB]
except couchdb.http.ResourceNotFound, e:
    print """CouchDB database '%s' not found.
Please check that the database exists and try again.""" % DB
    sys.exit(1)

def print_db_ids(db):
    for doc_id in db:
        doc = db.get(doc_id)
        print doc["from_user"], doc["from_user_id_str"],
doc["id_str"]
        try:
            print doc["text"].encode('cp932')
        except (UnicodeEncodeError):
            print 'UnicodeEncodeError'
        print doc["created_at"]
        print

print_db_ids(db)
```

図 3　データ出力プログラム例

3.1.2 データ概要

今回は、2012 年 12 月に行われた衆議院議員選挙にかかわるデータを対象に分析を行った。収集データの概要を表 1 に示す。収集したツイート件数は 38,205 件、ユーザー数は、18,794 人である。

表 1　データ概要

対象ユーザー	2012 年の衆議院議員選挙にかかわる日本語 Tweet をしたユーザー
使用データ	User ID, Screen Name, Tweet 文, Tweet 時間, Follower & Friend Ship, Profile
データ収集期間	2012 年 12 月 9 日～12 月 15 日 日本時間 AM8:00~AM9:00

3.1.3 データ収集条件

衆議院議員選挙にかかわるデータの収集を行うため、Tweet 文に含まれる用語および含まれない用語の指定を行った。それぞれの検索用語を表 2 に示す。

表 2　データ収集条件

Tweet 文に含まれる用語（OR 検索）	政権・政策・民主・自民・日本未来の党・公明・共産・みんなの党・維新・社民・減税日本・国民新党・新党大地・みどりの風・新党改革・幸福実現党
Tweet 文に含まれない用語（NOT 検索）	西尾維新・都知事

3.2 分析方法

3.2.1 Authority 値と Hub 値を使ったコミュニティ対象ユーザーの決定

有向リンク構造のネットワークに対し、HITS を用いてノードの Hub 値および Authority 値を算出し、その値からユーザーのタイプを決定する。

ここで出てくる Authority とは、情報を発信しているノードを指し、Hub とは、情報を発信している Authority に対してリンクを張っているノードを指す(図 4)。ノードが情報発信によって Authority としての価値を持ち、その値が決まっているとき、ノードの Hub 値は、出リンク先のノードが持つ Authority 値の合計から算出される値である。また、ノードが多くの出リンクを張ることで Authority と繋

がり、Hub としての価値を持つとき、Authority 値も同様に、入リンク先のノードが持つ Hub 値によって決まる値である。

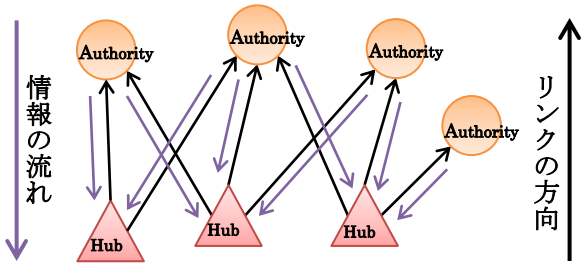

図 4　Authority と Hub

この算出結果より、Twitter 上で高い活動を示すユーザーは、Hub 値のみの高いユーザー(情報収集が目的)、Authority 値のみの高いユーザー(情報提供が目的)、Hub 値・Authority 値ともに高いユーザー(情報共有が目的)と、目的別に 3 タイプに分類することが出来る。

今回は、情報共有を目的として、互いに情報を流通し合うユーザー間での相互活動を対象としたコミュニティ抽出を目的とした為、Hub 値・Authority 値ともに高いユーザーのみを対象に、コミュニティ分析を行った。

3.2.2 コミュニティ抽出方法

SNS の活動記録から得られる情報は、巨大になる傾向がある。その結果、詳細な活動単位としてのコミュニティを、短時間で抽出することは困難である。

そこで、実時間内に計算可能である効率的なコミュニティ抽出方法を提案する。対象ユーザー間ネットワーク生成とコミュニティ抽出のフローを図 5 に示す。

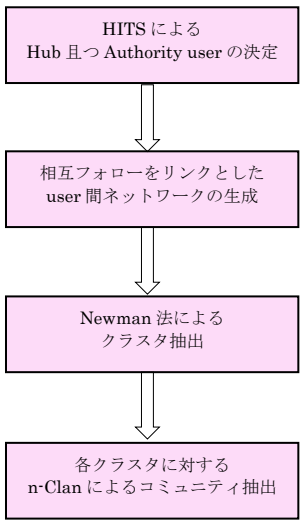

図 5 ネットワーク生成
コミュニティ抽出フロー

まず、HITS により算出した Hub 値および Authority 値ごとに、ユーザーのランキング付けを行う。その後、閾値を 0.05 として足切りし、最終的に Hub・Authority 共に名前の残っているユーザーのみを抽出して、コミュニティ対象ユーザー群リストを作成する。

次に、作成したユーザー群リストに対し、任意の 2 ユーザー間に相互フォローが認められる場合にのみリンクを張り、新たなユーザー間ネットワークを生成する。

生成したユーザー間ネットワークに対し、モジュラリティによってクラスタ抽出処理を行う Newman 法により、クラスタ抽出を行う。

さらに、各クラスタに対して、n=2 の n-Clan によりクラン抽出を行い、そこで抽出されたクランをコミュニティとする。

4. データの分析と考察

4.1 スケールフリー性

一日毎に、次数中心性の値と各次数の発生度数をパラメータとして、次数分布を調査した（図 6）。両軸を対数で表記すると、各日のデータがほぼ直線を描き、次数分布がベキ則に従うことを示した。つまり、少数のノードが、多くのノードと直接リンクを持つことで大きな次数を持ち、残りの大多数のノードが、少ないノードとの直接リンクしか持たないことで小さな次数を持つという、スケールフリー性を示した。

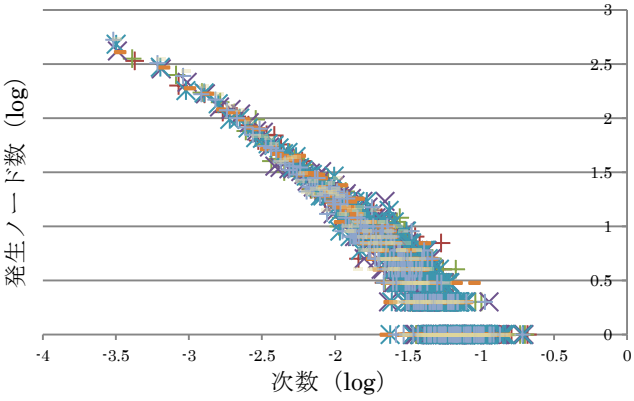

図 6　次数分布

4.2 クラスタ係数と平均経路長

各日ごとに、クラスタ係数および平均経路長を算出した（表 3）。

クラスタ係数は、ネットワーク内ノードの隣接ノード群において、その中で互いに隣接しているノード対の割合である。ここでは、ネットワーク内全ノードのクラスタ係数の平均値を扱っている。値が 1 になるとき、そのネットワークは完全グラフとなる。今回のデータに関しては、0.3 未満の値を取っていることから、完全グラフという視点から見ると、あまり密でないネットワークとして記述されている。

また、平均経路長は 4 未満の値を取っていることから、今回の収集データに関しては、比較的少ないノードを介すだけで、他のノードと繋がることが出来るということが分かった。

表 3　クラスタ係数と平均経路長

	クラスタ係数	平均経路長
12/9	0.270	3.252
12/10	0.263	3.364
12/11	0.250	3.317
12/12	0.263	3.327
12/13	0.242	3.412
12/14	0.250	3.58
12/15	0.239	3.621

4.3 コミュニティ抽出の処理

4.3.1 コミュニティ抽出時間

ケーススタディにおける、今回提案したコミュニティ抽出方法を用いたコミュニティ抽出の所要時間とデータサイズは、以下の通りである(表 4)。

表 4　コミュニティ抽出におけるデータサイズと所要時間

ノード数	リンク数	所要時間
93	1968	46ms
80	1092	78ms
67	845	31ms
107	3104	78ms
91	1978	48ms
106	3298	62ms
94	2104	62ms

※分析にかけるまでのデータ編集時間は含まない

一般的に、n-Clan の計算量は n をノード数とする時、O(2^n) である。このことから、分析データのノード数が増える程べき乗で計算量が大きくなり、計算時間がかかることが予想できる。

一方、今回提案したコミュニティ抽出方法で用いる Newman 法の計算量は、n をノード数, m を総リンク数とする時、O(m*n log n)である。これより、n-Clan に比べて Newman 法は計算量が少なく、計算時間が短くなることが予想される。

今回、分析に使用したデスクトップコンピュータでは、表 4 に示した程度のデータサイズであっても、n-Clan のみのコミュニティ抽出を試みると、計算量の多さから、実時間内でのコミュニティ抽出が出来ないという状況であった。

しかし、今回提案した Newman 法と n-Clan の 2 段階クラスタリングを行うことで、通常のデスクトップコンピュータを使用した場合でも、表 4 に示した程度のデータサイズであれば、1 秒以下のわずかな時間でのコミュニティ抽出が可能となった。

4.3.2 コミュニティ抽出結果

コミュニティの抽出結果を以下に示す（表 5）。

各日ともに、コミュニティ内の最大クラスタ係数が 0.8 前後の値を取り、平均媒介中心性が 0.01 前後の値を取っていることから、コミュニティ内のネットワークが比較的密であり、任意のメンバーを除いた場合でも、非連結になりにくいような構造となっていることが分かる。

表 5　コミュニティの抽出結果

	抽出コミュニティ数	最大コミュニティサイズ	ハブの最大次数	最大クラスタ係数	平均媒介中心性値
12/9	3	46	42	0.711	0.011665
12/10	3	50	42	0.784	0.009139
12/11	2	38	34	0.747	0.011257
12/12	3	49	46	0.760	0.013068
12/13	5	43	42	0.814	0.017009
12/14	2	53	50	0.757	0.005663
12/15	7	58	55	0.700	0.015685

4.4 コミュニティの特徴

4.4.1 コミュニティ内ハブユーザーの決定

抽出した各コミュニティに対し、次数中心性を基に、ハブユーザーの決定を行った。

次数中心性が高いということは、そのコミュニティ内において、より多くのコミュニティメンバーとの間に直接リンクを持つと見なすことが出来る。そこで、最も次数中心性の高かったユーザーを、そのコミュニティのハブユーザーとした。

4.4.2 コミュニティメンバー間の類似性

ハブユーザーの Tweet 文および Profile 情報から、ハブユーザーの属性を抽出した。

共通の興味・関心を持つ者同士が集まることで、コミュニティを形成していると考えると、より多くのコミュニティメンバーとの間に直接リンクを持つハブユーザーこそ、そのコミュニティを代表するような属性を持つと考えることが出来る。そこで、抽出したハブユーザーの属性を、そのハブユーザーが属するコミュニティ内における、コミュニティメンバー間の類似性と仮定した。ユーザーの属性パラメータには、衆議院議員選挙にかかわる政党名を採用しており、肯定・否定を含め、ある政党に対して、興味・関心を持っていると見受けられる Tweet 文や Profile 内の記載が確認できた場合に、該当する政党名の属性を与えている。

今回は、抽出したコミュニティのうち、Community A と Community B の 2 つのコミュニティに対し、ハブユーザーの属性からコミュニティメンバー間の類似性を仮定した。

Community A は、メンバー数 50、類似性パラメータは”生活の党”と”日本未来の党”とした。また、Community B は、メンバー数 55、類似性パラメータは”自民党”と”民主党”とした。

4.4.3 コミュニティメンバー間の類似性検証

2 つのコミュニティ Community A, Community B に対し、ハブユーザー以外のメンバーに対しても、ハブユーザーの時と同様に、Tweet 文および Profile 情報から属性の抽出を行った。その結果を表 6, 7 に示す。今回は、Tweet 文や Profile 情報から属性の特定が出来なかったユーザーや、ハブユーザーと異なる属性のみを持つユーザーに関しては、白抜きにしてある。

表 6　Community A のメンバー属性

A1		A12	日	A23	日	A34	生	A45	
A2		A13	日	A24		A35	生	A46	日
A3		A14	生	A25		A36	生	A47	日
A4	日	A15		A26	日	A37	日	A48	日
A5	日	A16	日	A27		A38		A49	日
A6	生	A17	日	A28		A39			
A7	日	A18	日	A29		A40			
A8		A19	日	A30	生	A41	日		
A9	日	A20	日	A31	日	A42	日		
A10	生	A21		A32		A43	生		
A11	日	A22	日	A33	生	A44	日		

表 7　Community B のメンバー属性

B1		B12		B23	民	B34	自	B45	自
B2	自	B13	自	B24		B35	自	B46	民
B3	自	B14	民	B25	自	B36	自	B47	自
B4	自	B15	自	B26	自	B37		B48	自
B5		B16	民	B27	民	B38		B49	民
B6	民	B17	自	B28	自	B39		B50	自
B7	自	B18	自	B29	自	B40	自	B51	自
B8	自	B19	自	B30	民	B41	自	B52	自
B9		B20		B31		B42	自	B53	民
B10	自	B21	自	B32		B43	自	B54	自
B11	自	B22	民	B33	民	B44	自		

ハブユーザーの属性からコミュニティメンバー間の類似性を仮定したところ、Community A では、49 ユーザー中 33 ユーザー, 約 67%のメンバー属性を網羅するという結果を得た。また、Community B では、54 ユーザー中 43 ユーザー, 約 80%のメンバー属性を網羅するという結果を得た。

5. まとめと今後の課題

SNS における活動の可視化と活動評価を実現するための第一段階として、今回は Twitter を対象に、データ収集の環境構築と、衆議院議員選挙に特化したデータ収集および Twitter におけるコミュニティの効率的な抽出方法と、ノード属性による各コミュニティの特徴づけについて検討した。

知人関係や Web のリンク数など、ネットワークが持つ性質として、幅広く観測されているスケールフリー性について、今回収集した Twitter の衆議院議員選挙にかかわる各日 1 時間分のデータに対しても調査をした。その結果、収集データに対しても、スケールフリー性を持つことが確認できた。

コミュニティ抽出において、HITS によるコミュニティ対象ユーザーの決定および Newman 法と n-Clan による 2 段階のクラスタ抽出を用いたコミュニティ抽出方法の提案と、ケーススタディにおける、提案手法を用いた実時間内でのコミュニティ抽出の実現性を確認した。提案手法を用いることで、ケーススタディで用いた程度のデータサイズであれば、極めて短時間でコミュニティ抽出が可能になるという結果を得た。

さらに、ケーススタディから、抽出したコミュニティに対し、次数中心性によるハブユーザーの決定と、2 つのコミュニティを対象として Tweet 文と Profile 情報からハブユーザーの属性を抽出すると、その属性により、ハブユーザーが属するコミュニティ内メンバーの属性を、かなり高い確率で網羅できる例を得ることが出来た。これにより、ハブユーザーの属性から、そのハブユーザーが属するコミュニティにおけるコミュニティの特徴づけが出来る可能性を得ることが出来た。

ただし、今回提案したコミュニティ抽出方法を用いることで、コミュニティ抽出における高速化が可能となった一方で、抽出・評価出来なくなったコミュニティが存在するという可能性が残っている。Newman 法が、モジュラリティを評価指標としてクラスタ抽出を行っていることから、ネットワーク内でリンクが密である大きな塊としてクラスタを抽出した上で、さらに詳細なコミュニティ抽出を行ってはいるが、この点に関しては、2 段階で抽出したクラスタがコミュニティとしての価値を持つのかという点と合わせて、コミュニティ抽出の精度という観点から、何らかの形で評価をする必要があると考える。

また、ノードの属性抽出に関しても、現時点では 1 つ 1 つ目視によって行っているため、主観に頼っている面がある。さらに、抽出にかなりの時間を要することから、選択した 2 つのコミュニティに対しての分析に留まり、データ全体および各抽出コミュニティに対する検証が行えていないというのが現状である。その為、例えば自然言語処理等を用いた、効率的な属性抽出方法および仮定される類似性の妥当性についての検討も、今後合わせて行っていく予定である。

参考文献

[1] COMPUTERWORLD, http://www.computerworld.jp/common/print/news/206880
[2] Mohsen Jamali, Hassan Abolhassani; “Different Aspects of Social Network Analysis”, Web Intelligence, 2006. WI 2006. IEEE/WIC/ACM International Conference, ISBN: 0-7695-2747-7, pp.66-72, (2007)
[3] 北原, 吉開, “アフィリエーションネットワークを用いた活動評価法の提案と評価”,信学技法 SITE2011-55 pp.317-322,2012.
[4] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment", Journal of the ACM, 46(5):604–632, 1999.
[5] A.Java, X.Song, T.Finin, B.Tseng, “Why we twitter: understanding microblogging usage and communities”, Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis Pages 56-65.
[6] Yang ZHANG, Yao WU, Qing YANG, “Community Discovery in Twitter Based on User Interests”, Journal of Computational Information Systems 8: 3 (2012) 991–1000
[7] NetMiner, http://www.netminer.com/index.php
[8] Matthew A. Russell, “入門 ソーシャルデータ-データマイニング、分析、可視化のテクニック” O’Reilly Japan (2011).