

Twitter の Favorite 機能を用いたユーザー嗜好分析

加藤 翔子[†] 伏見 卓恭[†] 斉藤 和巳[†]

[†] 静岡県立大学 経営情報学部 〒 422-8526 静岡県静岡市駿河区谷田 52-1

E-mail: †{b09032,j11507,k-saito}@u-shizuoka-ken.ac.jp

あらまし 消費者,あるいはユーザーの嗜好は現実社会における多様な経済活動の本質であり,そのモデリングは重要視されるべき事項である。しかし,潜在的には様々な指標が存在することが予想される反面,数値化され分析し得るものは多くはない。そこで本論文では,新たなユーザーの嗜好モデルの構築に向けた一歩として, Twitter のツイートをお気に入り登録して保管できる Favorite 機能を,有向ネットワークや2部グラフととらえて,ユーザー間の関係を探した。その結果,連結成分分解を行うことで大規模な連結成分の存在を,次数分布を調べることでスケールフリー性を,モチーフ分析を行うことで相互リンクを含むものに顕著な特徴がみられることを,2部グラフ可視化では Favorite されるユーザーがツイート内容により大きく二つに分けられることを,それぞれ確認した。

キーワード 有向ネットワーク, 連結成分分解, 次数分布, ネットワークモチーフ, 2部グラフ

Analyses of the Tendency for Users with the Favorite system of Twitter

Shoko KATO[†], Takayasu FUSHIMI[†], and Kazumi SAITO[†]

[†] School of Administration and Informatics, University of Shizuoka

52-1 Yada, Suruga-ku, Shizuoka, 422-8526 Japan

E-mail: †{b09032,j11507,k-saito}@u-shizuoka-ken.ac.jp

Abstract The tendency for the consuming public, or the users, is the essence of diverse economic activities, and modeling of those is an important matter. However, because of the high dependence of individual, we seldom obtain such data for analyses. In this paper, we analyzed the Favorites system of Twitter from the point of view of directed networks and bipartite graph, and clarified relations between the users to make the basis on which the new tendency model are grounded. As results of the analyses, we observed a large connected component in connected components decomposition, a scale free property in degree distribution, a conspicuousness of mutual links in network motifs, and two clusters of popular users which was divided according to their tweets in the bipartite graph.

Key words directed networks, connected components decomposition, degree distribution, network motifs, bipartite graph

1. はじめに

日本のような資本主義国家において,現実社会における消費者の嗜好や風潮,流行などの流れを把握し操作することは,経済的成功を収める為の重要な要因の一つである。従って,その動向を理解するために,実データとして収集し分析・モデリングを行うことも,重要視されている研究対象である。しかし,抽象的概念である嗜好を具体的な数値へデータ変換することは容易なことではなく,そのため存在するデータの数も限られており,新たなユーザー嗜好の指標の発見が待たれている。

本論文では, Twitter のツイートをお気に入り登録して保管できる Favorite 機能に着目し,新たなユーザー嗜好モデルの構築へ向けての基礎として,有向ネットワークや2部グラフの観

点から分析する。具体的には,連結成分分解することによりネットワークの構造を概観し,ノード間の相互関係を把握するためにモチーフ分析をし,2部グラフ化することにより Favorite するノードとされるノードに分けて Favorite 構造の動向を俯瞰する。

Favorite 機能と同様にユーザーの嗜好や興味を反映した情報をもつものに,ソーシャルブックマークが挙げられる。ソーシャルブックマークに関しては,太田ら [1] がソーシャルブックマーク分析のユーザーへの影響を調査し,その影響への対処として先行性を導入した分析を提案するなど,研究が盛んに行われているが,こうした情報をネットワークとして分析したものは既存の研究には見られず,従って本研究は意義のあるものと考えられる。

本論文の構成は以下となる．まず，本研究に用いた分析法について説明する．次に，Favorite 機能の概要や分析に用いたデータについて説明するとともに，評価実験による結果を報告する．最後に，本研究のまとめについて述べる．

なお，本研究で用いるネットワークを，本論文では「Favorite ネットワーク」と呼ぶ．また，後述の背景から，Favorite する・されるといったユーザーの行動をそれぞれ「ふぁぼる・ふぁぼられる」と呼ぶこととする．

2. 分析法

ユーザーをノード集合 $V = \{v_1, v_2, \dots, v_n\}$ とし，あるノード v_i が他ノード v_j をふぁぼったとき， v_i から v_j への有向リンク e_{ij} が生成されるとする．このリンクの集合を $E = \{e_{11}, e_{12}, \dots, e_{lm}\}$ としたとき，Favorite ネットワークは有向グラフ $G = (V, E)$ により表現される．本稿では，このグラフを分析する上で用いた，以下の 4 つの分析法について論じる．

- 連結成分分解
- 度数分布
- ネットワークモチーフ
- 2 部グラフ可視化

2.1 連結成分分解

グラフ上の任意の 2 ノードの間に道が存在するグラフを連結グラフという．ここで道とはノードの列であり，各ノードとその次のノードとの間にリンクが存在する．また，極大で連結な部分グラフのことを連結成分といい，リンクを有向・無向のどちらとみなすかによって強連結成分と弱連結成分に分類される．

有向グラフにおいて，グラフ上の任意の 2 ノード間に有向の道が存在するとき，そのグラフは強連結であると定義され，極大で強連結な部分グラフは強連結成分 (strongly connected components, SCC) という．

また，無向グラフとして考えたとき，同様にグラフ上の任意の 2 ノード間に無向の道が存在するとき，そのグラフは弱連結であると定義され，極大で弱連結な部分グラフは弱連結成分 (weakly connected components, WCC) という．

2.2 度数分布

無向グラフにおいて，グラフ上の任意のノード v_i に繋がっているリンクの本数をそのノードの度数 k_i という．また，有向グラフにおいては，任意のノードを指しているリンクの本数を入次数，ノードから出ているリンクの本数を出次数という．

2.2.1 スケールフリー性

現実世界に存在する多様なネットワークは，巨大で複雑な構造を有しているが，一定の共通する性質を見出すことができる．それら性質の一つとしてスケールフリー性（度数分布のべき乗則）がある．スケールフリー性とは，一部のノードが他の多くのノードとリンクで繋がっており，大きな度数を持っている一方で，大多数のノードはごく少数のノードとしか繋がっておらず，度数は小さいという性質である．出次数の大きなノードはハブとも呼ばれる．

スケールフリー性を持つネットワークの例としては，World-WideWeb が挙げられる [2] [3]．WWW ではごく少数の有名サ

イトが数百万単位のリンクを集めているが，大多数のサイトはわずかなリンク先からしかリンクされていない．

2.3 ネットワークモチーフ

ネットワークモチーフとは，複雑ネットワークを特徴づける指標である．3 ノードの場合のモチーフパターンを図 1 に示す．ここで，3 ノードをモチーフパターンに分類するために，全てのノードが少なくとも 1 本以上の入次数または出次数を持つことを条件とする．また，モチーフはグラフの同型も同時に考慮するため，3 ノードで構成されるモチーフは最終的に 13 パターンとなる [4]．

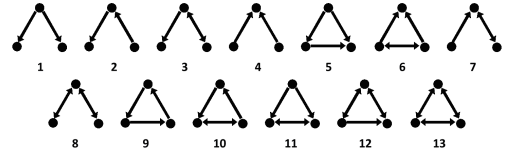


図 1 ネットワークモチーフ

2.3.1 ランダムネットワーク

ランダムネットワークは，実ネットワークのリンクをランダムに張り替えることにより作成される．以下に本研究に用いるランダムネットワーク作成条件を示す．

- ノード毎の入次数と出次数は固定する
- 全てのリンクを張り替える
- 張り替え先が自己リンク・多重リンクとなる場合は，張り替え先を変えて張り替える

2.3.2 Zscore

Zscore とは，実ネットワークとランダムネットワークとのモチーフパターンの出現数に有意差があるかを検定するための評価尺度である．以下に Zscore 値の求め方について説明する．実ネットワークにおけるモチーフパターン i の出現した回数を x_i ， j 番目に作成したランダムネットワークにおいてモチーフパターン i が出現した回数を $y_i^{(j)}$ とする．この時，モチーフパターン i の Zscore は次式で定義される．

$$Z_i = \frac{x_i - \mu_i}{\sigma_i} \quad (1)$$

ここで， $\mu_i = \frac{1}{L} \sum_j y_i^{(j)}$ ， $\sigma_i = \sqrt{\frac{1}{L} \sum_j (y_i^{(j)} - \mu_i)^2}$ である．ここで， L はランダムグラフの作成回数であり，本実験では $L=1000$ とした．

Zscore が正で大きいモチーフパターンほど，そのネットワークにおいて統計的に有意に頻出するといえる．すなわちそのネットワークの特徴的なモチーフとなる．逆に Zscore 値が負で絶対値の大きいモチーフパターンほど，有意に出現しないといえる．

2.4 2 部グラフ可視化

一度でもふぁぼったことのあるノード集合を V_A ，ふぁぼられたノード集合を V_B とし， V_A と V_B を構成するノードは異なるものとみなすと， V_A から V_B へのリンクのみとなり，2 部グラフ G' を構成する．

- $\forall u, v \in V_A, (u, v) \notin E$

- $\forall w, x \in V_B, (w, x) \notin E$

ここで、 $|V_A| = M, |V_B| = N$ とする。

2部グラフの可視化法として久保田らのSE (Spherical Embedding) 法 [5] を用いる。この可視化法は、類似したリンク関係を持つ、つまり似ているノード同士を近くに、異なるリンク関係を持つ、つまり似ていないノードを遠くに配置するという方法である。同SE法では、グラフ G' の隣接行列 $A = \{a_{ij}\}$ に対して中心化を施す [6]。隣接行列 A は、 $M \times N$ の矩形行列であり、 $\forall i \in V_A, \forall j \in V_B$ に対して $\{i, j\} \in E$ なら $a_{ij} = 1$ でそのほかは $a_{ij} = 0$ である。中心化行列 $H_M = I_M - \frac{1}{M} \mathbf{1}\mathbf{1}^T$ および $H_N = I_N - \frac{1}{N} \mathbf{1}\mathbf{1}^T$ と定義すると、左右より中心化された隣接行列は $B = H_M A H_N$ となる。ここで、 I_N は、 $N \times N$ の単位行列であり、 $\mathbf{1} = (1, \dots, 1)^T$ である。同SE法では、行列 B の要素を $\{b_{ij}\}$ とし、以下の目的関数を最大化するようにノード座標行列 $X = [x_1, \dots, x_M]^T$ および $Y = [y_1, \dots, y_N]^T$ を求める。

$$J(X, Y) = \sum_{m=1}^M \sum_{n=1}^N b_{m,n} \frac{\mathbf{x}_m^T \mathbf{y}_n}{r_A r_B} + \frac{1}{2} \sum_{m=1}^M \lambda_m (r_A^2 - \mathbf{x}_m^T \mathbf{x}_m) + \frac{1}{2} \sum_{n=1}^N \mu_n (r_B^2 - \mathbf{y}_n^T \mathbf{y}_n). \quad (2)$$

3. 評価実験

分析に用いたデータと実験結果、それに基づく考察を述べる。

3.1 ネットワークデータ

本論文で取り扱うデータは、Twitter の Favorite 機能を有向ネットワーク化ないし2部グラフ化したデータである。Twitter (<http://twitter.com/>) には自分や他人のツイートをお気に入り登録し、いつでも閲覧できる Favorite 機能が装備されている。この機能は全ユーザーに標準装備されており、相手のツイートを閲覧する権限を持っていれば、自分のツイートを含むどのツイートでも登録を行うことができる。Twitter はリアルタイムに情報を共有できる点に最大の特徴があるが、逆に過去のツイートを見たい場合は最新のものから遡っていく必要があるため、この Favorite 機能は、何度も見返したいツイートに遭遇した時などに有用な機能である。特に日本のユーザーには、この機能を用いてツイートをお気に入り登録することを指す「ふぁぼる」という造語が生まれるほど利用され親しまれている。この機能のユーザーをノード、「ふぁぼる・ふぁぼられる」の関係をリンクとしてネットワークを構築した。

データの取得期間は2011年7月1日から同月31日までであるが、後述の理由から3.2.4節では2011年7月1日から同月7日までのデータを用いる。このネットワークは74,658ノードと1,046,215リンクをもつ有向ネットワークで構成される。また、自己リンクは除外している。Favorite ネットワークのデータは「Favotter」(<http://favotter.net/>) の「今日の人気」ページから取得したものである。Favotter の仕様上、取得できるツイートは5人以上にふぁぼられたものに限られる。また、Favotter 上の表示でスクリーンネームに2バイト文字を有するノードが存在するが、このようなノードは、データを取

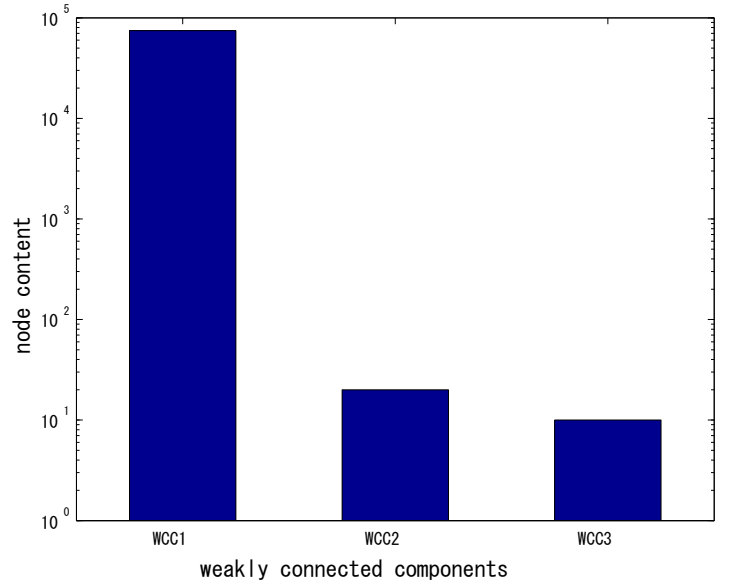


図2 弱連結成分

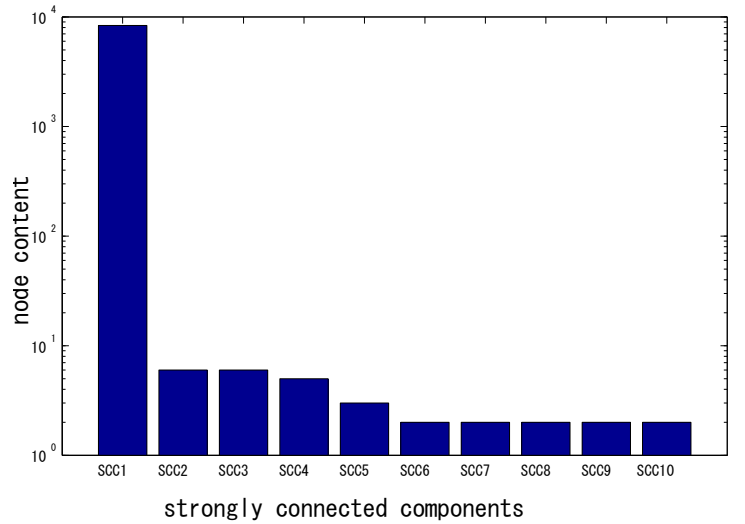


図3 強連結成分

得・分析するプログラムを作成する上で除外した。

3.2 実験結果

本節では、2.節で説明した各分析法を用い Favorite ネットワークを分析した結果とその考察を記す。

3.2.1 連結成分分解

図2と図3には、Favorite ネットワークを2.1で示した弱連結成分分解と強連結成分分解にかけた結果を、それぞれヒストグラムとして示す。ここで横軸はコンポーネント番号、縦軸は含有ノード数をそれぞれ指す。

図2より、大規模である Favorite ネットワークだが、全体の殆どを占める74,570ノードが最大弱連結成分に含まれることがわかる。従って、3.2.2以降の実験では最大弱連結成分を対象として分析を行った。また、図3の強連結成分は、すべて図2中のWCC1、すなわち最大弱連結成分に含まれることを記しておく。

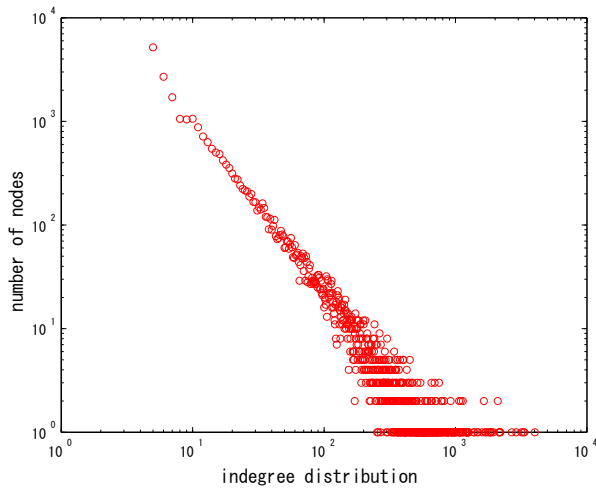


図 4 入次数分布

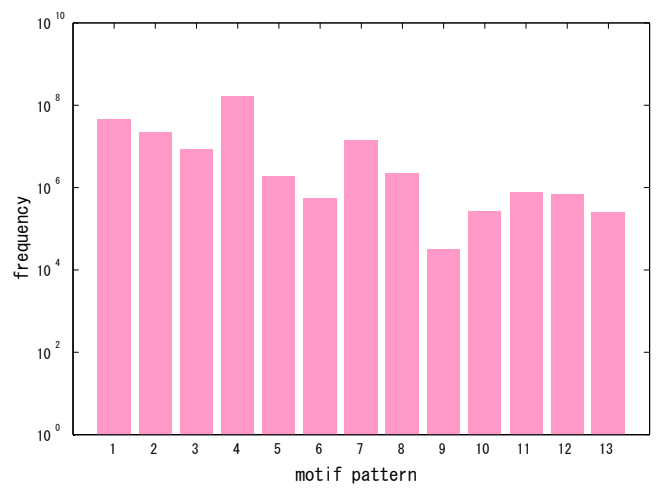


図 6 モチーフ分析 (出現頻度)

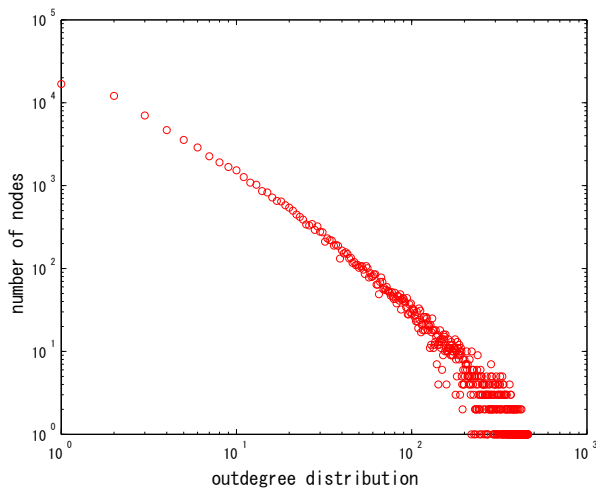


図 5 出次数分布

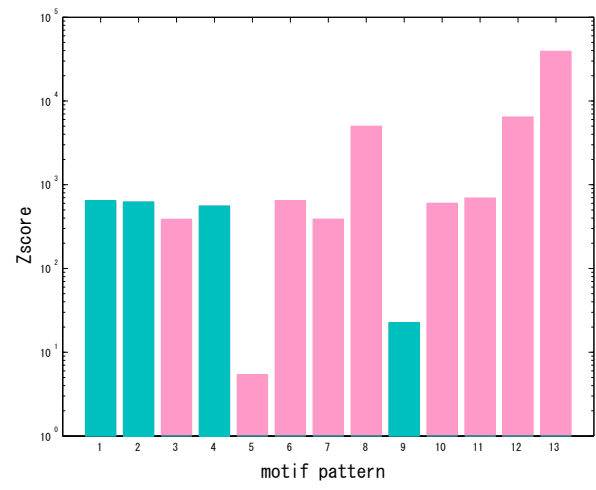


図 7 モチーフ分析 (Zscore)

3.2.2 次数分布

図 4 には、2.2 で示した入次数の分布を示す。ここで横軸は入次数、つまり何人にふぁぼられたかを、縦軸はノード数を指す。また、入次数が 4 のノードが 11 個あり、3.1 節で述べた「取得できるツイートは 5 人以上にふぁぼられたものに限られる」という制約と矛盾するが、これは同節で記した 2 バイト文字問題によって計算から除外されたノードがあるためである。

図 5 には、2.2 で示した出次数の分布を示す。ここで横軸は出次数、つまり何人に対してふぁぼったかを、縦軸はノード数を指す。

図 4 と図 5 より、入次数と出次数のどちらを見ても、2.2.1 で示したスケールフリー性を持つことがわかる。

3.2.3 モチーフ分析

図 6 と図 7 には、2.3 で示したネットワークモチーフにおける出現頻度と Zscore による統計的有意性を、それぞれヒストグラムとして示す。ここで横軸は図 1 で示したモチーフ番号を、図 6 の縦軸は出現頻度、図 7 の縦軸は Zscore の絶対値を指す。また、図 7 において青で表示されたモチーフは Zscore 値が負であること、赤で表示されたモチーフは Zscore 値が正

であることを指す。

図 7 より、8, 12, 13 のモチーフが統計的に有意に頻出していることがわかる。

3.2.4 2部グラフ可視化

図 8 には、2.4 で示した 2 部グラフ可視化を行った結果とノードの注釈を示す。なお、3.1 でも記したが、3.2.3 までの有向ネットワーク分析に用いたデータは大規模であり、リンク数が多く可視化する際にわかりにくくなるため、データの集計期間を短くして 2 部グラフ化している。ここで、外円に配置されたノードはリンクを出しているノード、すなわちふぁぼりノード V_A であり、内円に配置されたノードはリンクに指されているノード、すなわちふぁぼられノード V_B である。また、ふぁぼり数・ふぁぼられ数の多い上位 10 ノードを、それぞれ大きな点で表示し、多重度が 10 以上のリンクのみ灰色の直線で表示している。また、注釈における大カッコ内の半角英数字は代表ノードのスクリーンネームである。NHK_PR は NHK 広報局の公式アカウント、sakitoshinao は IT 分野を主とするフリージャーナリストのアカウントであり、いずれもその時々ニュースやそれに対する意見などをツイートしている。一方、kaiten_keiku や Satomii_Opera は、いずれも匿名性を有する

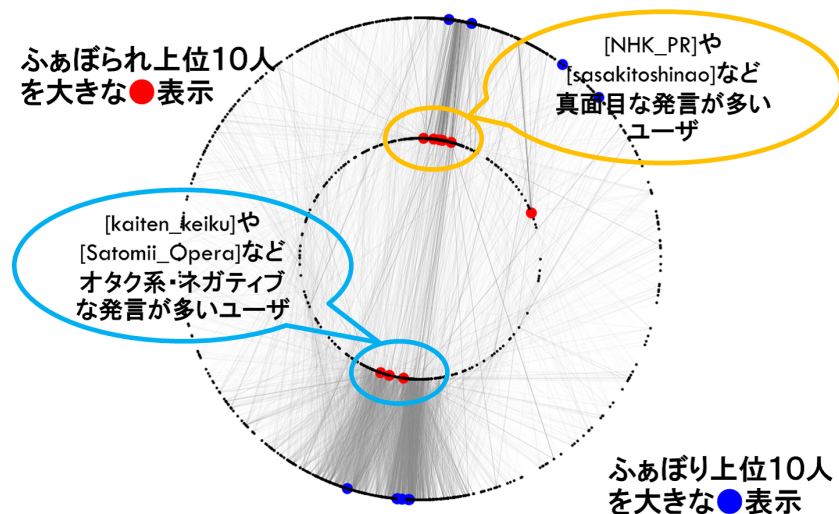


図 8 2部グラフ可視化

一般的な Twitter ユーザーの 1 アカウントであり、自虐的な内容で読者の共感や笑いを誘うツイートや、サブカルチャーに関するツイートを頻繁に投稿している。

この図より、ふぁぼられやすいノード群は大きく 2 つのクラスターに分けることができ、うち 1 つは真面目な内容を、もう一方はネガティブな発言やオタク系内容を含むツイートを投稿していることがわかる。

3.3 分析結果の考察

連結成分分解により、ノードの殆どが、ふぁぼる・ふぁぼられるいずれかの形で他ノードと連結していることが判明したが、これを、次数分布の調査で判明したスケールフリー性とともに考察すると、ハブ的役割のノードの存在が他のノード間を連結させていると予想できる。

また、スケールフリー性について、入次数においては、多くの支持を得られるユーザーは限られた数しか存在せず、一般的なユーザーが得られるふぁぼられ数は微々たるものであることが予想される。出次数においては、Favorite 機能に対する思い入れや熱心さがユーザーによって不均質であり、1 日にいくつもふぁぼるユーザーも少数存在するが、一般的なユーザーは一ヶ月間でも数える程度しか行わないことが予想される。

また、ネットワークの 3 者関係をみると、モチーフ分析から 8, 12, 13 のモチーフが統計的に有意に頻出しているが、図 1 に照らし合わせると、前述した 3 つのモチーフはいずれも 2 つ以上の相互リンクを含むことがわかる。さらに他のモチーフに着目すると、図 7 より Zscore 値が負であるモチーフは 1, 2, 4, 9 であるが、この 4 つのモチーフはいずれも相互リンクを含まないことがわかる。このことから、ふぁぼられた相手には自分もふぁぼり返すという慣習の存在可能性や、Favorite ネットワークが嗜好に基づくデータからなるため、嗜好が似たノード同士は相互リンクが生まれやすい可能性が考察される。

また、より支持を得やすいノード群は、真面目なツイートをするノード群とネガティブ・オタク系なツイートを発するノード群とに分けられたが、これは、Favorite 機能のユーザーに好ま

れやすいツイートを示唆しており、嗜好の流れを把握する上で意義のある結果となった。

4. おわりに

新たなユーザー嗜好モデルの構築への基盤として、Twitter の Favorite 機能を有向ネットワークと 2 部グラフの観点から分析した。その結果、最大弱連結成分が殆どのノードを網羅することが、入次数・出次数のいずれもスケールフリー性を有することが、相互リンクを含むモチーフが統計的に有意に頻出することが、多くの支持を得るノードはそのツイート内容で大きく二つに分けられることが、それぞれ明らかになった。

今後は、さらに多様な分析法を実施し、Favorite モデルの提案に向けて研究を進めていく予定である。

謝 辞

本研究は、豊田中央研究所との共同研究および、科学研究費補助金基盤研究 (C) (No. 23500312) の補助を受けた。

文 献

- [1] 太田飛鳥, 松澤智史, 武田正之, “ソーシャルブックマーク分析のユーザへの影響と先行性による解決”, 第 9 回情報科学技術フォーラム, 2010.
- [2] R. Albert, H. Jeong, A-L. Barabasi, “Diameter of the World-Wide Web”, Nature 401, 130-131, 1999.
- [3] A. Broder, et al. “Graph structure in the web”, Proc. of World Wide Web Conference9, 2000.
- [4] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii and U. Alon, “Network motifs: simple building blocks of complex networks.”, Science 298, pp. 824-827, 2002.
- [5] 久保田大和, 伏見卓恭, 斉藤和巳, “2 部グラフ可視化法の高速化”, 第 3 回データ工学と情報マネジメントに関するフォーラム, 2011.
- [6] W. Torgerson, “Theory and methods of scaling”, Wiley New York, 1958.