F-030

Twitter ユーザ間の興味の重なりの解析
Analysis of Interests Similarity between Twitter Users

遠藤 福富美† 武田 利浩† 平中 幸雄†
Fukutomi Endo Toshihiro Taketa Yukio Hiranaka

1 はじめに

近年，SNS などのソーシャルメディアによる情報伝播が注目されている。ソーシャルメディアのユーザ間の興味の重なりが分かれば，情報推薦などのサービスに応用できると考えられる。これまでにソーシャルメディアの解析が盛んに行われているが，主にソーシャルグラフの形状に関するもので，ユーザの日記や送受信したメッセージなどの言語情報を用いた解析がほとんど行われていない。

本研究の目的は Twitter[1]のツィート情報を用いてユーザ間の興味の重なりを明らかにすることである。本稿では，Twitter ユーザのツィート情報からソーシャルグラフとユーザの興味情報を抽出、解析までの一連の手法及び解析結果を報告する。

2 解析方法

本稿では，Twitter ユーザのツィート情報からユーザ間のソーシャルグラフ及びツィートしたユーザの興味を示す「タグ情報」を抽出して解析を行った。解析は大きく分けて２つの部分，「ソーシャルグラフの解析」及び「ユーザ間の興味の重なりの解析」で構成されている。

2.1 ソーシャルグラフの解析

ソーシャルグラフとは「人間関係図」である。グラフ中の「ノード」は「人間」を表し，「エッジ」は「人間関係」を表す。身近にあるものを例にすると，「家系図」がある種のソーシャルグラフに相当する。ソーシャルグラフを解析することによってそのソーシャルグラフの状態を定量化でき，より正確にグラフの性質を把握することができる。また，指標に基づいて解析した結果を数値で表すことによって，他のソーシャルグラフと容易に比較することができるなどのメリットがある。本稿では，まず Twitter ユーザのツィートからソーシャルグラフを抽出する。次に，従来で使われてきた指標を用いてソーシャルグラフを解析し、考察する。ここではノード数平均次数、最大経路長、平均経路長、平均クラスター係数という 5 つの指標で解析した結果を報告する。

Twitter ユーザのツィートからソーシャルグラフを抽出する方法を説明する。Twitter はツイート内容に「@ユーザ名」を付けることでシステムがユーザ名と認識し，相手のタイムラインに自分のツイートを表示させることができる。つまり，「@ユーザ名」を使用することで Twitter ユーザ間で相手を指定して話しかける事ができる。本稿はその特性を利用してソーシャルグラフを抽出した例えば，ユーザ A が「@B おはよう!」とツィートした場合，ユーザ A はユーザ B に対してツイートしたので２人のユーザはつながっているとする。

† 山形大学 Yamagata University

2.2 ユーザ間の興味の重なりの解析

興味の重なりを調査するために，まずユーザの興味を示す情報が必要。本稿では Twitter ユーザのツィートから抽出した名詞をそのユーザの興味を示す情報とし，「タグ情報」と言う。例えば，ユーザ A が「新しいパソコンを買った」とツィートした場合，ツィート中の名詞「パソコン」がユーザ A のタグ情報とする。タグ情報を比較することでユーザ間の興味の重なりを調べることができると考えられる。

ユーザ A とユーザ B のタグ情報がある場合，ユーザ A から見たユーザ B との興味の重なりは，ユーザ A とユーザ B のタグ情報が一致した数からユーザ A のタグ数で割った値と定義し，式（1）となる。

$$S_{AB}=\frac{T_{A\wedge B}}{T_A} \quad \cdots\cdots(1)$$

ここで，S_{AB} はユーザ A から見たユーザ B との興味の重なり，$T_{A\wedge B}$ はユーザ A とユーザ B のタグが一致した数，T_A はユーザ A のタグ数を表す。

逆にユーザ B から見たユーザ A との興味の重なりは S_{BA} となって，分母が T_B となる。視点を変えるとユーザ A とユーザ B 間の興味の重なりは違った値が出る。それはユーザがそれぞれ持つタグ情報の数が異なることを考慮したものである。

3 解析結果

3.1 ソーシャルグラフの解析結果

文献[2]によれば，一般的に SNS はスモールワールド性とスケールフリー性を持つ。スモールワールド性はソーシャルグラフが短い平均経路長（2.0 前後），高いクラスター係数（0.2 以上）を持つ事。また，文献[3]よりスケールフリーネットワークの特徴は，次数分布がべき乗則に従うことである。

表 1 に解析結果を示す。表 1 より本稿のソーシャルグラフでは平均経路長が 8.12 で，平均クラスター係数が 0.06508 となっているためスモールワールド性があると言えない。次に，次数分布からスケールフリー性を確認する。図 1 に次数分布を示す。実線は $f(x)=x^{-2}$ を示している。次数分布が正規分布よりは分布の裾が広いがべき則ほどではないことがわかる。よって，本稿のソーシャルグラフからスケールフリー性があると言えない。本稿のソーシャルグラフから「スモールワールド性」と「スケールフリー性」を確認できなかった。理由としては「データの期間が短い」と考えられる。Twitter ユーザの「@ユーザ名」を含むツイートからソーシャルグラフを抽出している。ソーシャルグラフの形が，データ収集の期間中にどのくらい多くの「@ユーザ名」をツイート

したのかに関係している。本稿で使用したデータは Twitter ユーザが一週間の間にツィートしたデータで，より長期間のデータからソーシャルグラフを抽出することでその問題が解決できると考えられる。もしくは，Twitter ユーザ間の「フォローしている」、「フォローされている」情報を使う事でも一般的な SNS ソーシャルグラフの特徴を持つソーシャルグラフが出来ると考えられる。

表 1:ソーシャルグラフの解析結果

ノード数	111109
平均次数	2.83
最大経路長	31
平均経路長	8.12
平均クラスター係数	0.06508

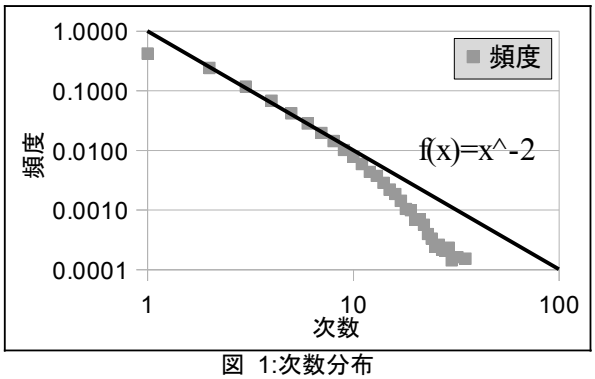

図 1:次数分布

3.2 ユーザ間の興味の重なりの解析結果

図２に，ユーザ間の距離と興味の重なりの関係を示す。横軸は距離で，ソーシャルグラフにおいて注目するユーザに対応する２つのノード間のエッジ数を用いる。縦軸の興味の重なりは式（１）で計算した値を全ユーザで平均値をとったもの。距離が１，即ち隣接ユーザ間の興味の重なりが最大で，7.2%となった。ユーザ間の距離が離れるにつれて興味の重なりも下がって行き，そして距離が 8 の時，興味の重なりがほぼ 0%に近い値となった。

図３に，隣接ユーザ間で興味の重なりを求めた時の分布を示す。横軸はユーザ間の興味の重なりの区間で，縦軸は興味の重なりがその区間で出現した回数を表している。例えば区間 5%<S<=10%となった回数は約 6 万回となっていて，興味の重なりが 0%となった回数に次いで二番目に高い値となっている。また，ほとんどの場合ユーザ間の興味の重なりが 3 割未満であることがわかる。

図４に，全ユーザが持つタグ数の分布を示す。横軸はタグ数を表していて，縦軸はそのタグ数を持つユーザの数を表している。4 個のタグを持つユーザ数が一番多く，それ以降はタグ数が増えるに連れてそのタグ数を持つユーザの数が徐々に減っていくことがわかる。全ユーザの 47［%］は持つタグ数が１０個以下で，全タグの種類 122520 に対して極めて少ない結果となった。ユーザの持つタグ数はそのユーザのツイート数と関係しているのでより長期間のデータからタグ情報を抽出ことでユーザの持つタグ数が少ない問題が解決できると考えられる。ちなみに，タグ情報を抽出出来なかったユーザ，即ちタグ数が 0 となっているユーザはソーシャルグラフから排除されているので，ここでは表記していない。

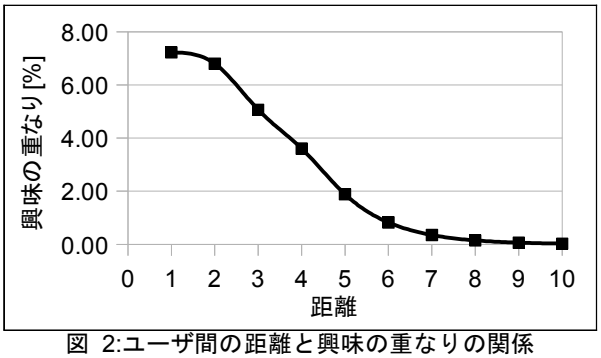

図 2:ユーザ間の距離と興味の重なりの関係

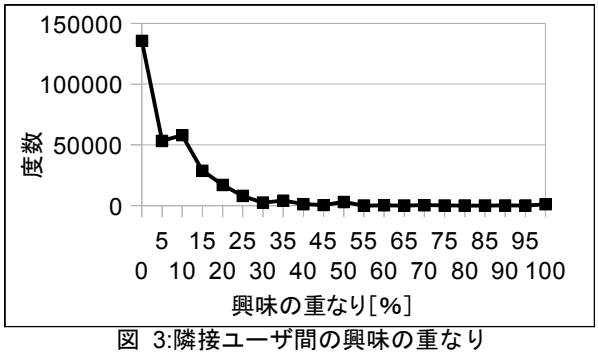

図 3:隣接ユーザ間の興味の重なり

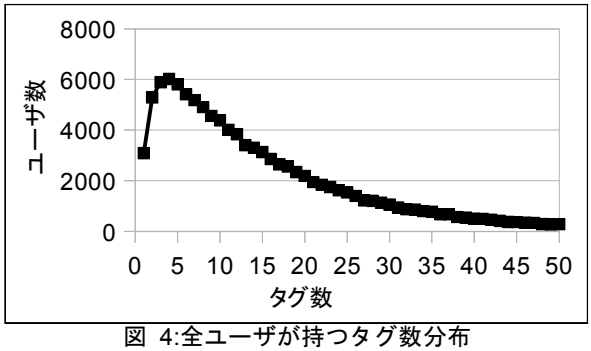

図 4:全ユーザが持つタグ数分布

4 まとめ

本稿では Twitter のデータからソーシャルグラフを抽出し，解析を行った。結果として，ソーシャルグラフから「スモールワールド性」と「スケールフリー性」を確認できなかった。「@ユーザ名」を含むツィートの数はソーシャルグラフの形に影響を与えるので，同一ユーザを長期間に渡って収集したデータからより理想的なソーシャルグラフを抽出することができると考えられる。また，興味の重なりの解析ではユーザ全体が持つタグ数が少ない、全体の重なりが低いといった問題があるものの，ユーザ間の距離が離れることにつれて興味の重なりが低くなることが確認された。

今後の課題として，タグの使用頻度が興味の強さを示すパラメータと考えられるので，興味の重なりを計算する際，タグの使用頻度を考慮した解析をする事で，解析結果を利用した情報推薦の精度が上がると予想される。

参考文献

[1] Twitter, http://twitter.com/about
[2] 鳥海不二夫, 山本仁志，諏訪博彦, 岡田勇, 和泉潔, 橋本康弘, “大量 SNS サイトの比較分析”, 人工知能学会論文誌, 25 巻 1 号 SP-I (2010 年).
[3] 増田直紀, 今野紀雄, “複雑ネットワーク―基礎から応用まで”, 近代科学社 (2010).