

ネットワーク構造に基づいたミニブログの統計分析

Statistical Analysis of Microblogging Based on Network Structure

瀬戸 秀隆 †

Hidetaka Seto

木村 昌弘 †, ‡

Masahiro Kimura

1 はじめに

最近、Web 上での新たなコミュニケーションツールとして、ミニブログが注目されている。米国での主要なミニブログサービスである Twitter の分析も、すでに始まっている [1]。近年、ネットワーク上での情報拡散現象に関心が集まっている [2] が、ミニブログは、1 行という短い記事のみをポストするため迅速なコミュニケーションが実現可能であるので、情報拡散の媒体としても重要である。したがって、ミニブログにおけるコミュニケーションの分析は重要な研究課題である。

本研究では、日本での主要なミニブログサービスである「もごもご」に注目し、そこでミニブログユーザ間のコミュニケーションについて調べる。「もごもご」には、「もごリンク」と呼ばれるミニブログユーザ間のブログロールリンクが存在し、「もごリンク」ネットワークが構築されている。また、ミニブログユーザは他のミニブログユーザがポストした記事にコメントを残すことができるので、ミニブログにはポストされた記事情報とともにそれらに付けられたコメント情報も蓄積されている。これらは、ミニブログユーザ間のコミュニケーション情報と考えられる。本論文は特に、「もごリンク」ネットワーク構造とミニブログユーザの被コメント数との相関関係について分析する。

2 分析データ

「もごもご」をクロールすることにより、2008年5月11日から2008年6月7日までの約1ヶ月間の「もごもご」内のミニブログデータを収集した。ミニブログユーザ総数は10,000、記事総数は16,547、コメント総数は26,735、「もごリンク」総数は15,482であった。

2.1 次数分布

まず、「もごリンク」ネットワークの次数について調べた。次数が0であるノードは7917個あった。すなわち、全体の約80%が孤立ノードであった。図1に「もごリンク」ネットワークの

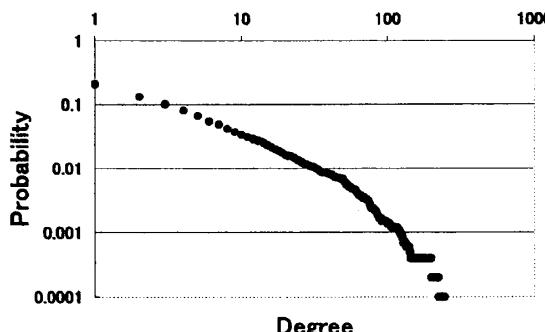


図1 「もごリンク」ネットワークの次数分布

† 龍谷大学大学院 理工学研究科 電子情報学専攻

‡ 龍谷大学 理工学部 電子情報学科

次数分布を示す。べき則的な性質が観察される。

2.2 被コメント数分布

次に、ミニブログユーザごとの被コメント数の累積確率を調べた。図2に被コメント数の累積確率分布を示す。指数分布的

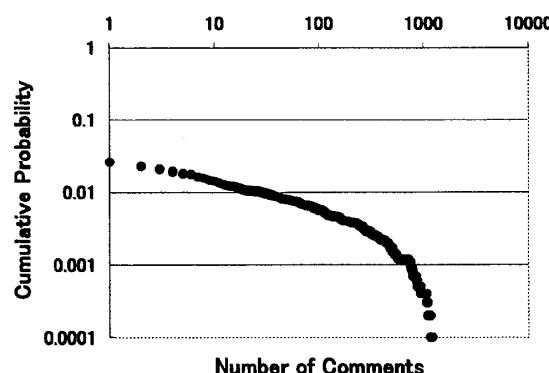


図2 被コメント数の累積確率分布

な性質が観察される。

3 相関関係分析

「もごリンク」ネットワーク構造とミニブログユーザへのコメント数との相関関係について調べた。特に、被コメント数が多いミニブログユーザを、複雑ネットワーク科学[3]において通常採用されるノード特性の指標に基づいたランキングにより同定可能かどうかについて分析した。

3.1 被コメント数に基づくランキング

まず、コメント数に基づくミニブログユーザのランキング結果が、時間的に安定かどうかについて調べた。2008年5月11日から2008年6月7日までのミニブログデータを、1週間ごとに、「5月11日から5月17日」、「5月18日から5月24日」、「5月25日から5月31日」および「6月1日から6月7日」の4つの部分期間に分割し、それぞれの部分期間におけるランキング結果と、全期間におけるランキング結果との類似性を分析した。

$A(k)$ および $B(k)$ をそれぞれ、全期間およびある部分期間におけるランキング結果の上位 k 位までのミニブログユーザの集合とする。 $A(k)$ と $B(k)$ の類似度 $F(k)$ を、 F 値を用いて、

$$F(k) = \frac{1}{k} |A(k) \cap B(k)| \quad (1)$$

で定義する。図3は、4つの部分期間に対して類似度 $F(k)$ を表示している。ここに、ダイヤ印は $F(50)$ 、四角印は $F(100)$ 、三角印は $F(150)$ をそれぞれ表している。図3より、どの部分期間に対しても上位ユーザの類似度は高いことが観察される。すなわち、コメント数に基づくミニブログユーザのランキング結果の上位は、時間的に安定と考えられる。

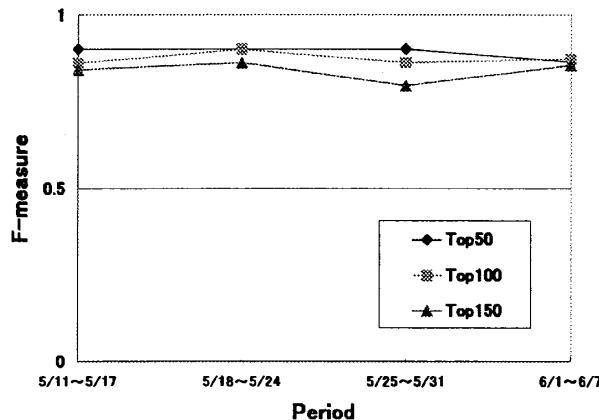


図3 被コメント数に基づくミニブログユーザのランキング結果の安定性

3.2 分析結果

ネットワークにおけるノード特性の指標として、複雑ネットワーク科学でよく用いられる、次数、PageRank値、betweenness値、クラスター係数およびcloseness値を考える[3]。「もごリンク」ネットワーク構造から決定される上記の各指標値に基づいてミニブログユーザをランクインし、被コメント数に基づくランクイン結果との類似性を調べた。異なる2つのランクイン結果の上位 k 位までの類似度 $F(k)$ は、3.1節と同様に F 値を用いて式(1)で定義する。 $50 \leq k \leq 200$ において、各指標に対する類似度 $F(k)$ の結果を図4に示す。図4より、指標として次数を用いたランクイン結果が最も類似度は高かったが、それでもそれほど大きな類似性は達成されなかったことが観察される。すなわち、複雑ネットワーク科学において通常採用される上記のような指標を用いて、被コメント数が多いミニブログユーザを同定するのは困難と考えられる。また、クラスター係数やcloseness値が高いミニブログユーザと、被コメント数が多いミニブログユーザとはほとんど関係がないと考えられる。

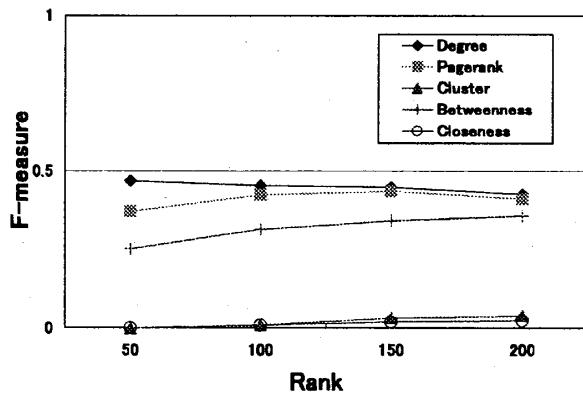


図4 被コメント数に基づくランクインとネットワーク構造に基づくランクインの類似度

ところで、NewmanとPark[4]は、社会ネットワークは、それ以外の実ネットワークとは異なり、一般に次の2つの統計的性質をもつということを観察している。それは、クラスター係数の平均値 C が、対応する configuration model [3] (ランダム

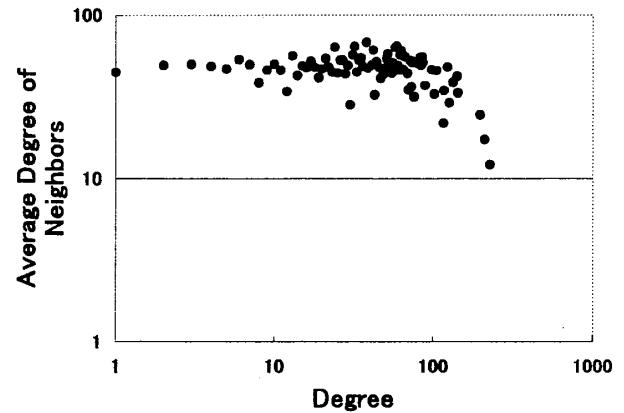


図5 「もごリンク」ネットワークの次数相関

ネットワークモデル)におけるその値に比べて非常に大きいという性質と、次数相関が正であるという性質である。「もごリンク」ネットワークにおいては、 C の値は、対応する configuration model では 0.13 であり、実測値では 0.35 であった。また、「もごリンク」ネットワークの次数相関[3]を図5に示す。これらの結果より、「もごリンク」ネットワークは、一般的な社会ネットワークとは多少異なる性質をもっていると推測される。これが、被コメント数の多いミニブログユーザを、複雑ネットワーク科学において通常採用される指標で同定が困難であった原因の一つではないかと、我々は考えている。

4 まとめ

ミニブログユーザ間のコミュニケーションを分析するために、「もごリンク」ネットワーク構造とミニブログユーザの被コメント数との相関関係について調べた。次数、PageRank値、betweenness値、クラスター係数、およびcloseness値という、複雑ネットワーク科学において通常採用されるノード特性の指標を用いては、被コメント数が多いミニブログユーザを抽出するのは困難であることを示した。特に、クラスター係数やcloseness値が高いミニブログユーザと、被コメント数が多いミニブログユーザとはほとんど関係がないことを示した。

謝辞

本研究は、科学研究費補助金基盤研究(C)(No.20500147)の補助を受けた。

参考文献

- [1] Java, A., Song, X., Finin, T., and Tseng, T.: Why we twitter: understanding microblogging usage and communities, *Proceedings of the SIGKDD Workshop on Web Mining and Social Network Analysis*, pp. 56-65 (2007).
- [2] Kimura, M., Saito, K., and Motoda, H., Minimizing the spread of contamination by blocking links in a network, *Proceedings of the 23rd AAAI Conference on Artificial Intelligence (AAAI-08)*, pp. 1175-1180 (2008).
- [3] Newman, M. E. J.: The structure and function of complex networks, *SIAM Review*, Vol. 45, pp. 167-256 (2003).
- [4] Newman, M. E. J. and Park, J.: Why social networks are different from other types of networks, *Physical Review E*, Vol. 68, 036122 (2003).