

PageRank 収束曲線を用いたノード多様性の定量化

伏見 卓恭†

斉藤 和巳†

風間 一洋‡

† 静岡県立大学大学院経営情報イノベーション研究科

‡ 和歌山大学システム工学部

1 はじめに

近年, Web のハイパーリンクネットワークや SNS 内でのユーザ関係ネットワークだけでなく, 電力網や道路網などをネットワークとして捉え, 分析する研究が盛んにおこなわれている。また, これらのネットワークにおいて各ノードは均質ではなく, 固有の立場や役割, 機能を有していると想定される。

本研究では, 現実世界で広くみられるようになったネットワークに対して, 各ノードのネットワークに対する機能に着目し, ノードの多様性を定量化することを試みる。ノードの機能的な特徴は PageRank 収束曲線のパターンとして現れ, 類似パターンのノードは同じ機能を持つと推測できる [1]。具体的には, ネットワーク構造に埋め込まれた各ノードの機能を PageRank スコアの収束曲線で表し, 収束曲線をノンパラメトリックベイズの枠組みにより加算無限個のクラスに分割することで, ノードの多様性を定量化する。すなわち, 同一の機能を持つノード群をクラスとした時, そのネットワークはいくつの機能を持つノードにより構成されているかを示す指標を提案する。評価実験では, 人工ネットワーク, Web ネットワーク, ソーシャルネットワーク, インフラネットワークなど, 様々なネットワークに適用し提案指標の有効性を確認する。

2 提案手法

ノード集合 V とリンク集合 E からなる単純無向ネットワーク $G = (V, E)$, 収束曲線の分類粒度を規定するパラメータ α を入力とし, 以下の手順で定量化する。

1. 各ステップでの PageRank スコアベクトル $\{y_1, \dots, y_S\}$ を計算;
2. 各ノードの収束曲線 $\mathbf{x}_u = (y_1(u), \dots, y_S(u))^T$ を構築;
3. ディリクレ過程に基づく K-means 法により全ノードの収束曲線を K_α 個のグループに分割;
4. ノード多様性指標 K_α を出力;

次節で, ノードの PageRank スコア収束曲線, および, クラスタリングについて説明する。

2.1 PageRank 収束曲線計算法

文献 [1] において, 各ノードの機能・役割はネットワーク構造に埋め込まれていると仮定している。ノードの機能として, ネットワーク内での階層的地位や相対的な位置, 次数や周辺ノードの次数, 周辺ノードとのつながり方などを意図しているが, これらが類似するノードどうしは, PageRank スコアの収束過程も類似すると推測できる。従って, ネットワーク構造上でのランダムウォークのモデルである PageRank [2] を用いて, 各ノードの機能を表す特徴ベクトルを計算する。

以下に, ネットワーク構造から窺い知ることのできるノードの機能を表す収束曲線の計算法を以下に示す。収束曲線は, 大域ジャンプなしの PageRank を用いて計算する。各ノードに 1 から $|V|$ までの整数値を一意に割り振る。ここで, $(u, v) \in E$ のとき $a(u, v) = 1$, それ以外 のとき $a(u, v) = 0$ とし隣接行列 $A \in \{0, 1\}^{|V| \times |V|}$ を定義する。各ノード $u \in V$ に対して, $\Gamma(u)$ をノード u の隣接ノード集合とする。すなわち, $\Gamma(u) = \{v \in V; (u, v) \in E\}$ となる。ここで, 行推移確率行列 P の各要素を $p(u, v) = a(u, v)/|\Gamma(u)|$ とする。各ノードの PageRank スコアを要素としたベクトル $\mathbf{y}$ は, $y(v) \geq 0$ で $\sum_{v \in V} y(v) = 1$ となる。初期ベクトルを $\mathbf{y}_0 = (1/|V|, \dots, 1/|V|)^T$ とし, 繰り返しステップ数 s を用い, PageRank スコアベクトル $\mathbf{y}$ は以下の更新式の極限分布として定義される:

$$\mathbf{y}_s^T = \mathbf{y}_{s-1}^T \mathbf{P} \quad (1)$$

ここで $\mathbf{b}^T$ は $\mathbf{b}$ ベクトルの転置を表わす。また, ノード u に注目すると,

$$y_s(u) = \sum_{v \in \Gamma(u)} y_{s-1}(v) \cdot p(v, u) = \sum_{v \in \Gamma(u)} \frac{y_{s-1}(v)}{|\Gamma(v)|} \quad (2)$$

で計算される。PageRank スコアの反復計算を所定の回数 S まで繰り返し, 各反復回数でのノード u の値を要素としたベクトルを $\mathbf{x}_u = (y_1(u), y_2(u), \dots, y_S(u))^T$ と定義する。このベクトル $\mathbf{x}_u$ をノード u の収束曲線と呼ぶ。各ノードの収束する値は, 各ノードの次数のみで決まるが, 一般に収束曲線は次数のみでは決まらない。周辺ノードの影響や周辺ノードとの相対的な位置関係, ネットワーク構造の影響を受ける。

Quantifying heterogeneity of nodes using PageRank convergence curves
†Takayasu FUSHIMI †Kazumi SAITO ††Kazuhiro KAZAMA
‡Graduate School of Management and Information of Innovation, University of Shizuoka

2.2 混合ディリクレ過程モデルによるクラスタリング

収束曲線群 $\{x_u, x_v, \dots\}$ および分類粒度を規定するパラメータ α を入力とし、以下の手順で計算する [3]。

1. $k \leftarrow 1, C_1 = \{x_u | u \in V\}, \mu_1 = \frac{1}{|C_1|} \sum_{u \in C_1} x_u$ と初期化；
2. 各ノードのクラスタ番号を $z_u \leftarrow 1$ と初期化；
3. z_u に変化がなくなるまで以下を繰り返す；
 - 各ノードの収束曲線 x_u に対し，
 - 距離 $d_{u,c}, c = 1, \dots, k$ を計算；
 - $\min_c d_{u,c} > \alpha$ なら， $k \leftarrow k + 1$ とし， $z_u \leftarrow k, \mu_k = x_u$ とする；
 - さもなくば， $z_u \leftarrow \arg \min_c d_{u,c}$ とする；
 - 各クラスタを $C_c = \{x_u | z_u = c\}$ のように生成；
 - 各クラスタの重心 $\mu_c = \frac{1}{|C_c|} \sum_{u \in C_c} x_u$ を計算；
4. $K_\alpha \leftarrow k$ を出力；

ここで $d_{u,c}$ は、ノード u の収束曲線 x_u とクラスタ c の重心 μ_c とのユークリッド距離を表す。当然、パラメータ α の値が小さいほど、クラスタ数 K_α は大きくなる。このクラスタ数 K_α を、ネットワークを構成するノードの多様性を表す指標として提案する。

3 評価実験

以下のネットワークを用いて提案指標の有効性を確認する。ソーシャルネットワークとして、ベンチマークとして用いられる Karate, Netscience, 論文の共著関係をリンクとした Nips, ブログのトラックバック関係をリンクとした Trackback, ウィキペディアの共起関係をリンクとした Wiki を採用する。Web ネットワークとして、大学、市役所の Web サイトをクロールして収集し無向化した Kendai, Hosei, Yamaguchi, ShizuokaW, HamamatsuW を採用する。インフラネットワークとして、静岡市、浜松市、東京 23 区、京都市の道路網 ShizuokaR, HamamatsuR, Tokyo23R, KyotoR, および、電力網 Power を採用する。人工ネットワークとして、Regular, Random, BA モデル [4] により構築した Scalefree, WS モデル [5] により構築した Smallworld, HN モデル [6] により構築した Hierarchical を採用する。

図 1 は、横軸にパラメータ α の指数 $\gamma, (\alpha = 10^{-\gamma})$ 、縦軸に提案指標値を示している。図 1 より、Web, ソーシャル, インフラネットワークの順に多様性が大きいことがわかる。実際に Web ネットワークのクラスタリング結果を見てみると、トップページ、コンテンツ一覧ページ、コンテンツページなど、その機能により比較的多めのクラスタに分類されている。一方、ソーシャルネットワークの各クラスタは、ハブノードやリーフノード、媒介ノードなど、いくつかのクラスタには分か

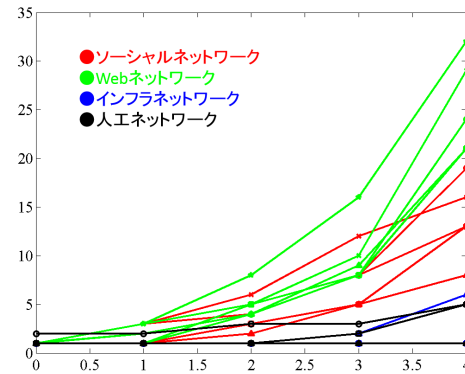


図 1: 多様性

れているが、各クラスタの区別が明確ではなく、どのクラスタの収束曲線も比較的類似傾向にあった。インフラネットワークの各クラスタでも、収束曲線に大きな違いは見られなく、比較的類似する曲線が多く存在した。ソーシャルネットワークとインフラネットワークでは、次数相関が比較的高いことが知られており、各ノードの収束過程 (PageRank 計算の過程) において PageRank スコアが緩やかに変化することが原因だと考えられる。すなわち、多様な曲線が生成されないことを意味する。道路網などでは各交差点の差異は小さいという直観に則した結果が得られたと考えられる。また、Regular, Hierarchical を除く人工ネットワークはランダム性を有しており、どのノードの機能も判別できないという意味において多様性が小さいという結果が得られた。

4 おわりに

本研究では、ネットワークにおけるノードの機能の多様性を定量化する指標を提案した。複数の種類のネットワークを用いた評価実験より、提案指標はある程度妥当な値を示すことを確認した。今後は、さらに多様なネットワークに適用可能かを検証するつもりである。

謝辞 本研究は、科研費 (No.25・10411) の補助を受けた。

参考文献

- [1] 伏見, 斉藤, 風間. ネットワーク機能コミュニティ抽出法. 日本データベース学会論文誌, Vol. 10, No. 3, pp. 13–18, 2012.
- [2] A. N. Langville and C. D. Meyer. Deeper inside pagerank. *Internet Mathematics*, Vol. 1, p. 2004, 2004.
- [3] B. Kulis and I. M. Jordan. Revisiting k-means: New algorithms via bayesian nonparametrics. *ICML*, 2012.
- [4] A. L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, Vol. 286, No. 5439, pp. 509–512, 1999.
- [5] D. J. Watts and S. H. Strogatz. Collective dynamics of ‘small-world’ networks. *Nature*, Vol. 393, No. 6684, pp. 440–442, June 1998.
- [6] E. Ravasz and A. L. Barabási. Hierarchical organization in complex networks. *Phys. Rev. E*, Vol. 67, No. 2, pp. 026112+, Feb 2003.