情報論的学習理論テクニカルレポート 2009
Technical Report on Information-Based Induction Sciences 2009 (IBIS2009)

ネットワーク科学の方法を用いたWebページネットワークの構造による分類

Classification of Web page network by the methods of network science

中川帝人* 鈴木泰博*
TEITO NAKAGAWA YASUHIRO SUZUKI

Abstract: Recently, many kinds of network structure properties have been found in network science. But many studies of network science in existence have analyzed only a single network but not many networks to compare them and the relative property of networks is not clear. So, this study aims at collecting many web-page networks and comparing and classifying them by the methods of network science with less computational complexity and meaningful feature vector as against existing methods. As a result, the web-page networks were classified into two categories. A one of them has the structure like a complete graph and not scale-free. The other has scale-free tree like structure.

Keywords: Complex Network, Link Mining, Graph Mining, SOM

1 はじめに

Watts et al.[1]、Barabasi et al.[2]らの論文以降、従来の社会ネットワーク分析も含む、ネットワーク科学の手法の発展により、様々な対象のネットワークが研究対象となってきた。しかしながら、それらの研究は単一の巨大なネットワークを各種統計的指標を用いて解析し、モデル化をするというアプローチが主流であり、現実のネットワークの構造にどのようなクラスが存在するかを調べるために、これを多数集めて比較解析した研究は少ない。

そこで、本研究ではWebページのネットワークを対象として、これをドメイン毎に複数収集することによって、多数のネットワークの収集を行う。そして、ネットワークの各種統計的指標が全てのネットワークでどのような分布を示すのか、またどのようなネットワーク構造のクラスが存在するかを調べることを目的とした。このために各種ネットワーク科学の解析法を用いて、各ネットワークの統計的特性量を抽出、特徴ベクトルを作成し、これを基に自己組織化マップを用いてクラスタリングを行い、ネットワークをクラス分類した。

2 先行研究

2.1Webページのネットワーク

Webページのネットワークとは各Webページをノード、その間に貼られたリンクをエッジとする有向グラフであるネットワークであり、Page Rank[3]等の様々な応用研究がなされている。Webページのネットワークに関しては次数分布がべき分布に従うスケールフリーネットワークであること[4]、次数対クラスタリング係数の分布がべき分布に従う階層性ネットワークであることが分かっている[5]。ただし、この性質はドメイン間リンクを含んだWebページのネットワーク構造である。本研究では一つのドメイン内に存在するホームページを一つのネットワークと考えており、これは各ホームページ管理者がどのようなリンクレイアウト方略でホームページを設計しているかを示したネットワークである。

2.2 ネットワーク構造の類別

中川、鈴木[6]は、Webページのネットワークを対象として、これをドメインごとに収拾し、各ドメインのネットワークを比較解析した。この研究では各種ネットワークの統計的指標から特徴ベクトルを作成し、これを基にクラスタ解析を行い、各ドメインのネットワークを、1. スケールフリーライクなネットワーク、2. 完全グラフ、3. 木の 3 つに類別した。しかしながら、これらはネットワークのノード数であるサイズとの相関が強く、1,2,3 の順にサイズが小さくなっていくという結果になった。中川、鈴木はこの 3 つのネットワーククラスの差異はサイズの違いに起因しているのではないかと解釈している。つまり、ネットワークが大きくなればなるほど、完全グラフに必要なエッジの数は急速に増えるため、大きなサイズのネットワークではエッジ数の少ないスケールフリーライクな構造が見られたということになる。そこで、本研究ではより本質的なWebページ

*名古屋大学情報科学研究科 複雑系科学専攻, 464-8601 名古屋市千種区不老町, tel. 052-789-4270
e-mail: nakagawa.teito@b.mbox.nagoya-u.ac.jp

のネットワークの構造の違いを知るためにサイズ数が同程度のネットワークの構造の違いを調べることを目的とした。

2.3 機械学習分野との関連

上記のようなネットワークを類別するという問題は統計的機械学習の分野においてはグラフ類別問題と呼ばれる。ただし、ネットワークとグラフはともに要素間の結合関係を表す同一の構造である。Getoor and Diehl.[7]はグラフを扱うデータマイニングであるリンクマイニング(Link Mining)に関して、これを８つのタスクに類別している。そして、その一つにグラフ類別問題(Graph Classification)を挙げている。これは入力である複数のグラフ{x1, x2, x3, …., xn}に何らかのアルゴリズムを用いてクラス変数{C1, C2, C3, … ,Ck}を割り当てる問題であり、上述したネットワークの類別問題に等しい。一般にグラフのような構造データを統計的に処理することは困難であり、このような問題に対処するために、グラフ類別問題には、対象とするグラフの特性を活かした様々なアルゴリズムが存在する。例えば、Kashima and Inokuchi.[8]はラベル付きグラフを対象として、カーネル法という学習法を用いてグラフの類別を行った。この時、グラフ上の仮想的なランダムウォークを用いてグラフ間の内積を定義している。また、Wilson et al.[9]はグラフラプラシアンに対して対称多項式を用いることにより、グラフの特徴ベクトルを抽出し、多変量解析を行う方法を提案している。しかし、これらはどれも計算量が大きく、現実的な解析を行うにはサイズが大きくても 100 程度が限界であり、またその特徴量からの意味づけの把握が困難であるという欠点を抱えている。

3 データと解析指標

3.1 データセット

今回はデータセットとして、スパム判別のコンテストである Web Spam Challenge uk-07 のデータセットを用いた。このデータセットには uk ドメインに所属するドメインのリストが含まれている。そこで、このリストから各ドメインを対象として、Web ページのネットワークを抽出した。データセットから得たドメイン名もしくはドメイン名の後に index.html を追加したもののどちらかをトップページとしてこれを中心として幅優先探索によるクローリングを行い、Web ページのネットワークをサンプリングした。ただし、この時ドメイン外へのリンクは無視した。

こうして得られたネットワークよりネットワークのサイズが 100～500 の Web ページを解析対象とすることにした。これはサイズが小さすぎると各種ネットワークの解析指標が意味を持たなくなること、サイズが大きい Web ページのネットワークが少ないために比較できるほどのサンプル数を得られないためであった。

その結果、1090 のネットワークが解析対象となった。以下の図 1 はネットワークのサイズ N とエッジ数 M のヒストグラムである。

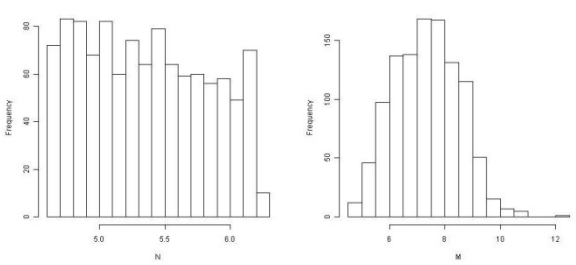

図 1：解析対象となったネットワークのサイズ数 N およびエッジ数 M のヒストグラム。双方とも対数をとっている。サイズ数が対数をとると一様分布のような分布に従うのに対して、エッジ数は対数をとると正規分布のような分布に従っている。

3.2 ネットワーク解析の統計的指標

本研究では以下のネットワークの統計的指標を用いた(表 1)。これらの統計的指標は意味づけが容易であること、多くのネットワークの統計的解析指標において計算量が低いことより選んだ。

1. 平均次数は全てのノードの次数の平均であり、個々の Web ページが持っているリンクの数を表している。サイズ数の違いを消すために理論上の最大値であるサイズ数 N-1 で割って正規化した。対数を取ると正規分布のような分布に従うために対数を取った(図 2)。
2. クラスタリング係数は各ノードのローカルなクラスタリング係数の平均であり、Web ページ同士がリンクを共有している度合いを表す。今回のデータでは、0～１まで、一様分布のようなきれいな分布に従っていることがわかった(図 2)。

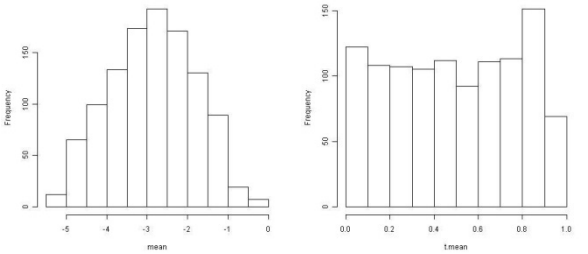

図 2:平均次数(対数)(左)とクラスタリング係数のヒストグラム(右)。平均次数は正規分布のような分布に従っており、クラスタリング係数は一様に分布している。

3. 平均最短経路長は任意の２ノード間のホップ数の平均であり、すべてのコンテンツから別のコンテンツへのアクセスに必要なリンク数を表す。ほとんどのネットワークでは 1～3 の間に収まった(図 3)。

4. コミュニティ数は Newman-fast 法[10]により、モジュラリティ Q を最大にするノードの密な集合を取り出した際の集合の個数であり、これをノード数で割ることによって正規化した。この指標は Web ページの分割度合を表している。ほとんどのネットワークでノード数の 2～3%の範囲内に収まっており 2、3 のコミュニティしか存在しないということが明らかになった(図 3)。

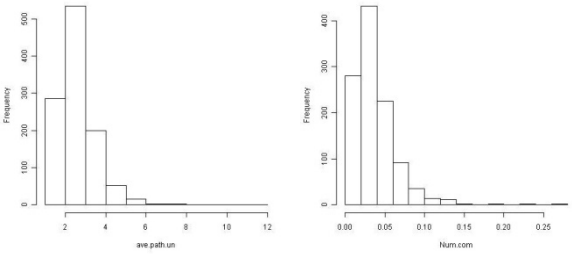

図 3:平均最短経路長(左)、コミュニティ数(右)のヒストグラム、平均最短経路長は大部分が 2，3 の範囲で分布していること、コミュニティ数はノード数に対して 5%以内の数で分布していることがわかる。

5. 次数相関はエッジのノード対の相関係数である。次数分布が正規分布に従わないことよりスピアマン相関を用いた。これは高ければ高いほどハブ同士、次数の低いノード同士がつながっていることを意味していることより、ハブページから次数の低い末端のコンテンツページへのアクセス可能性を意味していると考えられる。-0.3 を中心として、正規分布のよう分布に従っていることがわかり、平均として次数相関は負であることがわかった(図 4)。
6. リーフとは次数が 1 であるノードであり、その比率を表したものがリーフの比率で、これを理論上の最大値であるノード数－1 で割って正規化した。次数が 1 のノードは jpeg や mpeg などの動画像ファイル等、末端のコンテンツを表していると考えられる。得られたデータよりリーフの数が 1 割にも満たないページが全体の 3 分の 1 を占めているがそれ以上では一様分布のように一定の比率に従った(図 4)。

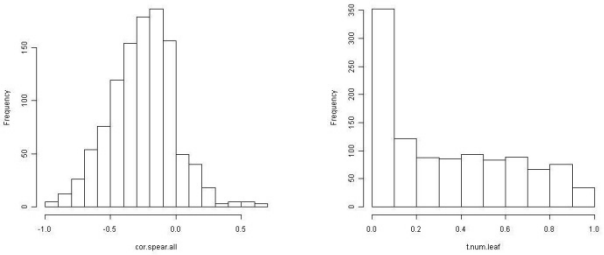

図 4:次数相関(左)とリーフの比率(右)のヒストグラム。次数相関は-.3 を頂点として、正規分布のように分布している。リーフの比率は 1 割にも満たないものが 3 分の 1 を占め、残りで一様に分布している。

表 1：本研究で用いたネットワークの統計的指標

解析指標	グラフ理論的な意味	Web ページにおける意味
1.平均次数 (対数)	エッジの多さ	各ページが持っているリンクの数
2.クラスタリング係数	ローカルな結合性	リンクの共有の度合
3.平均最短経路長 (無向グラフとしての)	グローバルな結合性	コンテンツのアクセスに必要なリンクの数
4.コミュニティ数	うまく分割できるノードの集合の数	ページの別れ度合
5.次数相関	ノードの次数のスピアマン相関	ハブからの末端のコンテンツへのアクセス可能性など
6.リーフの比率	次数が 1 のノードの比率	末端のコンテンツファイルの比率

4 ネットワーク構造の類別問題

ネットワークを構造によって類別するために、ネットワークの統計的特性量を特徴ベクトルとして、これをユニット数 15×15 の自己組織化マップを用いて類別した。以下の図 5 は各ユニットに所属する個体数をプロットしたものである。これより、左下の 1-15 ならびに 15-1 を中心としてネットワークを大きく二つに類別できる。

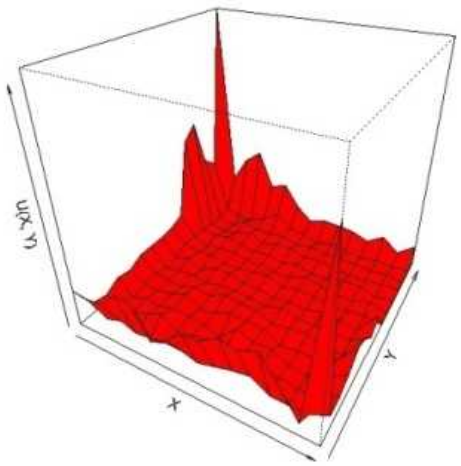

図 5：自己組織化マップの各ユニットの個体数。高さが個体数を

表す。1－15 と 15－1 を中心として大きく 2 つに分かれている。

これらのユニットに所属する特徴ベクトルの平均は以下のようになっている。これより、1-15(n=68)のユニットに関しては平均次数が高く、クラスタリング係数が高く、平均最短経路長が短く、リーフの数が少ないという結果が見られた。15-1(n=85)のユニットはそれに対して、平均次数が低く、クラスタリング係数が低く、平均最短経路長が短く、リーフの数が多いという結果が見られた。また、それぞれのユニットに所属するネットワークの次数分布を調べてみたところ、1－15 はべき分布に従わないものが多く、15－1 はべき分布に従うものが多数見られた。

これらの結果より、1－15 に属するネットワークはより完全グラフに近い、スケールフリー性を持たないネットワークであり、15－1 に属するネットワークは木構造に近いスケールフリーな構造を持っていると考えられる。

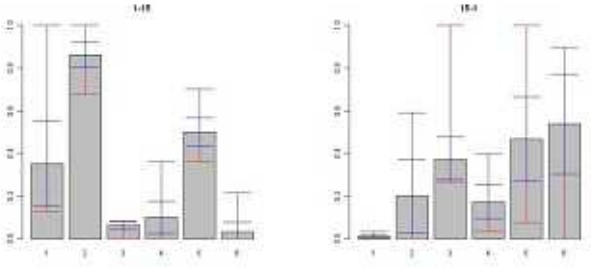

図 6：ユニット 1-15(左)とユニット 15-1(右)の特徴ベクトル。各統計量に関して、1090 のサンプル内の最大値を 1、最小値が 0 となるように正規化している。エラーバーはそれぞれユニット内の最小値から最大値の範囲(赤)、標準偏差(青)を表している。

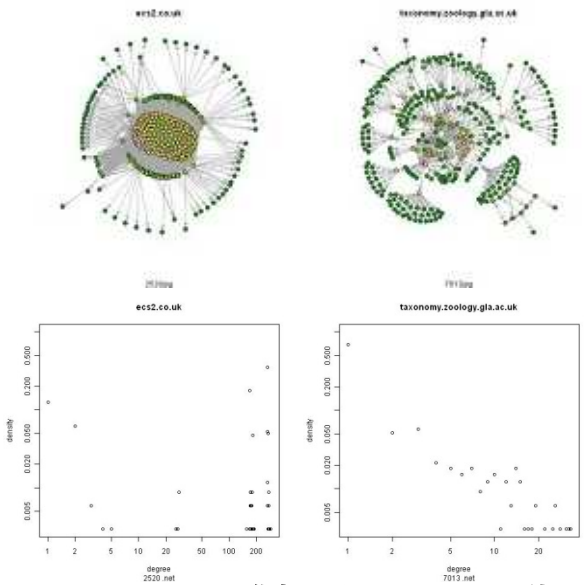

図 7： それぞれのユニットを代表するネットワークの KK 法[11]による可視化図(上)と両対数軸による次数分布(下)。左の図が 1－15 を右の図が 15－1 に典型的なネットワークである。

5 最後に

本研究では同程度のサイズ数の Web ページのネットワークを対象としてネットワーク科学の方法による特徴ベクトルを用いてネットワークの構造として 2 つの異なるクラスを発見することが出来た。今後はこれらのネットワークとしての構造的な差異がホームページの内容の差異にどのように影響を与えているのか詳細に調べたいと考えている。

実際に筆者による目視では 1－15 に属するネットワークのホームページはデザインが優れており、企業や組織などのホームページであることが多く、15－1 にぞくするネットワークのホームページであることが多かった。今後このような違いをより客観的な指標で表すためにラベル付きデータやアンケートによる調査を実施したいと考えている。

参考文献

[1] D.J.Watts.and S.H.Strogatz.: Collective dynamics of ‘small-world’ networks, Nature, vol.393, pp. 440-442, (1998)

[2] A.L.Barabasi. and R.Albert.: Emergence of Scaling in Random Networks, Science, vol. 286, pp. 509-512, (1999)

[3] L Page, S Brin, R Motwani, and T Winograd.:*The PageRank Citation Ranking: Bringing Order to the Web*, Technical Report. Stanford InfoLab. (1999)

[4] Reka Albert, Hawoong Jeong. And Albert-Lazlo Barabasi.:Diameter of the World-Wide Web, Nature vol.401, pp.130, (1999)

[5] Eezsebet Ravasz and Albert-Lazlo Barabasi.:Hierarchical organization in complex networks, Physical Review, vol. E67, 026112, (2003)

[6] **中川帝人、鈴木泰博**: **複雑ネットワークの方法を用いたネットワークの構造による分類**, **第 5 回ネットワーク生態学シンポジウム**, (2009)

[7] Lise Getoor. and Christopher P. Diehl.:Link Mining: A Survey, ACM SIGKDD Explorations Newsletter, vol.7, 2, pp. 3-12, (2005)

[8] Hisashi Kashima. and Akihiro Inokuchi.:Kernels for Graph Classification, In ICDM Workshop on Active Mining, (2002)

[9] Wilson, R.C., Hancock, E.R. and Bin Luo:Pattern vectors from algebraic graph theory, Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 27, 7, pp. 1112-1124

[10] Aaron Clauset, M. E. J. Newman, and Crisopher Moore:Finding Community Structure in Very Large Networks, Physical Review, vol.E70, 066111, (2004)

[11] Tomihisa Kamada and Satoru Kawai. :An algorithm for Drawing General Undirected Graphs, Information Process Letters 31, 7-15, (1989)