# 大学サイトモチーフの類似度ネットワークによる分析法

白賀　昌宏† 　伏見　卓恭†† 　大久保誠也†,†† 　武藤　伸明†,†† 　斉藤　和巳†,††

† 静岡県立大学経営情報学部経営情報学科　〒 422–8526 静岡県静岡市駿河区谷田 52-1
†† 静岡県立大学経営情報学研究科　〒 422–8526 静岡県静岡市駿河区谷田 52-1
E-mail: †{b08059,j09118,s-okubo,muto,k-saito}@u-shizuoka-ken.ac.jp

**あらまし**　大学ウェブサイトのハイパーリンク構造を題材に類似度ネットワークによるモチーフ分析法を探求する．モチーフ検出のための $Z$score の類似度に基づき，各大学ウェブサイトをノードとしたネットワークを構築する．ネットワーク構成法として，類似度の高いサイトペア順に単一連結成分となるまでリンクを付与する方法（単純類似ネットワーク），サイト毎に最類似の $k$-サイトに対しリンクを付与する方法（$k$-近傍ネットワーク），及び，任意のサイトペアに対し他に両者の類似サイトが存在しない場合にリンク付与する方法（相対近傍ネットワーク）を比較する．計算機実験により，相対近傍ネットワークによる分析法が有望であることを示す．

**キーワード**　ネットワークモチーフ，相対近傍ネットワーク，$k$-近傍ネットワーク，単純類似ネットワーク

# Analysis method by similarity networks of university sites motifs

Masahiro SHIRAGA†, Takayasu FUSHIMI††, Seiya OKUBO†,††, Nobuaki MUTO†,††, and Kazumi SAITO†,††

† School of Management and Information, University of Shizuoka
52-1 Yada, Suruga-ku, Shizuoka, 422–8526 Japan
†† Graduate School of Management and Information, University of Shizuoka
52-1 Yada, Suruga-ku, Shizuoka, 422–8526 Japan
E-mail: †{b08059,j09118,s-okubo,muto,k-saito}@u-shizuoka-ken.ac.jp

**Abstract**　By using hyperlink structure of the university websites as our subject matter, we explore motifs analysis methods by constructing similarity networks based on $Z$scores. Namely, by regarding a university website as a node, we compare the network composition methods based on the similarity of $Z$score vectors. In this paper, we focus on the three network construction methods: the method of giving links until becoming a single connected component in order of similarity of site pairs (Nearest Neighborhood network), the method of giving the links to the most similar $k$-site from each site ($k$-Nearest Neighbor network), and the method of giving the links only when each site pair is the relatively neighbour in comparison to the other sites (Relative Neighborhood Graph network). In our experiments, we show that the analysis method based on the Relative Neighborhood Graph is promising.

**Key words**　network motifs, Relative Neighborhood Graph network, $k$-NN network, NN network

## 1. はじめに

近年，大規模な複雑ネットワークの研究が盛んに行われている．人間関係ネットワークや WWW ネットワークにおいて，重要ノードを抽出するための手法は広く知られている．一方で，大規模ネットワークの多くはスケールフリー性 [1] やスモールワールド性 [7] という特徴的な構造を持つことが明らかとなり注目を集めている．しかし，実ネットワークは大規模かつ複雑な構造を有するため，その構造を完全に解明することは困難である．そこで，ネットワークの構造を特徴付けるものとして，ネットワークモチーフが提案されている [2]．ネットワークモチーフは，複雑ネットワークの基本構成要素であり，複雑ネットワークの機能を理解することを目的に提案されたものである．ネットワークの特徴的なモチーフパターンを知ることにより，ネットワーク内で頻出するノード間の相互関係パターンやネットワーク上での現象への影響について，知見を得ることができる [3]〜[5]．

本稿では，大学ウェブサイト群のような複数のネットワークに対するモチーフパターン分析を対象にする．これらのモチーフパターンを分析すれば，類似特徴を有するネットワークや，

異なった特徴を有するネットワークが存在すると自然に想定できる．したがって，特に，ある程度以上のネットワーク群を分析対象とするケースでは，このようなモチーフパターンの類似関係を系統的に調べるための方法論が必要になる．そのため，本研究では各ネットワークをノードと見なし，モチーフパターンの類似度に基づきリンクを付与する，メタレベルでの類似度ネットワークによる分析方法論について探究する．このような類似度ネットワークの構成法としては，類似度の高いサイトペア順に単一結合成分となるまでリンクを付与する方法（単純類似ネットワーク），サイト毎に最類似の $k$-サイトに対しリンクを付与する方法（$k$-近傍ネットワーク），及び，任意のサイトペアに対し他に両者の類似サイトが存在しない場合にリンク付与する方法（相対近傍ネットワーク）[6] などが考えられる．

具体的には，大学ウェブサイトのデータセットを対象に，ネットワークモチーフ検出のために $Z$score による類似度を算出し，単純類似ネットワーク，$k$-近傍ネットワーク，及び，相対近傍ネットワークの 3 手法により類似度ネットワークを構築する．3 手法によって構築された類似度ネットワーク分析法を比較し，ネットワークモチーフの類似度によるウェブサイトの分析法の提案を試みる．

本論文の構成は以下の通りである．2 章でネットワークモチーフについて触れ，3 章で類似度に基づくネットワーク構築法について説明する．4 章で大学ウェブサイトデータを用いた評価実験を行い，最後に 5 章で本研究のまとめを述べる．

## 2. 諸　定　義

本章ではネットワークモチーフの基本的概念を述べる．また，ランダムネットワークの作成法および $Z$score 計算法についても説明する．

### 2.1 ネットワークモチーフ

ネットワークモチーフ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は，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が条件である．そして，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 パターンとなることが示されている [2]．

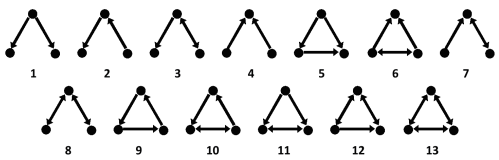


図 1　ネットワークモチーフ

### 2.2 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法を示す．

- 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 $Z$score

$Z$score 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 $Z$score 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 $Z$score は次式で定義される．

$$Z_i = \frac{x_i - \mu_i}{\sigma_i}. \tag{1}$$

ここで，1,000 個のランダムネットワークを用いる場合，$\mu_i = \frac{1}{1000}\sum_j^{1000} y_i^{(j)}$，$\sigma_i = \frac{1}{1000}\sum_j^{1000}(y_i^{(j)} - \mu_i)^2$ である．

$Z$score 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 $Z$score 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．

## 3. 類似度ネットワーク構築法

各オブジェクト（本論文では各大学サイト）ペアに対して，任意の類似度が定義できたと仮定する．本章ではオブジェクトをノードとし，類似度の高いノードペア間にリンクを張りネットワークを構築する手法を説明する．

### 3.1 単純類似法

単純類似法（以後 NN 法と呼ぶ）は，類似度の高いノードペアから順にリンクを張り，全ノードが 1 つの連結成分になるまで繰り返す．詳細な手順を以下に示す：

(i)　全ノードペア間の類似度を計算する
(ii)　類似度で降順に並び替える
(iii)　類似度順位が高いノードペア順にリンクを張る
(iv)　全ノードが 1 つの連結成分になるまで (iii) を繰り返す

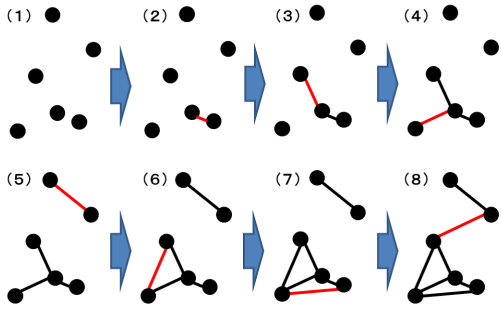


図 2　NN 法によるネットワーク構築

NN 法によるネットワーク構築のモデル図を図 2 に示す．図 2 では，ノード間の類似度を 2 次元のユークリッド距離で表している．図 2 中の（1）の 6 つのノードに対して，類似度の最

も高いノードペア間にリンクを張る（2）．次に類似度の高いノードペア間にリンクを張る（3）．この処理を全ノードが 1 つの連結成分になるまで繰り返す．NN 法により構築したネットワークを NN ネットワークと呼ぶ．

### 3.2 $k$-近傍法

$k$-近傍法（以後 $k$-NN 法と呼ぶ）は，パターン認識の分野でしばしば用いられる単純な機械学習アルゴリズムである．詳細な手順を以下に示す．

(i) 全ノードペア間の類似度を計算する

(ii) 各ノードに対して，自身との類似度で降順に相手ノードを並び替える

(iii) $k \leftarrow 1$ とする

(iv) 各ノードは，自身との類似度が $k$ 位のノードとリンクを結ぶ

(v) 全ノードが 1 つの連結成分になるまで $k \leftarrow k+1$ とし，(iv) を繰り返す

$k$-NN 法によるネットワーク構築のモデル図を図 3 に示す．図 3 では，$k=3$ とし，ノード間の類似度を 2 次元のユークリッド距離で表している．図 3 中の赤いノードはそれぞれ，自身の最近傍である 3 つのノードにリンクを張っている．この処理をすべてのノードに対して行う．$k$-NN 法により構築したネットワークを $k$-NN ネットワークと呼ぶ．

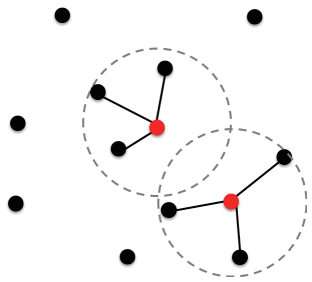

図 3 $k$-NN 法によるネットワーク構築

### 3.3 相対近傍法

相対近傍法（以後 RNG 法と呼ぶ）は，任意のノードペアに対し，他に両ノードの類似ノードが存在しない場合にリンク付与する．ノード集合を $\{p_1, p_2, \cdots, p_N\}$ とする．任意のノード間の距離として類似度から算出されたユークリッド距離 $d(p_i, pj)$ を採用する．詳細な手順を以下に示す．

(i) 任意のノードペア $\{p_i, p_j\}$ について，ノード間の距離を半径とする超球を，ノード $p_i, pj$ のそれぞれを中心として描く

(その重なった部分を Lune と呼ぶ)

(ii) Lune にその他のノードが存在しない場合に（図 5），ノード $p_i, p_j$ 間にリンクを張る

(iii) （ii）を全ノードペアに対して行う

これは，2 つのオブジェクト $p_i$ と $p_j$ が相対近傍であること，すなわち，全ての $p_k \notin \{p_i, p_j\}$ について次式が成立するときリンクが生成される．

$$d(p_i, p_j) \leqq \max_{k \notin \{i,\, j\}} \{d(p_i, p_k), (p_j, p_k)\} \qquad (2)$$

図 4 と 5 に，RNG 法において 2 つのノード $p_i$ と $p_j$ の間にリンクが生成されない場合とされる場合の例をそれぞれ示す．図 4,5 では，ノード間の類似度を 2 次元のユークリッド距離で表している．

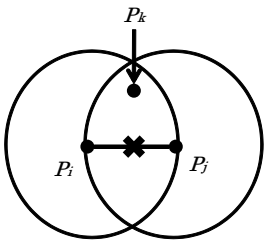


図 4 リンクを結ばない例

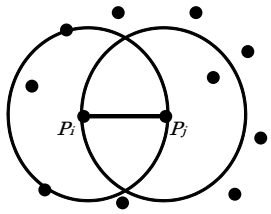


図 5 リンクを結ぶ例

RNG 法により構築したネットワークを RNG ネットワークと呼ぶ．

## 4. 評価実験

本研究では，日本の国立大学 85 大学のウェブサイトを対象に評価実験を行う．

### 4.1 実験方法

全国立大学 85 大学(注1)のウェブサイト内のページを取得し，各ウェブサイトのハイパーリンク構造からウェブサイトネットワークを構築する．図 6 に，実験に用いたウェブサイト群のノード数とリンク数の散布図を示す．全ウェブサイトの平均ノード数は 3,111.8 ，平均リンク数は 47,891.07 である．

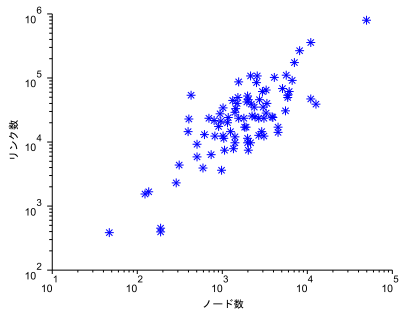


図 6 85 大学のノード数とリンク数

ウェブサイトネットワークは，このように大規模・複雑であるため，各ウェブサイトネットワークの $Z$score を見ているだけでは各大学のウェブサイトネットワーク間の関係を把握することは困難である．そこで隠れた構造や関係を明らかにするために，以下の手順で分析することを試みる．

(i) モチーフ頻度の $Z$score を計算する (2 章)

(ii) $Z$score ベクトルの類似度に基づいて 3 手法により類似度ネットワークを構築 (3 章)

(iii) 大学サイトの関係を分析する．

(注1)：全 86 大学中，GNU の Wget によって収集できない 1 大学を除き 85 大学

すなわち，各大学のウェブサイトネットワークを 1 つのノードとし，その特徴量 $Z$score ベクトルの類似度によりリンクを張り 3 種の類似度ネットワークを構築する．

### 4.2 実験結果と考察

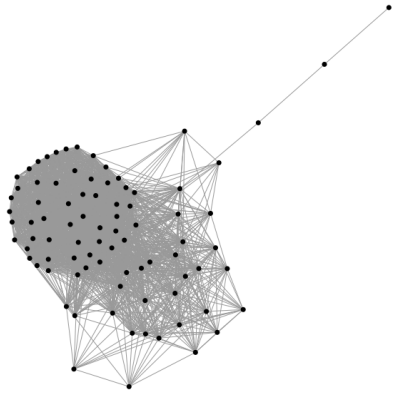

図 7　NN ネットワーク

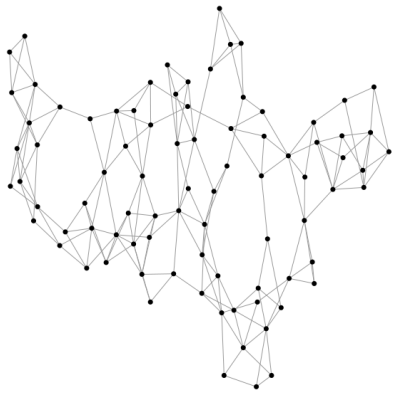

図 8　3-NN ネットワーク

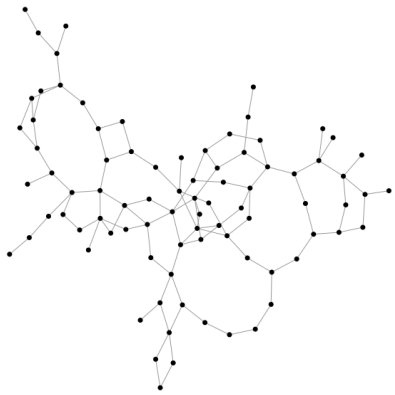

図 9　RNG ネットワーク

図 7〜図 9 に 3 種の類似度ネットワークの可視化結果を示す．3 手法の違いを定量的に把握するために，表 1 に各種類似度ネットワークの基本統計量を示す．なお，$k$-NN ネットワークは $k = 3$ で全ノードが 1 つの連結成分になったため，3-NN ネットワークと呼ぶことにする．表 1 より，平均次数とクラスター係数では NN ネットワークが最も高く 3-NN ネットワーク，RNG ネットワークの順に小さくなっていることが分かる．また，平均ノード間距離とグラフ直径を比較すると平均次数などとは逆に RNG ネットワークが最も高く 3-NN ネットワーク，NN ネットワークの順に小さくなっている．

3 手法それぞれの可視化結果を比較する．NN ネットワークでは端に位置しているノードを識別することができるが，それ以外のノードは密結合していて視覚的にとらえるのが難しい．1 つの連結成分になるまでリンク付与の処理を繰り返すため，他と比べてリンク密度が非常に高くなっている．3-NN ネットワークを見ると各ノードに対して類似度の高いノード同士で比較的密な構造になっている．すべてのノードが最低 3 本のリンクを有しているため，NN ネットワークと異なり外れ値を検出することができない．さらにネットワークのいたるところがほぼ同様な密度であるため，ネットワークの構造を把握することが困難である．また，ネットワークの全体構造として丸みを帯びていることが見て取れる．一方，RNG ネットワークの構造に着目すると，類似度の高いノード同士は比較的密に連結しているが，類似度の小さいノード間では疎なネットワーク構造しか作れないため，リンクの粗密により大域的にデータを分類できる．また，類似度の低いノード同士は辺を結ばないためノード間の関係を簡略化できる．

比較の結果，NN ネットワークの外れ値を検出できる長所と 3-NN ネットワークの全体的にリンクが疎である点をあわせもつ RNG ネットワークは，大規模なウェブサイトネットワーク間の特徴をとらえるために有望な分析法であることが示唆された．

表 1　基本統計量による比較

| データ名 | 平均次数 | 平均クラスター係数 | 平均ノード間距離 | 直径 |
|---|---|---|---|---|
| **NN** | 49.035 | 8.248813e-01 | 1.532493e+00 | 5 |
| **3-NN** | 4.494 | 4.381793e-01 | 5.371989e+00 | 13 |
| **RNG** | 2.565 | 0.000000e+00 | 6.930812e+00 | 19 |

## 5. おわりに

本論文では，大学ウェブサイトのハイパーリンク構造を題材に類似度ネットワークによるモチーフ分析法を探求するために，3 手法による類似度ネットワークを構築し比較した．計算機実験より，RNG 法による類似度ネットワークは，他の手法の類似度ネットワークよりも視覚的に特徴をとらえやすいことを示した．今後は，さらに多様なウェブサイトデータを対象に本論文の分析法を適用し，その有望性を評価したい．

**謝辞** 本研究は，科学研究費補助金基盤研究 (C) (No. 22500133) の補助を受けた．

## 文　献

[1] A.-L. Barabasi and R. Albert, “Emergence of scaling in random networks.” Science 286, pp. 509-512, 1999.

[2] 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.

[3] 高田寛喜，斉藤和巳，上田修功，“時系列情報を考慮したモチーフパターン変化の分析”，電子情報通信学会第 7 回 Web インテリジェンスとインタラクション研究会， 2006.

[4] 島田諭，小出明弘，斉藤和巳，佐藤哲司，“QA2 部グラフにおけるモチーフを用いたコミュニティの経時的変化に関する分析”，知識共有コミュニティワークショップ (KSCWS)， 2010.

[5] 小出明弘，斉藤和巳，佐藤哲司，“モチーフによる QA 二部グラフの構造分析”，Web とデータベースに関するフォーラム (WebDB Forum)， 2010.

[6] J. K. Supowit, “The relative neighborhood graph, with an application to minimum spanning trees.” Journal of the ACM (JACM) 30, pp. 428-448, 1983.

[7] D. J. Watts and S. H. Strogatz, “Collective dynamics of ‘ small-world ’ networks.” Nature 393, pp. 440-442, 1998.