F-056

ネットワーク構造による類似探索性能の分析法の提案
Analyzing performance of similarity search depending on network structure

外岡 達也† 小出 明弘† 斉藤 和巳†
Tatsuya Tonooka Akihiro Koide Kazumi Saito
青山 一生‡ 澤田 宏‡ 上田 修功‡
Kazuo Aoyama Hiroshi Sawada Naonori Ueda

1．はじめに

スモールワールド性は，友人関係ネットワーク上でのメッセージ配信に関する Milgram の実験[4]や，その構成法の一つを数理モデルとして定式化した Watts-Strogatz の研究[1]を契機に，多様な現実のネットワークに共通する性質として注目されている．その代表的な特徴の一つは，ネットワーク上のリンクを辿り，どのノードからも別のノード（ターゲット）へ，比較的短いステップ，概して 6（six degrees of separation）程度で到達するパスが存在する点である.ただし，Watts-Dodds-Newman, および Kleinberg のその後の研究[2,3]では，スモールワールド性のもう一つの本質的な特徴を指摘している．すなわち，Milgram の実験でも見られるように，ネットワーク上でのターゲットの位置を明確に知らない各ノードが，その近傍の局所情報のみを利用するだけで，ほぼ最短に近いパスを見つけ出し，効率の良いメッセージ配信（ルーティング）を実現できる点である．

Watts-Dodds-Newman は，想定する階層的人間関係ネットワークに対して，人物間の類似度に基づく社会距離（厳密には三角不等式を満たさない非類似度）の導入で，上述したような局所ルーティング機能が実現できることを示した．一方，Kleinberg は，ユークリッド空間において格子状に配置したノードの最近傍を結合したネットワークに対して，ある確率モデルに基づきリンク群を付与すれば，効率の良いルーティングが実現できることを理論的に示した．しかしながら，これらの研究では，理論的な実現可能性の追求に主眼が置かれ，多様な実問題を対象とした応用でのスモールワールド性の有効利用について触れられず，応用可能性の追求は重要な研究課題として残されている．

本研究では，工学的な視点に基づき，メッセージ配信をオブジェクトの類似探索問題と捉え，効率の良い探索を実現するネットワーク構成法に向けて，探索におけるネットワークの性質を分析する方法論を提案する．詳細には，あるオブジェクトペアの距離と比較して，双方のオブジェクトに近い第三のオブジェクトが存在しない場合、前記オブジェクトペアを結合させて構成する RNG ネットワーク（Relative neighborhood graph）[5]と，オブジェクト毎に k 個の最類似オブジェクトを結合させて構成する k-NN (Nearest Neighbour) ネットワークの性質を比較する．実験では，新聞記事データを用いた分析結果を報告する．

†静岡県立大学 University of Shizuoka
‡NTT コミュニケーション科学基礎研究所

2．問題設定と解法

2．1 アトラクタ数による分析法

いま，探索対象のオブジェクト集合を X とし，オブジェクトペア x, y ∈ X に対する類似度 s(x, y) とする．また，類似度に基づきオブジェクトを結合させて構成した探索用ネットワークを G とし，クエリとして与えられるオブジェクトを q とする．オブジェクト x の（無向）ネットワーク上での隣接オブジェクト集合を Γ(x) で表す．これらを構成要素とする 4 つ組 (X, s(), G, q) において，オブジェクト x がアトラクタとは，以下の条件を満たす集合と定義する．

$$A(q) = \{x : s(x,q) \geq \max_{y \in \Gamma(x)} \{s(y,q)\}\}. \quad (1)$$

一般に，アトラクタの個数が多ければ，探索方法やその終了条件によらずに，ネットワーク上での探索は難しいと想定できる．明らかに、少ないリンク数でアトラクタ数も少ないネットワークの構築が望まれる．一方，リンク数が同程度なら，アトラクタ数の少ない（ユニモーダルに近い）ネットワークが一つの指標として望ましいと考えられる．

2．2 ベイスンのサイズによる分析法

いま，オブジェクト u はアトラクタでないとする．このとき次に探索するオブジェクトを以下で求まるオブジェクトとするのは，最急勾配法の視点からも極めて自然である．

$$v = \arg\max_{y \in \Gamma(u)} \{s(y,q)\}. \quad (2)$$

このような (u, v) の有向リンク集合のみからなるサブグラフはネットワークの部分木 T_F となる．有向リンク (u, v) の向きを (v, u) と反転させた有向部分木を T_B とする．ここで，(X, s(), G, q) において，アトラクタ x のベイスンとは，有向部分木 T_B 上で，アトラクタ x から到達可能なオブジェクト集合 B(x) と定義する．

一般に，最良アトラクタ（最類似オブジェクト）x*の持つベイスンが広い（B(x*) に含まれるオブジェクト数が多い）場合には，比較的に探索は容易と言える．逆にベイスンが狭ければ，一般に，探索方法やその終了条件によらずに，ネットワーク上での探索は難しいと想定できる．アトラクタ数と同様，少ないリンク数でベイスンの広いネットワークの構築が望まれる．一方，リンク数が同程度なら，ベイスンの広い（ユニモーダルに近い）ネットワークが一つの指標として望ましいと言える．

2. 2 可到達アトラクタによる分析法

クエリ集合 $\{q_1, ..., q_N\}$ に対して，それぞれの最類似オブジェクト集合 $\{x_1^*, ..., x_N^*\}$ を求めれば，各オブジェクト u に対して，以下のような可到達アトラクタ集合を定義することができる．

$$R(u) = \{x_i^* : u \in B(x_i^*)\}. \quad (3)$$

明らかに，R(u) のサイズが大きいオブジェクトは，最良アトラクタ（最類似オブジェクト）のベイスンの共通部分にあり，探索を開始するオブジェクトとして望ましい性質を持つと言える．つまり，アトラクタを求めるまでの探索ならば，R(u) の要素数の多いオブジェクトを探索開始点とした方が，期待値として最良解の求まる回数が多くなる．さらに，R(u) の要素数の多いオブジェクトがネットワーク上に多数存在すれば，探索開始オブジェクトの選定が一般に容易になる．

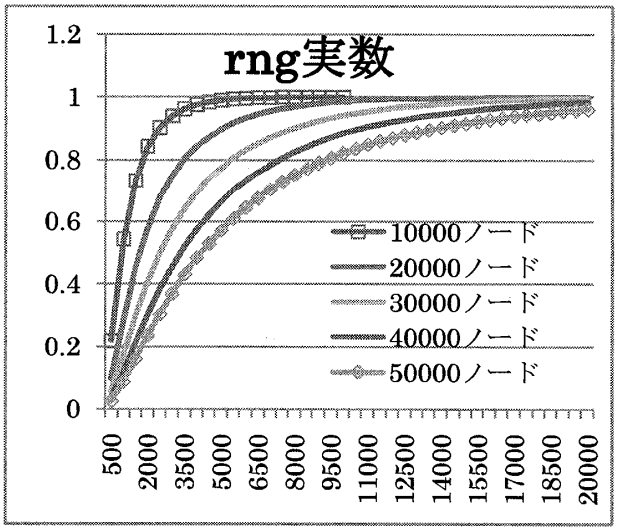

図 1:RNG におけるアトラクタ数(実数)

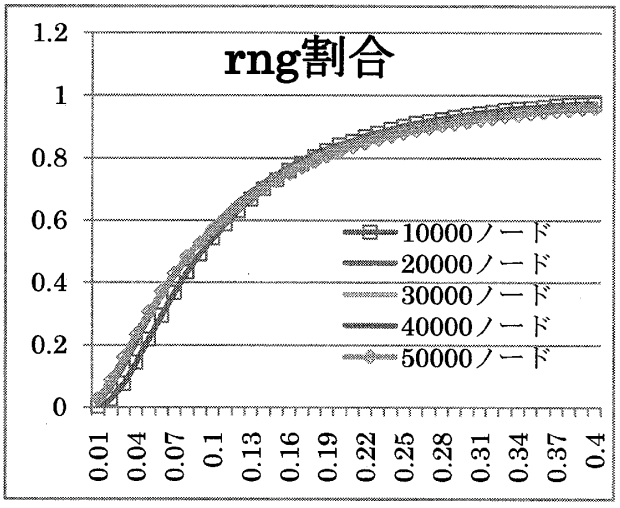

図 2:RNG におけるアトラクタ数(割合)

3. 実験による評価

評価データとして、1992 年 1 月~2001 年 12 月までの 10 年間の毎日新聞国際面記事を用いる[6]. RNG ネットワークと k-NN ネットワークの平均次数を同程度の値にするため，k-NN ネットワークの k の値を 4 とした．下記での分析ではオブジェクト数を一万単位で変化させた．

3. 1 アトラクタ数の分析結果

基本問題設定として，評価データの各オブジェクトをクエリ $q \in X$ とし，各クエリに対するアトラクタ数の分析を行った．図 1 と図 2 にはそれぞれ式(4)と式(5)で定義する RNG ネットワークのアトラクタに関する累積分布を示す．同様に図 3 と図 4 にはそれぞれ式(4)と式(5)で定義する k-NN ネットワークのアトラクタに関する累積分布を示す．

$$P(a) = |\{q : |A(q)| \le a\}| / |X|. \quad (4)$$

$$P(a) = |\{q : |A(q)| / |X| \le a\}| / |X|. \quad (5)$$

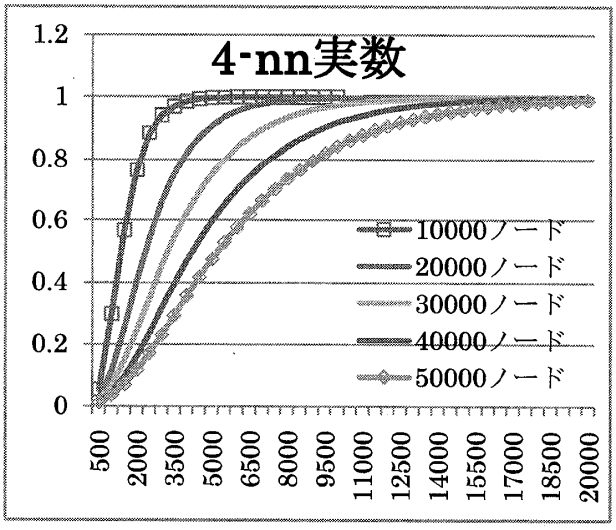

図 3:k-NN におけるアトラクタ数(実数)

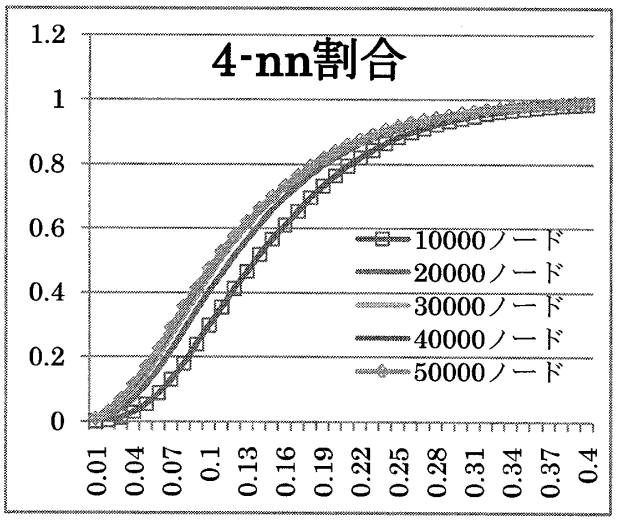

図 4:k-NN におけるアトラクタ数(割合)

これらの図から，オブジェクト数に対するアトラクタ数の割合はネットワーク構成手法に寄らず，オブジェクト数が大規模になってもほぼ同程度の値を示すことが分かる．

3．2　ベイスンサイズの分析結果

上記と同じ問題設定で，評価データの各オブジェクトをクエリ $q \in X$ とするので，各クエリがアトラクタとなるため，そのベイスンサイズの分析を行った．図 5 と図 6 にはそれぞれ式(6)と式(7)で定義する RNG ネットワークのベイスンサイズに関する累積分布を示す．同様に，図 7 と図 8 にそれぞれ式(6)と式(7)で定義する k-NN ネットワークのベイスンサイズに関する累積分布を示す．

$$P(b) = |\{q : |B(q)| \le b\}| / |X|. \quad (6)$$

$$P(b) = |\{q : |B(q)| / |X| \le b\}| / |X|. \quad (7)$$

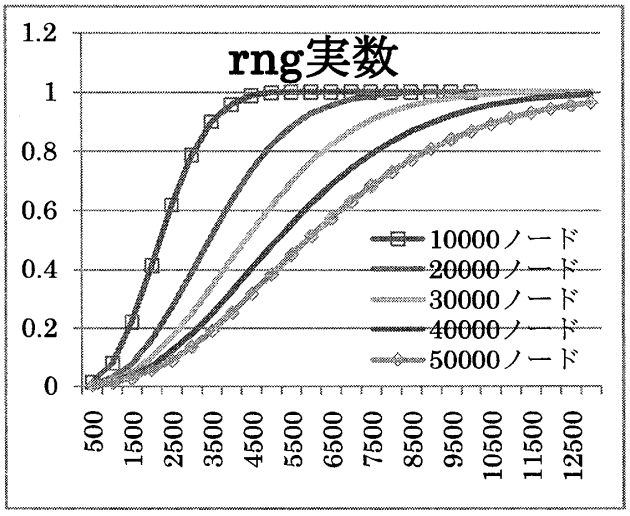

図 5:RNG におけるベイスンサイズ(実数)

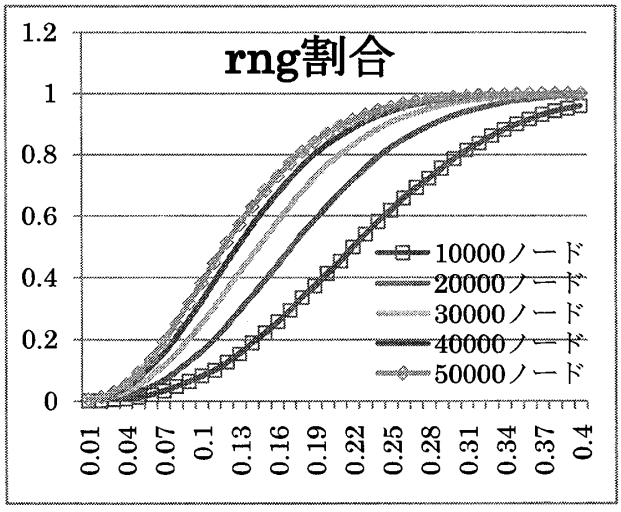

図 6:RNG におけるベイスンサイズ(割合)

図から，ベイスンサイズはオブジェクト数の増加につれ相対的に小さくなることが分かる．RNG のベイスンサイズは k-NN のそれよりも大きくなる傾向にあることが分かる．

3．3　可到達アトラクタの分析結果

上記と同じ問題設定で，可到達アトラクタ数の分析を行った．図 9 と図 10 にはそれぞれ式(8)と式(9)で定義する RNG ネットワークの可到達アトラクタ数に関する累積分布を示す．同様に，図 11 と図 12 にはそれぞれ式(8)と式(9)で定義する k-NN ネットワークの可到達アトラクタ数に関する累積分布を示す．

$$P(c) = |\{u : |R(u)| \le c\}| / |X|. \quad (8)$$

$$P(c) = |\{u : |R(u)| / |X| \le c\}| / |X|. \quad (9)$$

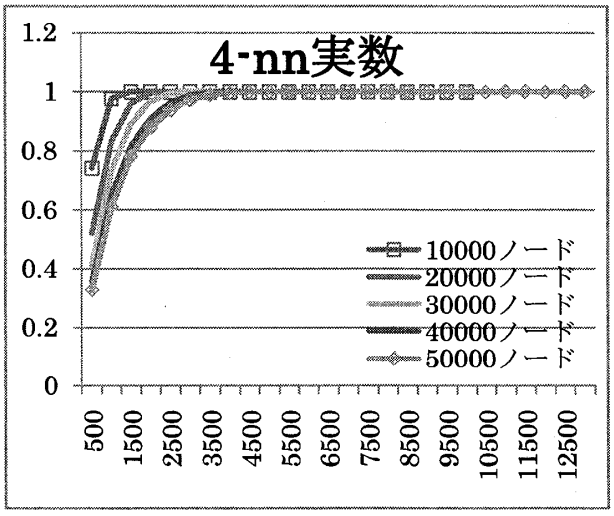

図 7:k-NN におけるベイスンサイズ(実数)

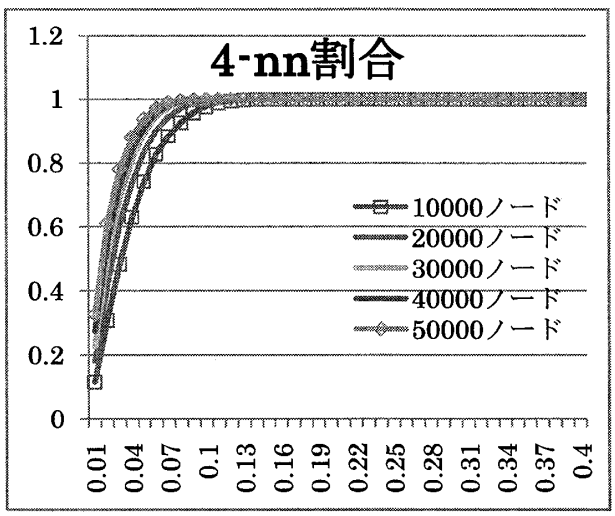

図 8:k-NN におけるベイスンサイズ(割合)

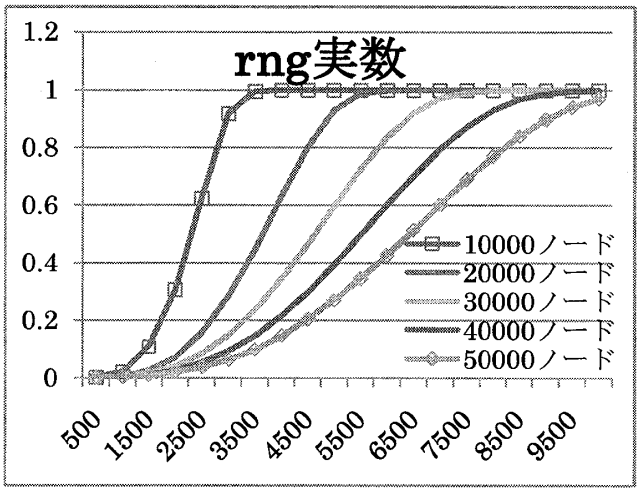

図 9:RNG における可到達アトラクタ数(実数)

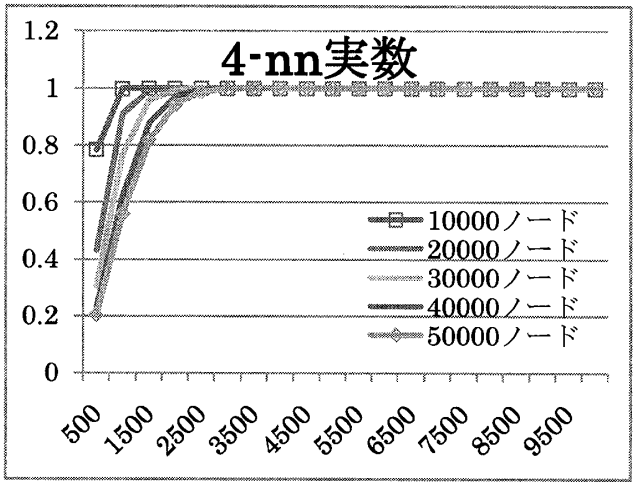

図 11:k-NN における可到達アトラクタ数(実数)

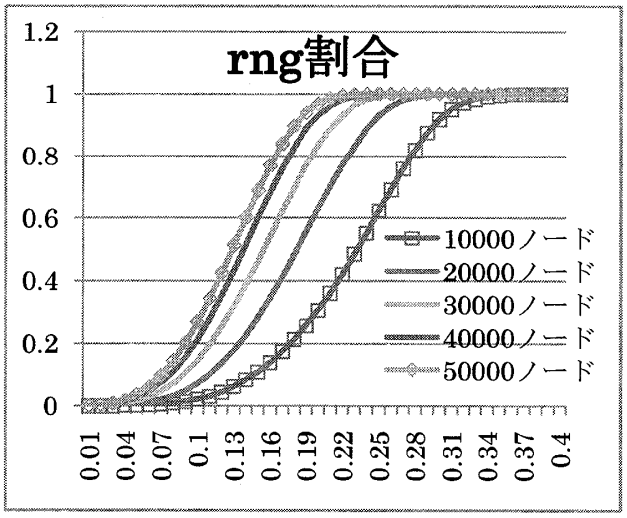

図 10:RNG における可到達アトラクタ数(割合)

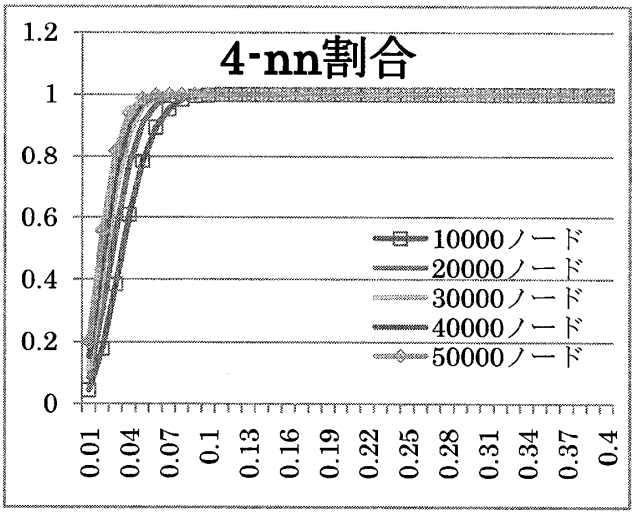

図 12:k-NN における可到達アトラクタ数(割合)

これらの図から，可到達アトラクタ数もベイスンサイズと同様にオブジェクト数が増加するにつれ相対的に少なくなる傾向にあることが分かる．また，RNG の可到達アトラクタ数も k-NN のそれより多くなる傾向が示されている．

4．おわりに

今回の実験では，異なる手法で構成されたネットワークを用いて類似探索問題に必要な三種の性質を比較した．その結果，k-NN ネットワークに比べ RNG ネットワークの方が類似探索性能の有効性が高いことが示唆された．今後は，さらに多様なデータへの適用を通して，ネットワーク構造と探索性能との関係解明を進める．

謝辞　本研究の一部は科研費(20500109) の助成を受けた．

参考文献

[1]Watts, D. J., Strogatz, S. H.: Collective dynamics of `small-world' networks. Nature 393, 440--442 (1998)
[2]Kleinberg, J.: Complex networks and decentralized search algorithms. Proc. Int'l. Congress of Mathematicians, (2006)
[3] Watts, D. J., Dodds, P. S., Newman, M. E. J.: Identity and search in social networks. Science 296, 1302--1305 (2002)
[4]Milgram, S.: The small world problem. Psychology Today 2, 60--67 (1967)
[5] Supowit , J. K.　The relative neighborhood graph, with an application to minimum spanning trees. Journal of the ACM (JACM)　30, 428—448 (1983)
[6]小出 明弘,外岡 達也,斉藤 和巳,青山 一生,澤田 宏,上田 修功:オブジェクト集合に依存した RNG の特性分析. 第 9 回情報科学技術フォーラム(投稿中)