社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2004－ICS－136 (3)
2004／8／4

ウェブ構造はどこまでスケールフリー/スモールワールドか？ ～ウェブ構造のモデル化に向けて～

福田 健介
NTT 未来ねっと研究所

本研究では，ウェブ構造の空間的・時間的モデル構築のために，ドメインレベルでの，ウェブ構造の統計的解析を行った．その結果，従来より指摘されているスケールフリーネットワークの他に，機械的リンク生成によって生成された，少なくとも２つの統計性の大きく異なるタイプのネットワークが存在することがわかった．また，時間的発展モデルに関しては，スケールフリー性を満たすac.jpドメインでは，ログスケールでネットワークの直径が大きくなるが，局所的な性質であるクラスタ率は大きく変化しないことが明らかになった．

Analysis of Statistical Properties of WWW

Kensuke Fukuda
NTT Network Innovation Laboratories

We analyze statistical properties of the structure of World Wide Web (WWW) in order to construct a spatial and temporal model for domain-level WWW. We find that there are at least two types of network structure largely different from the scale-free network due to an effect of the semi-automatically generated web pages. Also, we demonstrate that the WWW structure in ac.jp domain, which is well modeled by the preferential attachment, still satisfies the small world property robust against the growth of the domain.

1 はじめに

近年のネットワーク解析の結果より，World Wide Web (WWW) のページおよびリンクからなるネットワーク構造はスケールフリー性やスモールワールド性を持つことが指摘されている [1, 2, 3]. スケールフリー性とは，ページあたりの参照リンク数およびその被参照リンク数に平均的なサイズが存在しないことを意味している [4]. また，スモールワールド性とは，その名のとおり，「世間は狭い」という特徴– 小直径, 高クラスタ率–を持つネットワークのことである [5]. つまり，従来予想されていたよりも，ウェブの空間はコンパクトであり，かつリンク数には非対称性があることを意味している. これらの性質は，興味深いことに，ウェブ構造だけの特徴に限らず，人間関係や自然界の連鎖等にも広く見られる特徴である [6, 7].

スケールフリー性およびスモールワールド性の生成要因に関する解析，モデル化もなされているが，ウェブ構造に関しては，ページを作る際のユーザの挙動が強く反映すると考えられている (preferential attachment)[8]. つまり，人気のあるページにはリンクが張られやすい傾向にあるが，大多数のページにはほとんどリンクが張られないという結果が，統計性を決定しているというわけである. 一部のユーザがウェブを利用していた時代には，このモデルは広く受け入れられるものであったと考えられる. しかしながら，最近ではテンプレートを用いるページ作成ツールが生成したページ群や，ショッピングサイトのように大規模に自動生成されたリンクを持つページ群を内包したサイトも増えている. しかしながら，このようなページ群が，ネットワーク構造の統計性にどのような影響を与えているかについては，いまだ十分な研究がなされていない.

また，例えば，大学やネットワークプロバイダのような多数のユーザが個人のウェブサイトを持っている場合には，preferential attachment が成立しやすい状況であると考えることができる. しかしながら，規模の小さなドメインから大きなドメインまで，一律に preferential attachment モデルを適用することが可能であるかどうかは実証されていない. 言い換えると，ウェブ構造の成長は単一のモデルで記述可能かどうかはわかっていない.

そこで本研究では，検索エンジンロボットが収集した，日本のウェブ空間からドメインレベル (xxx.yy.jp) でのウェブ構造を抽出し，統計的な解析を行った. その結果，(1) 入出次数分布は，従来指摘されているスケールフリー構造の他に，指数型，固定型，ベキ減衰より緩やかな減衰を持つネットワークが存在することがわかった. 特に後者の 2 つのネットワークでは，自動的にリンクを生成するプログラムの影響によるものであることがわかった. (2) ウェブ空間の成長に関しては，ac.jp ドメインの解析により，ネットワークの直径は大学の (ウェブ空間上での) 規模に対して $O(\log)$ で増加するものの，局所的な性質であるクラスタ率はほぼ変化がないことがわかった. これはページ数が一定数以上の規模の大学ドメインでは，ウェブ空間を preferential attachment でモデル化可能であることを意味している.

2 スケールフリーネットワーク，スモールワールドネットワーク

この章では，ネットワーク構造を定量化する上で重要となる二つの指標–スケールフリーおよびスモールワールド–について説明する.

2.1 次数分布

ネットワークの解析では，あるノードから他ノードへのリンク数 (出次数) および，他ノードからあるノードへのリンク数 (入次数) に着目することが多い. そして，この入出次数の分布は，ネットワークを特徴づける重要な指標である. 近年の各種解析によって，多くの実ネットワークでは，入次数 (出次数)k の次数分布 $P(k)$ はいくつかの分布に分類されることが報告されている [4].

1. ベキ的減衰: $P(k) \sim k^{-\gamma}$.
2. ベキ的減衰 + 指数的カットオフ: $P(k) \sim k^{-\gamma}\exp(-k)$.
3. 指数 (またはガウス) 的減衰: $P(k) \sim \exp(-k)$.

ここで注目すべきは，(1) のようにベキ的減衰が現れる場合である. ベキ的減衰は，多くのサンプルを調査すればするほど，(確率的には) 大きな値の次数が現れることを意味する. つまり，入出次数には典型的なサ

イズが存在しない．他の見方をすれば，大多数のノードは少数のリンクしか持たないが，少数のノードは多数のリンクを持つという，非対称的な構造であると言える．このようなべキ的減衰を伴う次数分布で特徴づけられるネットワークは，スケールフリーネットワークと呼ばれている．(2) はスケールフリー構造になんらかの制約 (フィルタ) が加えられている場合に対応する [10]．それに対して，(3) は分布の裾が短い分布であり，入出次数に特徴的なサイズがあることを示している．

Albert らは， Nortre Dame 大内のウェブ空間を解析し，その入出次数分布がべキ的減衰 ($\gamma \sim 1.1$) にしたがうことを報告している [1]．同様に， Broder らは，ウェブページ収集ロボットが収集したウェブページ空間を解析し，やはり入次数分布にべキ的減衰が現れることを示している．

2.2 平均距離，クラスタ率

以下では， ネットワーク構造の成長を特徴づける，2 つの量–距離およびクラスタ率– について説明する．

ネットワーク上の 2 つのノード間の最短経路長とは，あるノードからもう一方へのノードに到着する際に必要なリンク数の最小値である．そして，ネットワークの直径は，任意の 2 ノード間を経由する最短経路のうちで，リンク数が最大のものと定義される．部分木が複数存在する場合には，最も多くのノードを含む部分木の直径をネットワークの直径とする．直径は，ダイクストラのアルゴリズムを全てのノードに適用して，その最短経路長の最大値を求めることで計算できる．同様に，直径と関係する量として，平均距離は，ネットワーク上の任意の 2 ノード間の最短経路長の平均値と定義される．直径および平均距離は，ネットワークの大域的な性質を表している．

次に，局所的な性質であるクラスタ率について説明する．クラスタ率は，直観的には，「自分の友達同士が友達である」確率に対応する．つまり，自ノードに隣接するノード間の接続度合いを示している．k 本のリンクを持つノード n のクラスタ率の計算方法は以下のとおり: ノード n からリンクされた k 個のノード間の最大可能リンク数 k_{max} は，$k_{max} = k(k-1)/2$ 本である (完全結合)．k 個のノード間で張られている実際のリンク数を k_{exist} とすると，ノード n でのクラスタ率 C は，$C = k_{exist}/k_{max}$ と定義される．C は 0 から 1 の間の値を取り，ネットワークが完全結合の場合には，$C = 1$ となる．クラスタ率は，ネットワークの局所的な性質を示していると考えられる．

一般に，ネットワークの直径が小さく，クラスタ率が大きいネットワークはスモールワールドネットワークと呼ばれる．それに対して，レギュラーラティスではネットワークの直径は大きくクラスタ率も大きい．また，ランダムネットワークではネットワークの直径は小さくクラスタ率も小さい．つまり，スモールワールドネットワークは両ネットワークの中間的な性質を満たしているとも言える．

3 データセット

本研究では， ウェブページ収集ロボット [9] が収集した，2001 年 12 月の日本のウェブ空間のスナップショット (総ページ数 182,770,391，総リンク数 608,790,185) を使用した．なお， このスナップショットは当時の日本の総ページ数の 5 割程度に相当すると推定される．この元データより， ページに含まれるリンク情報を抽出し， 同一ドメイン内のリンク関係–リンク元ページおよびリンク先ページ –のリストをドメインごとに作成した．その際，末尾が “/index.html”, “index.htm” で終わるものに関しては，“/” に置換した．

4 ウェブ構造の空間的パターン

4.1 ユーザの嗜好による効果

この節では，ページ生成の際にユーザの嗜好が反映されやすいと考えられる，ac.jp ドメインでの例を示す．図 1 は keio.ac.jp 内のウェブ空間での入出次数分布である．繰り返しになるが， 入次数は， あるページに対しての被参照リンク数， 出次数はあるページから出る参照リンク数である．入次数は $1 < k < 20,000$ で， 出次数では $10 < k < 1,000$ で，分布はべキ的減衰となっている．言い換えると， 大多数のページは他のページからはあまり指されていないが，一部のページは非常に多数のページから参照されていることに対応する．ac.jp ドメインでの結果は， preferential attachment の仮定– 多数のユーザが， それぞれの嗜好によってリンクを張る– が成立しやすいことがわか

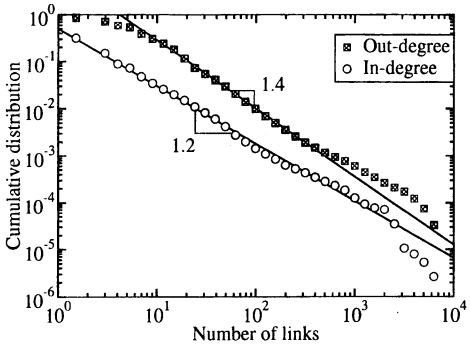

図 1: 入出次数分布 (keio.ac.jp).

る. ac.jp に属するドメインの入次数分布は, 一定以上のページを持つドメインでは, ベキ的減衰が観測されやすいこと, また, 規模の小さなドメインの出次数分布では, 指数的な減衰が見られることが指摘されている [11].

4.2 機械的リンク生成による効果

次に, 機械的なリンクが生成されやすい, 商用サイトの影響を調査した. 図 2 は amazon.co.jp のサイト内

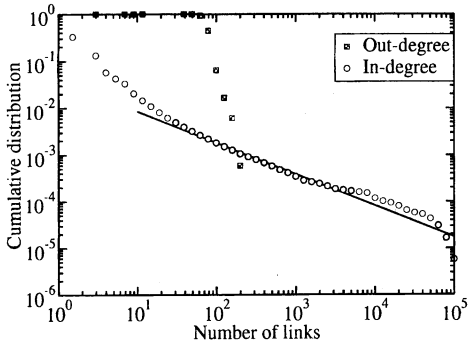

図 2: 入出次数分布 (amazon.co.jp)

ウェブ空間における入出次数分布である. 大学内のウェブ空間とは異なり, 2 つの分布は異なる傾向を持つことがわかる. まず, 出次数分布は, 10^2 程度に集中している. すなわち, 大部分のページは 100 本程度のリンクを持ち, それ以外のページはほとんど存在しないことを意味している. これは, 大部分のページは個々の商品を表す定型型のページであることを示唆している. それに対して, 入次数分布は, $50 < k < 50,000$ の範囲で, ほぼベキ的減衰となっている. 個々の商品を表すページを指すリンクは少ないため, $k < 50$ では分布がベキからはずれていると考えられる. 逆に, ベキ的減衰が成立している領域は, 個々のページをまとめているページ (各種分野のトップページ) 等に対応しているものと考えられる. このような傾向は, 他の電子モールやニュースサイト等でも観測されている.

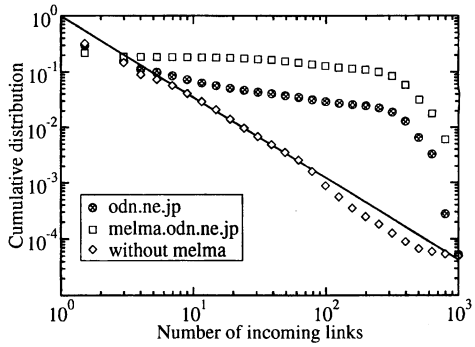

図 3: 入次数分布 (odn.ne.jp)

次に示す例は, インターネットプロバイダ内のウェブ空間の入次数分布である (図 3). 直感的には, プロバイダでは顧客であるユーザが個々のページを持っているため, ac.jp ドメインのネットワークの持つ性質と同様の性質が観察されると期待できる. しかしながら, 興味深いことに, プロバイダの空間全体で見た場合には, 次数分布はベキ的減衰よりもさらに緩やかな減衰となっている. つまり, 小さな入次数のページから大きな入次数のページまでが一定の確率に近い状態で存在することを意味している. この傾向は, 出次数分布でも同様の結果であった.

データを詳細に解析したところ, このプロバイダは, ユーザの持つウェブページ群の他に, メールマガジンのアーカイブサイトを持っていることがわかった. このサイトの各々のページは, メールマガジンの一回のメールに対応しており, 各号のページから全てのバックナンバーに対してのリンクが存在していた. つまり, 各メールマガジンの過去の発行回数を n とすると, 生成されるネットワークはノード数 n からなる完全結合ネットワークになっている. メールマガジンの数は数

千のオーダであり，これらの存在が入出次数分布のテイルを引き延ばす効果として働いていることがわかった．実際，図3にあるように，メールマガジンサイトだけを分離した空間の分布は裾の長い分布になっており，その補空間の分布は，ac.jpドメインのサイトの結果と同様にベキ的減衰を持つ結果となった．このベキ的減衰は，プロバイダーが提供する個人のウェブサイトの空間構造が反映されているものと考えられる．完全結合に近いネットワークは，他にもランキングサイト等で観測されており，ウェブサイトの構築の仕方によっては，特に珍しいものではないと言える．

4.3 まとめ

ウェブネットワークの入出次数分布に関しては，Amaralらが指摘しているような3つのタイプの他に，少なくとも，ほぼ固定サイズの分布を持つタイプおよび，完全結合に近いタイプが観測された．表1に，今回の解析で観測された入次数および出次数分布の関数型の組み合わせを示す．実際にはある程度の規模以上のサイトでは，入次数にはベキ的減衰が観測され，それ以外の関数型はほとんど観測されない．例外は，ベキ的減衰よりもテイルが緩やかな減衰である．それに対して，出次数分布では各種分布が散見される．これは，出次数にはサイト構成の特色が現れやすいことを示している．

表1: 入出次数分布に現れる関数型; power (ベキ的減衰), exp (指数的減衰), fix (均一なサイズ), full (完全結合).

in / out	power	exp	fix	full
power	○	○	○	×
exp	×	○	?	×
fix	×	×	×	×
full	×	×	×	○

5 ウェブ構造の成長モデル

この節では，規模の小さなドメインがどのような過程を経て大きなものへと変遷していくのかについて考える．本研究で使用したデータは，ある時点でのウェブ空間のスナップショットであるため，このデータのみからでは，成長の過程について議論することはできない．しかしながら，ドメイン内でのページ(リンク)生成方法の差違が非常に小さいものであると仮定できるのではあれば，より多くのページを持つドメインは，少ないページ数を持つドメインと比べて成長が進んでいると考えることは妥当であろう．前節で示したように，商用サイトではpreferential attachmentの仮定が成立しないような，リンク構造が散見されている．しかし，ac.jpに属するドメインのウェブ空間は，スケールフリー性が成り立つ傾向にあることが報告されている[11]．そこで，本解析ではページ生成方法の差違が少なく，ユーザの挙動が反映されやすいac.jpドメインに着目し，その変遷の過程を解析する．

5.1 解析結果

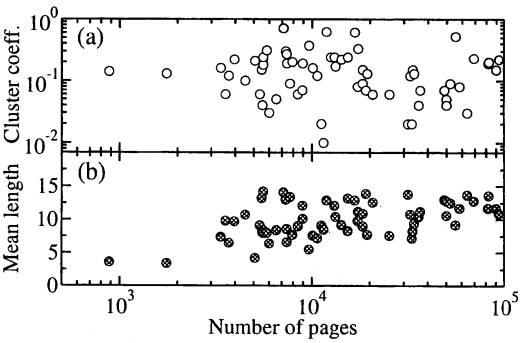

図4: ドメインの規模とクラスタ率(a)および平均距離(b).

ac.jpドメインに含まれるドメインのウェブ空間の平均距離およびクラスタ率を計算した結果を図4に示す．計算にあたっては，個々のパラメータの計算量を減らすために，incoming linkが0のノード(誰からも指されていない)および，outgoing linkが0のノード(誰も指していない)を枝刈りしたネットワークを使用した．グラフ中の個々のプロットが個々のドメインに対応している．前述のように，直径はウェブ空間の広域的な性質を示し，クラスタ率はウェブ空間の局所的な性質を示している．

図より，平均距離はページ数の増加に対してほぼロ

グスケールで増加していることがわかる．たとえば 10^5 程度のページを持つ組織では，あるページから他のページに移動するために平均 10 回程度リンクをたどる必要がある．ログスケールでの増加は直観的にはネットワーク構造が階層的になっていることを意味する．それに対して，クラスタ率はページ数に依らずほぼ $O(10^{-1})$ となることがわかる (縦軸も log スケールであることに注意)．つまり局所的な構造はどのドメインを見ても，統計的な違いが少ないことを意味している．

これらドメインのネットワークが果たしてスモールワールドと言えるかどうかであるが，[11] によれば，同規模のノードおよびリンクを持つレギュラーラティスやランダムネットワークでは，直径およびクラスタ率のオーダーはそれぞれ，$O(C_{regular}) \sim 10^{-1}$，$O(L_{regular}) \sim 10^3$，$O(C_{random}) \sim 10^{-3}$，$O(L_{random}) \sim 10^1$ である．以上の結果より，ac.jp ドメインを対象とした，ネットワークの成長モデルは，スモールワールド性を満たしていると考えることができる．

6 おわりに

本研究では，ウェブ空間のモデル化を行うために，ドメインレベルでのウェブ空間の統計的な解析を行った．その結果，空間的な特徴としては，従来のスモールワールドネットワークに関する研究で指摘されているネットワーク構造の他に，ウェブ特有な現象である商用サイト等に見られるリンクの自動生成による影響が，統計性を大きく変える可能性があることを示した．また，時間的な特徴としては，ac.jp ドメインに属するドメインを解析した結果，ネットワークの平均距離はログオーダーで増加するものの，局所的なクラスタ率はほとんど変化しないことがわかった．

今後の方針としては，(1) 入出次数分布の関数型パターンの出現頻度の解析 (2) ac.jp ドメイン以外の時間発展の特徴解析を進める予定である．

参考文献

[1] Albert, A., Jeong, H., and Barabasi, A.-L. Diameter of the World Wide Web. *Nature* **401**, 130–131 (1999).

[2] Broder, A., *et al.* Graph structure in the web. *Proc. of World Wide Web Conference 9*, (2000).

[3] Adamic, L. A. The small world web. *Proc. ECDL'99*, LNCS 1696, pp.443-452, (1999).

[4] Amaral, L.A.N., Scala, A., Barthelemy, M., and Stanley, H.E. Classes of small-world networks. *Proc. Natl. Acad. Sci. USA* **97**, 11149–11152 (2000).

[5] Wattts, D.J. and Strogatz S.H.: Collective dynamics of 'small-world' networks. *Nature*, vol.393, pp.440–442 (1998).

[6] Watts, D.J.: *Small Worlds*, Princeton Univ. Press, (1999).

[7] Barabási, A.-L.: *Linked: The New Science of Networks*, Perseus Publishing, (新ネットワーク思考 (青木訳), NHK 出版). (2002).

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

[9] ODIN project. `http://www.ingrid.org/odin/`.

[10] Mossa, S., Barthelemy, M., Stanley, H.E., and Amaral, L.A.N.: Truncation of power law behavior in "scale-free" network models due to information filtering. *Phys. Rev. Lett.*, vol.88, 138701, (2002).

[11] 福田 他: ac.jp ドメインにおけるドメイン内 Web リンク構造の解析. *Proc. FIT2003*, 情報処理学会 (2003).