

World Wide Web 情報空間のシミュレーション

高 橋 良 昌[†]

本稿では、World Wide Web (Web) 情報空間を確率的に時間変化する有向多重グラフとしてモデル化し、そのモデルを用いて行なったシミュレーションの結果について報告する。このモデルでは、有向多重グラフのとり得るすべての状態からなる空間 Ω の中に明白なアトラクタが複数存在する。シミュレーションの結果によると、限られた場合を除いて Ω は明白なアトラクタのベイスンで覆われていること、アトラクタに収束するまでの遷移時間はサイトが張るリンクの数に比例し、その比例係数はサイトの数によって増加することが示唆された。

Simulations of World Wide Web Information Space

YOSHIMASA TAKAHASHI[†]

This paper describes a model of World Wide Web (Web) information space. The model is represented as the directed multigraph which evolves stochastically. And also this paper reports results of simulations using the model. About this model, a space Ω which is composed by all states such that the directed multigraph can be in, has obvious attractors. Results of simulations suggested that except extreme cases, Ω is covered with basins of obvious attractors, and that transient times to attractors are proportional to the number of links of a site, and its rates increase as the number of sites.

1. はじめに

現在 World Wide Web (以後 Web) 上で多くの情報が公開されており、それら情報間のリンクにより情報空間が形成されている。今後より多くの情報が Web に追加されるにつれ、リンクにより形成される情報空間の大きさも増すと予測される。この成長する Web 情報空間の中から目的の情報を検索するため、Web 情報空間の構造に関する研究がなされている。

例えば Mukherjea ら¹⁾ は局所的な Web 情報空間を Overview Diagram というグラフ構造を用いて可視化し、視覚的な検索システムを提案した。また Woodruff ら²⁾ は Web を構成するページについて広範囲に調査し、各種の統計量を報告している。そして中川ら³⁾ は日本における 1996 年 1 月のリンクデータをもとに基本的な統計量を算出した。

これらは Web 情報空間のある時刻での構造についての研究であり、Web 情報空間の時間変化を扱ったものではない。しかし Web 情報空間を知識の集積所として考えた場合、その構造の将来像を予測すること

は興味深い課題である。そこで本稿では、Web 情報空間の将来について定性的な示唆を得ることを目的とする。

2. モデル

現在 Web 上で多くの個人または団体により情報が公開されている。それら情報発信者による情報のまとまりを、ここではサイトと呼ぶ。各サイトの情報は HTML⁴⁾ により記述され、その仕様の一部である href 属性を用いたリンクにより、ある情報から自他のサイトの別の情報を参照することができる。この参照関係には明かに参照元のサイトから参照先のサイトに向かう方向性がある。本研究ではこのサイト間の参照構造に着目する。

2.1 多重グラフによる記述

サイト間の参照関係を多重グラフ $G = (V, E)$ で記述する。ここで $V = V(G)$ はグラフ G の頂点集合、 $E = E(G)$ はグラフ G の辺集合である。1 つの頂点 $v \in V$ は Web 上の 1 つのサイトを表す。あるサイトは href 属性を用いて、他のサイトや自分自身の内容を参照する。1 つの辺 $e \in E$ はこの参照 1 つを表す。ある辺 e は参照元を表す頂点、および、参照先を表す頂点を識別する情報を持つ。すなわち、 G は有向グラ

[†] 北海道大学大学院理学研究科
Graduate School of Science, Hokkaido University

フである。

2.2 仮定

サイト間の参照関係は定常でなく、サイトの更新により書換えられる。すなわち、Web 情報空間を記述する有向多重グラフ $G = (V, E)$ の構造は時間変化する。その時間変化のルールを以下の仮定を用いて定める。

仮定 1 頂点 $v \in V(G)$ は生成も消滅もしない。

仮定 2 ある 1 つの頂点 $v \in V(G)$ について、 v 自身や他の頂点へ向かう辺の合計数は一定で変化しない。

仮定 3 $\forall v \in V(G)$ について、 v 自身や他の頂点への辺の数は等しい。

仮定 4 すべての $v \in V(G)$ は一定間隔で更新され、 G の構造を変える。

仮定 5 すべての $v \in V(G)$ の更新は同期的に行われる。

仮定 6 ある頂点 v_i から別の頂点 v_j への 1 つの辺について、頂点 v_i の更新により、この辺は頂点 v_j が隣接する頂点の 1 つへ向かう辺に書換えられる。

仮定 7 仮定 6 において、 v_j が隣接する頂点を $v_{j,1}, \dots, v_{j,n}$ 、 v_j から $v_{j,k}$ ($1 \leq k \leq n$) へ向かう辺の数を $r_{j,k}$ とする。このとき v_i の新たな近傍として $v_{j,k}$ は確率

$$P_{j,k} = \frac{r_{j,k}}{\sum_{l=1}^n r_{j,l}}$$

で選ばれる。

仮定 8 ある頂点 v_i が更新されるとき、 v_i から他の頂点への辺はすべてが同時に書換えられる。

仮定 4, 5, 8 はサイトあるいは参照を更新する順番の影響を取り除く目的で導入する。非同期的なサイトあるいは参照の更新の方法は無数に考えられる。しかし、それらの更新方法の違いがシミュレーションの結果に与える影響は予測できない。そこで本研究では、サイトあるいは参照の更新を同期的に行うことにより、更新の順番付けの影響をなくした。

有向多重グラフ G を実際に時間変化させるのは、仮定 6, 7 によるルールである。これらのルールは、 G を構成する各頂点およびその近傍に対する、局所的な変化を規定しているにすぎない。しかしこの局所的なルールの結果、後述するアトラクタによる大局的な変化の構造が生じる。

2.3 アトラクタ

前項の仮定のもとで Web 情報空間が時間変化するとき、Web 情報空間がとり得る状態すべての中で、アトラクタとなる状態が存在する。これを示すため、まず、ドミナント、参照容量、占有率、および隣接行列を定義する。

定義 1 (ドミナント) ある時刻での Web 情報空間を記述した有向多重グラフ $G = (V, E)$ を考える。ここで $V = V(G)$ は G の頂点集合、 $E = E(G)$ は G の辺集合である。いま、 V の元を自然数で添字付ける。

$$V = \{v_1, v_2, \dots, v_n\}$$

ここで n は G の頂点の個数である。 v_j から v_i へ向かう辺の数を $r_{j,i}$ とし、 $\forall v \in V$ から v_i へ向かう辺の数を s_i とする。ここで

$$s_i = \sum_{k=1}^n r_{k,i}$$

である。 s_i ($1 \leq i \leq n$) の最大値を $s_{max} = \max\{s_1, s_2, \dots, s_n\}$ としたとき、 $s_i = s_{max}$ を満足するすべての v_i を G のドミナントと呼ぶ。

定義 2 (参照容量) 定義 1 において、

$$c_j = \sum_{k=1}^n r_{j,k}$$

を $v_j \in V(G)$ の参照容量と呼ぶ。

定義 3 (占有率) 定義 1 において、 G のドミナントの 1 つを $v_i \in V = \{v_1, v_2, \dots, v_n\}$ とする。このとき v_i の占有率 pr を

$$\begin{aligned} pr &= \frac{s_i}{\sum_{k=1}^n s_k} \\ &= \frac{s_{max}}{\sum_{k=1}^n s_k} \end{aligned}$$

で定義する。

定義 4 (隣接行列) 定義 1 において、 j 行 i 列成分が $r_{j,i}$ である n 次正方形行列 R を G の隣接行列と呼ぶ。

Web 情報空間が 2.2 項での仮定に従って時間変化する、「明白なアトラクタ」が存在するようになる。この「明白なアトラクタ」をいままで述べた定義を用いて定義する。

定義 5 (明白なアトラクタ) 頂点の個数 n 、すべての頂点の参照容量が c である有向多重グラフのとり得るすべての状態からなる集合 Ω を考える。ある頂点から別の頂点に向かう辺が複数あるとき、各辺 1 つ 1 つを区別しなければ、すべての $G \in \Omega$ はその隣接行列 R と 1 対 1 に対応する。さらにすべての $G \in \Omega$ は、その隣接行列 R の n 個の行を 1 列に並べたベクトルを作ることにより、 n^2 次元の離散ベクトル空間上の点に 1 対 1 に写像することができる。これにより Ω の各要素間の距離を定義することができる。

いま 2.2 項での仮定に従う Ω 上の流れを考えると、ドミナントの占有率が 1 であるような $G^* \in \Omega$ に着目する。 G^* はただ 1 つのドミナントを持ち、そのド

ミナントを含む G^* のすべての頂点からの辺が集中する構造を持つ．このような G^* は明らかに固定点である．その隣接行列 R の成分 $r_{i,j}$ は，ドミナントを v_d , ($1 \leq d \leq n$) とすると

$$r_{i,j} = \begin{cases} c & \text{if } j = d \\ 0 & \text{if } j \neq d \end{cases}$$

となる．これらの G^* は安定な固定点である．そこでこれら G^* を明白なアトラクタと呼ぶ．

3. シミュレーション

ここでは，前節で提案した概念モデルを用いたシミュレーションの結果について述べる．

3.1 シミュレーション例

図 1 にシミュレーションの実行例を示す．この例では頂点数 20，参照容量 19 である．

外側の円周上に並んでいる円が頂点を表す．頂点のうち黒く塗りつぶされているものが，ドミナントである．ある頂点から別の頂点へ向かう辺は 1 本の矢線で表し，その濃度の濃淡で辺の数の多少をそれぞれ表す．ある頂点から自分自身へ向かう辺は，その頂点のさらに外側に付く小さな円状の矢線で表す．

最初ある頂点から出る辺はすべての頂点に対して等しい確率で向かうように初期化される（図 1 (a)）．時間とともにすべての辺は一部の頂点に向かうようになり（図 1 (b) (c)），最後には 1 つの頂点に対してすべての辺が集中する明白なアトラクタへと収束する（図 1 (d)）．

頂点数および参照容量を変えて実験すると，多くの場合上記の例と同様に最終的には明白なアトラクタへと収束した．そこで頂点数と参照容量をパラメータとして，様々な初期状態から出発したときに明白なアトラクタへと収束する確率と，収束する場合について遷移時間の期待値を実験により求めた．

3.2 明白なアトラクタへの遷移実験

実験は各頂点数，参照数ごとに 2×10^3 回の試行を繰り返して行った．各試行の初期状態は乱数を用いて定め，仮定 4 で定めた単位時間を 1 ステップとする，最大 10^3 ステップまで有向多重グラフを時間変化させ，その時間内に明白なアトラクタに収束するものを計数した．また収束した試行については収束までに要した時間を記録しておき，頂点数，参照数ごとの遷移時間の平均値を求めた．

図 2 に参照容量を変化させたときの明白なアトラクタへ収束する確率の変化を示す．横軸に参照容量，縦軸に収束確率をとる．

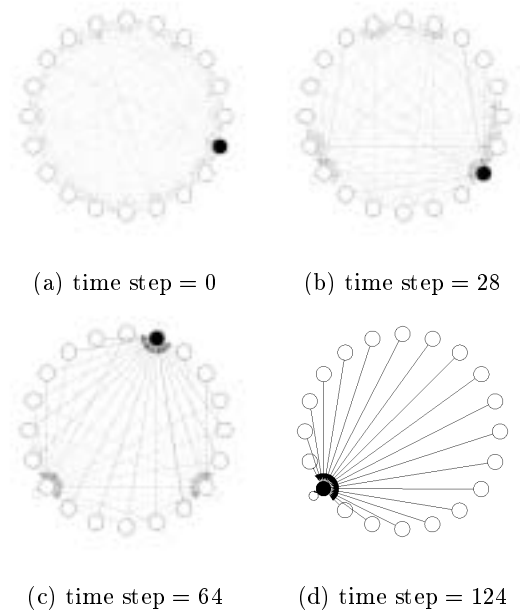


図 1 シミュレーションの実行例
Fig. 1 Example of a simulation.

図 2 には頂点数 (a) 100 (b) 50 (c) 30 (d) 10 の 4 通りについての結果を示してあるが，ほぼ 1 本のグラフの線に重なっている．これは収束確率が頂点数に依存しなかったことを表わしている．参照容量が極めて小さい場合を除いて，有向多重グラフのとり得る状態のほとんどは明白なアトラクタのベースンに覆われていることが分かる．

収束する場合の遷移時間について，図 3 に参照容量による変化を，図 4 に頂点数による変化を示す．

図 3 には頂点数が (a) 100 (b) 50 (c) 30 (d) 10 の 4 通りの場合についての結果を示す．横軸に参照容量，縦軸に遷移時間をとる (a) から (d) のいずれの場合も参照容量と遷移時間がほぼ比例しており，その比例係数は頂点数とともに増加している．

一方図 4 には参照容量が (a) 100 (b) 50 (c) 30, (d) 10 の 4 通りの場合についての結果を示す．横軸に頂点数，縦軸に遷移時間をとる (a) から (d) のいずれの場合も遷移時間は頂点数が増加するとともに増加したが，そのグラフの線は上に有界であるように見える．

4. 議 論

ここでは前節で得られたシミュレーションの結果と現実の Web 情報空間との対応を考察する．

図 2 によると，あるサイトの自分自身や他のサイト

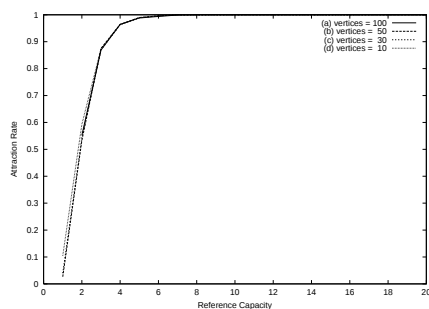


図 2 明かなアトラクタへの収束確率

Fig. 2 Probabilities of the convergence to obvious attractors.

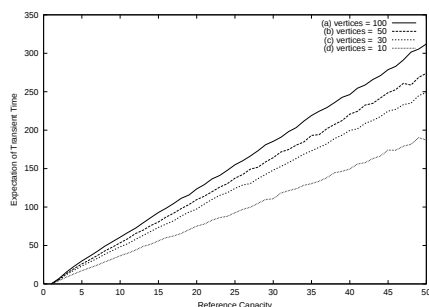


図 3 遷移時間の参照容量変化

Fig. 3 Dependence of the transition times on the reference capacities.

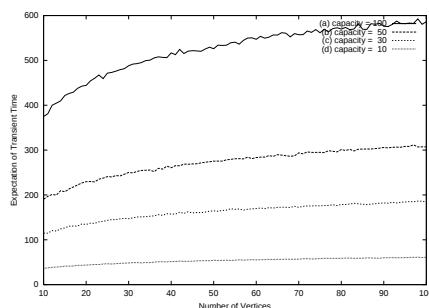


図 4 遷移時間の頂点数変化

Fig. 4 Dependence of the transition times on the numbers of vertices.

に対する参照の数が極端に少ない場合を除いて、最終的には 1 つのサイトがすべての参照を集めることが予想される。そして図 3、図 4 によれば、サイト数が十分多ければ、そのような最終状態になるまでにかかる時間は、サイト数によらず各サイトの参照数に比例して長くなる。

本研究ではサイトが持つ情報の内容までは考慮に入れていない。しかし仮に参照をより集めているサイトが情報のより豊富なサイトであるとする、以上の結果から Web 情報空間は十分先の将来においてただ

1 つのサイトに情報が集積され、そのサイトを他のすべてのサイトが参照する構造になると予測される。

しかしこの予測は、Web 上に情報は分散している、という直観に反している。そこで個々のシミュレーションの結果をより詳細に検討する。

中川ら³⁾によれば、1996 年 1 月の日本の Web 情報空間の構造の特徴として、

- (1) 全体の相互参照率が低く、サイト間の参照は片方向が多い
- (2) 同一サイト内の参照が多く、他のサイトが提供する情報への参照は少ない
- (3) 全体の傾向として、あるサイトは少数の特定サイトにリンクを張る

が挙げられる。本研究でのシミュレーションは 1, 3 の 2 点についてはこの特徴を良く再現できた。2.2 項の仮定 6, 7 による局所的相互作用から、このような大局的な構造がどのようにして生じるのかについてはまだ分かっておらず、今後の研究課題である。また 2 については、多くの場合自サイトへの参照を保存できずにすぐに他サイトへ向かう参照に変化してしまうため、良く再現することができなかった。これは仮定 7 において、新たな近傍へと遷移する確率の与え方に問題があるためと考えられる。本節で先に述べたように、仮により参照されることがより情報のあることであるとする、この仮定ではあるサイトの管理者が実際より自サイトの情報を軽んじるようにしていることを意味している、と考えられる。この点に関しては今後モデルの改良が必要である。

参 考 文 献

- 1) Mukherjee, S. and Foley, J.: Visualizing the World-Wide Web with the Navigational View Builder, Available as http://www.igd.fhg.de/archive/1995_www95/proceedings/papers/44/mukh/mukh.html. (1995).
- 2) Woodruff, A., Aoki, P., Brewer, E., Gauthier, P. and Rowe, L.: An Investigation of Documents from the World Wide Web, Available as http://www5conf.inria.fr/fich_html/papers/P7/Overview.html. (1996).
- 3) 中川格, 石塚英弘, 山本毅雄: 日本の World Wide Web 情報空間: 1996 年 1 月のリンクデータ解析, Available as http://www.dl.ulis.ac.jp/DLjournal/No_9/1_itaru/1-itaru.html. (1997).
- 4) W3C: *HTML 4.01 Specification* (1999). Available as <http://www.w3.org/TR/html401/>.