

outdegree 図に着目したウェブリンク構造変容分析による デバイド状態の評価

Assessment of states of digital divide using Web linkage structure analysis focusing on outdegree graphs

熊澤 健人[†]
Kento Kumazawa

中平 勝子[†]
Katsuko T. Nakahira

北島 宗雄[†]
Muneo Kitajima

1 はじめに

ウェブ利用者がウェブから情報を得ようとする際、言語やウェブ利用者の情報探索能力、用いるハードウェア、インターネット環境など様々な障壁によって得られる情報には格差が生じる。インターネット環境や経済的に発展した国などでは、普段十分な情報を得られることが多いため、この格差を利用者が感じる機会は多くはない。しかし、実際にこの格差は存在していてこの格差を測ることができる指標を示し、こういった要素によって格差が生じるのかを示す必要がある。本稿では、outdegree 図を用いてこの指標を示し、シミュレーションを用いて outdegree 図の生成し形状の変容からデジタルデバイドの指標と成り得るパラメータを示し、デジタルデバイドの指標とすることを目的とする。

2 outdegree 図を用いたデジタルデバイドの評価

本研究では、ウェブ空間からウェブ利用者が得ることができる情報量の差によって生み出される格差をデジタルデバイドと捉える。デジタルデバイドは「基盤」「利用」「効用」という各段階や複合的な段階において生じると考えられる [1]。

これらの 3 要素の内「基盤」「効用」は、ハードの普及や電子商取引の額などといった統計的データから観測することができる。そのため、今回は観測することが難しい「利用」に焦点を当てて「利用」によってデジタルデバイドがどのように生じるかを考察する。デジタルデバイドを考えるとときに考慮すべき項目として以下を挙げる。まず考えられる問題がウェブページ数である。ウェブページ数が多いウェブ空間と少ないウェブ空間では、多い場合の方が求めている情報がウェブ空間に存在している可能性は高くなる。しかし、ウェブ利用者のインターネット利用環境や使用言語によってウェブ利用者にとって有用なウェブページが絞り込まれる。

続いて、その絞り込み条件の一つである言語の問題である。ウェブ利用者はウェブ上から情報を得る際に自らが理解できる言語で情報を検索する。それは、自らが情報を得るためには、自分の理解できる言語で情報を見る必要があるためである。自分の理解できない言語で情報が存在していても情報を読み取ることができない。言語によってウェブ上に存在している情報量には格差が存在する。それは、使用言語ごとのウェブの利用者数などによって異なり、利用者が多ければ情報が情報が多く、利用者が少なければ情報も少ないということになる。ウェブの利用分野においても顕著にその傾向は現れる。学術分野では、主に英語が用いられるなど特定の言語に情報が集約される傾向もある。

最後にウェブ上のリンク構成である。インターネットの利点として情報の双方向性がある。情報の送受信において情報を受

信したい時、求めている情報の流通度が関係してくる。ウェブ上に情報が存在していたとしても、その情報に辿り着くことができなければ情報を取得することができないが、流通度の高い情報は取得することが容易である。情報の流通度の指標として、今回はリンクを用いる。

これらの要素によってデジタルデバイドが生じると考える。

本稿で指標として用いる outdegree 図とは、横軸に出次数、縦軸に URL 数をとる両軸対数グラフのことを指す。

入出次数分布を用いてウェブ空間のリンク構造を観測するとネットワークのフリースケール性によってベキ乗的減衰が現れる。多くの商用サイトなどで用いられる機械的リンク生成によって出次数には縦一直線の五月雨的な構造が現れるといった 2 点が福田らによって示唆されている [2]。(本稿では、機械的リンク生成を無意識的リンク生成と捉える。) 加えて、石原らによって出次数 50 程度までは水平構造が現れることがあるということも示唆されている。この水平構造は個々のウェブサイトの巨大化による内部リンクの増加によるものとされる [3]。

これらの特徴は、全てウェブの成長し個々のウェブページが有する出次数の増加に伴って現れる特徴である。ウェブが成長し出次が増加したウェブ空間では、ウェブが成熟しておらず出次数が少ないウェブ空間に比べて、求めた情報に辿り着くことが容易であると捉えることができる。求めた情報に対するアクセスの容易さの違いは、ウェブ利用者が得られる情報量に格差を生じ、この格差がデジタルデバイドを生み出す。

そのため、上記の 3 つの特徴はデジタルデバイドを表す指標として用いることができると考えられる。

3 シミュレーションによる outdegree 図の再現

今回のシミュレーションではリンク生成に関与すると考えられるいくつかの属性を持ったウェブページで構成されたウェブ空間を擬似的に作り出し、そのウェブ空間から outdegree 図を構成する。持たせる属性を以下のように設定する。

【ウェブページに使用される主要言語】リンクの構成され易さには言語の種類が関与することが考える。具体的には、読むことができる言語から成るウェブページは、読むことができない言語から成るウェブページに比べるとリンク生成され易いということが予想される。しかし、意識的リンク生成ではこのような性質が予想されるが、無意識的リンク生成によってリンク生成される場合この限りではない。

【ccTLD】これは情報の探し易さの観点からデジタルデバイドに関与する属性である。国内リンクしか辿ることができないか国外リンクも辿ることができるかによるが、国内に情報がある場合と、国外に情報がある場合では国内に情報がある場合の方が情報に対してのアクセスが容易であると考えられる。そして、ある国の ccTLD にどの程度ウェブページがあるのかということは、その国のウェブ空間にどの程度の情報があるのかを表すことができる。実際 ccTLD とサーバの所在国が同一であ

[†] 長岡技術科学大学

るとは限らないが今回は ccTLD とサーバの所在国は同一であると仮定し、ccTLD をサーバの所在、アクセスのし易さの指標として用いる。

以上の 2 つの属性が今回設定するデジタルデバインドに直結すると考えられる属性である。

【意識的リンク生成/無意識的リンク生成】そして、outdegree 図を再現するにあたって無意識的リンク生成の存在を外すことはできない。無意識的リンク生成は、成熟したウェブ空間と未成熟なウェブ空間から得られるそれぞれの outdegree 図に現れる違いの 1 つとして挙げられる五月雨的構造に關与する属性であると考える。

【ウェブページ生成方法】ウェブページの生成方法によってリンク構成の様相が変わっていくことが予想される。それには、ツールを用いた生成方法ではリンクの生成や情報発信が容易になりウェブページ数やリンク数が増加し情報の流通度が上昇するため情報流通に關与する属性と考える。

続いて、大量のリンクが構成されるウェブ生成は、五月雨的構造の形成に關与する。ツールを使わない人の手によるウェブ生成は、ベキ乗則に従うとする。

言語や、ccTLD、などの存在比率はベキ乗則に従って存在するものとし、属性を設定し構造体を生成する。言語と ccTLD の關係として、ccTLD の国で用いられる主要言語の存在比率が高いものと仮定して設定する。言語、ccTLD の種類は 5 種類程度で実装を行い、順次拡張を行うものとする。ウェブのリンクを生成するにあたっては以下の方法でリンクしているか、リンクしていないかを判定する。

出次を有する pURL(親 URL) が持つ ccTLD を C_p 、ページ内で使用されている主要言語を L_p 、リンク生成方法を M_p 、ウェブページ生成方法を G_p とする。リンク先となる cURL(子 URL) が持つ ccTLD を C_c 、ページ内で使用されている主要言語を L_c とする。リンク生成方法を表す M_p は、無意識的リンク生成を用いている場合には 1、意識的リンクのみの場合は 0 とする。今回は、ページ作成者が自らページを作成しリンクを構成する場合、プログツールなどを用いウェブページを作成し、ツールを利用したリンク構成をする場合、ホームページビルダーなどで多くのリンクが自動生成される場合の 3 つのパターンを考慮した。ccTLD の一致や言語の一致などによってリンクが構成されるかされないかの確率には乱数を発生させてノイズを含ませる。乱数を発生させる際には式 (1) を用いる。式 (1) は R_{min} で確率の最小値、 R_{max} で確率の最大値を任意に設定し、リンクの容易さを表すパラメータとして調整することができる。リンクの生成され易さについては、ccTLD が異なっている場合に比べて、言語が異なっている場合の方がリンクが生成されにくいと考えられる。そのため、それぞれ一致していない場合のリンクが生成される確率を ccTLD よりも言語が異なっていた場合の方が低く設定する。

$$R(R_{min}, R_{max}) \quad (1)$$

リンクの生成方法は式 (2) のように定義する。無意識的リンクが生成されている場合には 1、意識的リンクのみから構成される場合には 0 とする。この無意識的リンクを持っているページのウェブ空間での存在比率を調整することで、五月雨的構造の再現をできることを期待する。

$$L(m_p) = \begin{cases} 1 & (\text{無意識的リンク}) \\ 0 & (\text{意識的リンク}) \end{cases} \quad (2)$$

ウェブページ生成方法は式 (3) のように定義する。今回は、ページ作成者が自らページを作成しリンクを構成する場合、プ

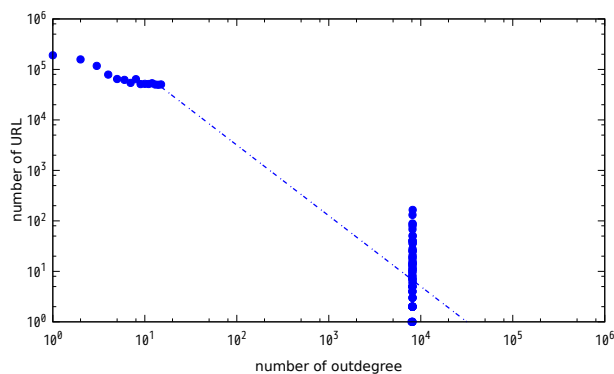


図 1 シミュレーションによる outdegree 図

ログツールなどを用いウェブページを作成し、ツールを利用したリンク構成をする場合、ホームページビルダーなどで多量のリンクが自動生成される場合の 3 つのパターンを考慮した。

$$B(G_p) = \begin{cases} 2 & (\text{多量リンク構造で作られる}) \\ 1 & (\text{プログツールなどを用いて作られる}) \\ 0 & (\text{人の手のみで作られる}) \end{cases} \quad (3)$$

以上の手法から求められた、ccTLD、言語、リンク生成方法、ウェブページ生成方法などの關係からそれぞれのリンクする確率を求める。実際にこれらの確率を等価で扱っているのか議論の余地があるが、今回はこれらの確率を等価であると見なし、平滑化することでリンク生成の判定を行うとする。

以上の手法でパラメータを調整して得られる outdegree 図の形状の変化を観察することでどのパラメータがデジタルデバインドに關与しているのかを明らかにすることができる。

4 まとめと今後の課題

本稿では、ウェブ空間の成長に伴ってウェブ空間から得られる情報量に格差が生じると考え、ウェブ空間の成長度合いによってデジタルデバインドを見ることができるとした。そして、ベキ則、水平構造、五月雨的構造といった outdegree 図の形状がウェブ空間の成熟度を示す指標として捉え、シミュレーションを行うことでこういった要素がこれらの outdegree 図の特徴に作用するのかを示すことでデジタルデバインドの指標とすることができるということを示した。本稿では、シミュレーションの基礎を構成するために簡易的なシミュレーションに留めた。今後の研究では、シミュレーションに用いた擬似的ウェブページに持たせる属性を増やし、パラメータの調整を行うことで、こういった要素がデジタルデバインドに關与しデジタルデバインドを示す指標として用いることができるのかを示していく。

参考文献

- [1] 上嶋 智大: デジタルデバインドの評価指標についての一提案, FIT2007(第 6 回情報科学技術フォーラム), p481 p484, 2007.
- [2] 福田 健介: 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~, 情報処理学会研究報告, 136 号, p17 p22 2004,
- [3] 石原 直幸: ccTLD を単位とした Web コミュニティ構造の分析, FIT2007(第 6 回情報科学フォーラム), p115 p118, 2007.
- [4] 三上 喜貴, 中平 勝子, 児玉 茂昭: 言語天文台から見た世界の情報格差, 慶応大学出版会, 2014.