論文の内容の要旨

非有基的集合論を用いたウェブ構造の分析

堀江 郁美

一貫性を保つことは、ウェブサイトのユーザビリティを向上するためにもっとも重要なものの一つであると言われている [1]. これは，読者はサイトを渡り歩いた経験則に従って直感的に行動することが多いため，ウェブサイトが一貫性を保つことによって読者の航行を助ける必要があるからである．しかし，データ量の増大に伴い一貫性の維持が難しくなり，ウェブサイトのリンク構造が難解になる傾向が出てきた．この一貫性に従わないリンクや，この様なリンクを持つページを不規則構造と呼ぶ．このような不規則構造があるウェブサイトの場合，読者は今どこを読んでいるのか，次にどこへ進めばいいのかがわからない様な状態に簡単に陥いる可能性がある．ウェブサイトには各ページを辿ってみるまで構造がわからないといった特徴があるため，読者が迷うことのないウェブサイトを作成するためには不規則構造のないウェブサイトが必要となる．そこで，本研究では不規則構造の検出を目的とした．

本研究ではウェブサイト作成中，または，作成後に，ウェブのリンク構造を分析し不規則構造を探し出す手法を提案している．前もって構造を決めてしまうのではなく、ウェブサイト作成後である理由としては次の 2 点があげられる．ウェブサイトに適した構造は ウェブサイトの内容によって定まるために，書いているうちに適した構造が変化してしまうことがある．また，複数の著者でウェブサイトを作成する場合，各著者によって適していると思う構造が違うために，共通の構造を限定することができない可能性がある．

Broder ら [2] はウェブをグラフとみなして分析を行ったが，対象はインターネット全体であり，ウェブサイトの構造については議論していない．Wang と Liu [4] は ウェブサイトの構造の出現頻度によって典型的な構造を調査したが，不規則構造の検出は行わなかった．Botafogo と Shneiderman [3] は “Lost in Hyperspace” 問題解決のためにハイパーテキストの構造分析を行った．しかし，数値的指標を用いて分析を行ったために，求める構造や数値的指標の特定を前もって行わねばならなかった．

本研究では不規則構造発見のために集合論の外延性の公理に着目し，ウェブサイトを集合を用いて表現する方法を採用した．外延性の公理を用いることにより，同じ構造を持つページを

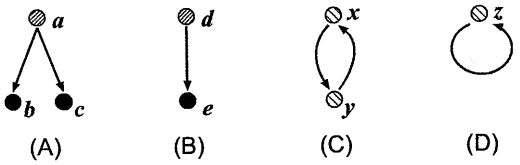

図 1: 集合の標準的なグラフ表示

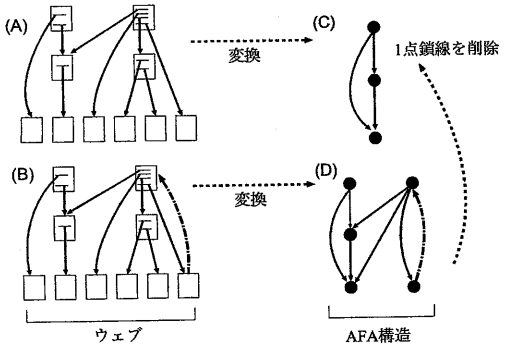

図 2: 不規則構造とその AFA 構造への影響

同一視することができるため，分析対象となるウェブ構造自身を基準にして分析が行えるようになるからである．この特徴のおかげで，ウェブサイト作成後に基準となる構造や値を必要とせず，不規則構造の発見ができるようになった．

しかし，ウェブは循環構造を含むので，基礎の公理によって循環を明示的に禁止している従来の集合論を採用することができない．それで，本研究では循環を認める反基礎の公理 (AFA)[5, 6] に基づいた非有基的集合論を採用した．

非有基的集合論では，集合は頂点によって，$b \in a$ の所属関係は弧 $a \to b$ とみなすことによって，所属関係を有向グラフで表すことが多い．グラフの頂点は AFA に従って矛盾が起こらない範囲でできるだけ同一にされる．この同一視された後の構造のことを AFA 構造と呼ぶ．

例えば，図 1(A) の集合 a, b, c は，$a = \{b, c\}$ と記述することができるが，集合論では $b(= \{\})$ と $c(= \{\})$ は等しくなる．要素のない集合である葉は，$\emptyset(= \{\})$ と一致する．図 1(B) の集合 d，e はそれぞれ $d = \{e\} = \{b\} = a$，$e = \emptyset = b = c$ となり，図 1(C) は，AFA の定義により $x = y$ となる．図 1(D) の z は，図 1(C) の x, y に等しくなる．これは最も単純な循環集合であり，Ω と呼ぶ．本研究では，ウェブを AFA 構造に変換したものに対して分析を行う．

不規則構造を持つウェブサイトは非有基的集合論で表すと頂点数の多いより複雑な AFA 構造になる傾向がある．例えば，図 2(A) は不規則構造を持たないウェブサイトであり，図 2(B) は典型的な構造に従わないリンク (破線リンク) を持つウェブサイトである．これらの AFA 構造 はそれぞれ，図 2(C)，図 2(D) であり， 不規則なリンクを持つウェブサイトの AFA 構造 (図 2(D)) は，不規則構造を持たないウェブサイトの AFA 構造 (図 2(C)) より頂点数の多い複雑なものとなる．本研究ではこの AFA 構造の性質を基に不規則構造を探し出す手法として，簡約度分析 (図 3) と高階ランク分析 (図 4) を提案した．

簡約度分析は，弧検出，弧選択，簡約化の 3 つの操作からなり (図 3)，典型的な構造に従わないリンクを不規則構造として検出する．AFA 構造から 弧を一本抜くことによって AFA 構造 が単純化される場合，その弧を不規則構造の候補であると定義する．この様にして弧検出

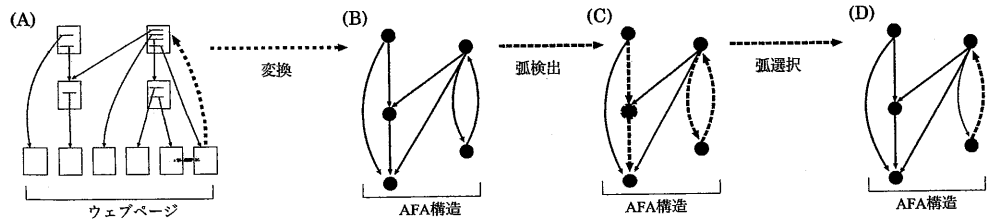

図 3: 簡約度分析: 弧検出, 弧選択

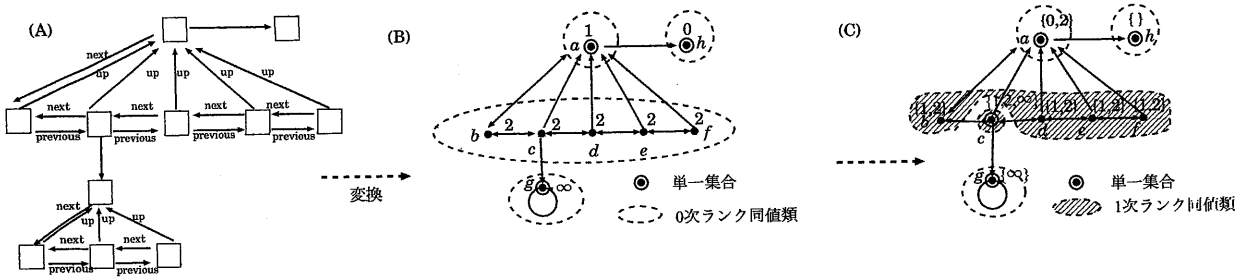

図 4: 高階ランク分析：ウェブ，0 次ランク, 1 次ランク

は不規則なリンクの候補を探し出す．図 3(B) の AFA 構造の場合，点線で示された 4 つの不規則構造の弧の候補 (図 3(C) を見付け出した．そこで数を絞りこみ不規則構造を見つけ出す方法として弧選択を考案した．弧選択では，各不規則構造の候補弧を削除した時に等しくなる頂点に着目した．そこで，弧選択では，他の弧を削除した時に等しくなる頂点を含み，さらに，等しくなった頂点数が最大のものを不規則構造として選び出す．そして，弧選択で選択された不規則構造を削除し，再度同一化を行うことにより，簡約化された AFA 構造を作成する操作を簡約化と呼ぶ．

高階ランク分析 (図 4) では，各頂点の周りの構造を示す指標である高階ランクを用い，不規則構造を検出する．リンクを辿って，そこから先がない葉と辿るべき弧を持つ頂点は構造上異なると考えられる．また，子として葉を持つ頂点と，葉を子として持たない頂点も構造上異なる．この構造上の違いを表す指標として，頂点から葉までの最短距離を示すランク[1]を導入する．図 4(A) に示したウェブサイトの AFA 構造は図 4(B) のようになる．頂点の右上の数字は 0 次ランクを表す．弧を持たない頂点 h のランクは 0，ランクが 0 の頂点を持つ頂点 a のランクは 1，ランクが 1 の頂点を持つ頂点 b, c, d, e, f のランクは 2 となる．弧を持たない頂点へパスを持たない自己参照型の頂点 g のランクは ∞ となる．この例から，同じランクを持つ頂点の子は，同じランクを持つことがわかる．

ランクは，組み合わせることでさらに頂点を分類することができる．ランク 1 の頂点は，必ずランク 0 の頂点を子として持つが，その他にランク 1 やランク 2 の頂点を子に持つことができる．ランク 1 の頂点を子に持つ頂点と，ランク 1 の頂点を子に持たない頂点を区別するために，高階ランクを導入する．同じ n 次ランクを持つ頂点の集合を，数学的用語を用いて n 次ランクの同値類と呼ぶ．

本研究では不規則構造を見つけるために，ランクの同値類が分割される過程に着目した．同じ構造を持つ頂点は同じランクを持つが，次数をあげることによって子のランクの影響を受け，要素数が 1 の単一集合が分離されることがある．単一集合は他に同じ構造を持つ頂点がな

[1] 子孫に葉を持たない頂点のランクは ∞ とする．

いことを意味する．よって，本研究ではこの高階ランクの同値類として得られる単一集合の頂点は，不規則構造と関係があると仮定し，この仮定のもとで 不規則構造を発見した．図 4(C) では，a, c, g, h が不規則な頂点となる．

本研究ではこれら 2 つの分析をそれぞれ実際に使用されているウェブサイトに適用し，それぞれで不規則構造の抽出に成功した．選び出された不規則構造に該当するリンクやウェブページを調べたところ，著者が単なるミスで間違って張ってしまったリンクや，著者の趣味やウェブサイトの内容に影響を受けて作成された異質なリンク構造[2]などがあった．また，サイドメニューを持つ典型的な構造と異なり，サイドメニューを持たないページや，サブカテゴリーのトップページ，スタイルが確立される前に作成された古い資料なども見つかった．また，二つの分析手法はどちらも不規則構造を検出するが，高階ランク分析の方が簡約度分析に比べて効率がよく，簡約度分析は高階ランク分析で検出できない不規則構造間の関係が検出できたり，構造の概略図を作成できることがわかった．

以上のように，ウェブサイトを非有基的集合論で表現し，不規則構造を発見する方法として，簡約度分析と 高階ランク分析を提案し，それぞれの手法を実際のウェブサイトに適用し，不規則構造を発見することに成功した．本研究では対象となるウェブ自身を基準として用いたために，ウェブサイト作成後に，あらゆるウェブサイトの構造に対して不規則構造を発見できるようになった．この結果は従来の構造分析手法では得られないものであり，画期的な成果である．

参考文献

[1] Jacob Nielsen, "Designing Web Usability: The Practice of Simplicity," New Riders Publishing, ISBN 1-56205-810-X, 1999.

[2] Andrei Broder, Ravi Kumar, Farzin Maghoul, Prabhakar Raghavan, Sridhar Rajagopalan, Raymie Stata, Andrew Tomkins, Janet Wiener, "Graph structure in the web", In Proceedings of the 9th International World Wide Web Conference, pp. 247-256, 2000.

[3] Rodrigo A. Botafogo and Ben Shneiderman, "Identifying aggregates in hypertext structures," Hypertext'91 Proceedings, pp.63-74, 1991.

[4] Ke Wang and Huiqing Liu, "Discovering Typical Structures of Documents: A Road Map Approach," SIGIR'98, pp.146-154, 1998.

[5] Keith Devlin, "The Joy of Sets: Fundamentals of Contemporary Set Theory," Springer Verlag, 1993.

[6] Peter Aczel, "Non-well-founded Sets: CSLI Lecture Notes Number 14," Stanford, 1988.

[2]例えば，サブサブシーケンスや，階層的構造とシーケンシャル構造の合併版などであった．