高階ランクを用いたウェブ構造の分析

堀江郁美[†] 山口和紀[††] 柏原賢二[††]

ウェブサイトにおいて，一貫性のない構造は読者を混乱させるということが指摘されている．そこで，本研究では一貫性のない構造を発見する方法として高階ランク分析を提案する．高階ランク分析は，ウェブを有向グラフと見なし，非有基的集合論を基にした高階ランクを用いて，ウェブサイトから一貫性のない構造を発見するものである．この高階ランク分析の有効性を確かめるために，4 つのウェブサイトに高階ランク分析を適用し，各々のウェブサイトにおいて一貫性のない構造として発見されるページが特殊なものや誤ったものであることを検証した．

Higher-order Rank Analysis for Web Structure

Ikumi Horie,[†] Kazunori Yamaguchi[††] and Kenji Kashiwabara[††]

An irregular structure that differs from the typical structure in a Web site might confuse readers, thus reducing the effectiveness of the site. In this paper, as a method for detecting such irregular structures, we propose higher-order rank analysis. In the analysis, viewing the Web as a directed graph and employing a higher-order rank based on the non-well-founded set theory, we are able to detect irregular structures differed from the typical structure of a target site. To test the effectiveness of our method, we applied it to several Web sites in actual use, and succeeded in identifying irregular structures within the sites.

1. はじめに

ウェブ構造に一貫性がないということは **Lost in Hyperspace 問題**[1),2)] の原因の 1 つとされている．これは，読者は予想に反するような構造があると，すぐに自分がどこにいるか，次にどこにいけばいいのか分からない状態に陥ってしまう可能性があるからである[3)]．

図 1 に一貫性のない構造の例を示す．図 1 のウェブ構造には，5 つのページとリンクからなるウェブの部分構造が複数あるが，一番右のウェブ構造は鎖線のリンクがあり他と異なるので一貫性を乱している．このウェブ構造の一貫性をなくしている部分構造を**不規則な構造**と呼ぶことにする．たとえば，図 1 の例では鎖線のリンクとそのリンク元のページが不規則な構造である．ここでは不規則な構造については直感的な理解にとどめ，不規則な構造として具体的に定義されるものについては 2.5 節および 2.6 節で言及することとする．

† 津田塾大学
Tsuda College
†† 東京大学
The University of Tokyo

オーサリングツールには，構造があらかじめ決まっていれば構造の一貫性を保つことができるものがある．しかし，最適な構造と内容は互いに補いあうべきものであり，内容が増えるに従って構造も発展して複雑になることが多いので，前もって構造を決めるのは難しい[3)]．よって，ウェブサイトを作成した後で，不規則な構造を探す方法が必要となる．そこで，本論文では不規則な構造を発見する方法を提案する．

本論文では，リンク構造の表現にはページを頂点，リンクを弧で表した有向グラフを採用した．そして，不規則な構造を見つけるために，各頂点付近の構造を表す指標である高階ランクを導入し，高階ランクに基づいた分析方法として高階ランク分析を開発した．我々はこの高階ランク分析を，Google Japan，Yahoo! JAPAN，外務省，首相官邸の 4 つのウェブサイトに適用し，この方法が有効であることを示す．

先行研究との関連は以下のとおりである．

Broder ら[4)] はウェブのリンク構造から，ウェブ全体の大まかな構造が蝶ネクタイ型のグラフであることを発見した．この研究で用いた手法はウェブ構造全体の特徴をとらえることはできるが，本研究のように不規則な構造を検出することはできない．

Botafogo ら[1),2)] はリンク構造を用いて中心となる

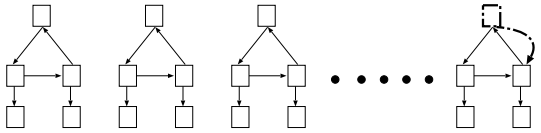

図 1 不規則な構造
Fig. 1 Irregular structure.

ページや階層構造を表す数値的な基準を作成した．McEneaney [5),6)] はハイパーテキストのユーザの動きを数値的な基準を用いて分析した．Botafogo らや McEneaney は数値的基準を用いてウェブサイトの連結性や階層性を評価した．これに対し，本研究はあらかじめ基準を与える必要がない点が特徴である．

Amitay ら[7)] は構造的パターンを用いてサイトを自動的に分類する方法を提案した．この研究はウェブサイトの機能を分類する方法の提案であり，提案された方法でウェブサイトの不規則な構造の発見に利用することはできない．

Chen [8)] は入出次数を用いたリンク構造の類似性の分析を行っている．Weiss ら[9)] は階層構造型サーチエンジン HypurSuit で共通の祖先・子孫を考慮に入れた 2 つの文書の類似性を計算している．これに対し本研究では各ページからすべての子孫までの構造を考慮に入れた類似性を用いて分析している．

Wang ら[10)] は半構造データをラベル付きデータモデル OEM で表し，ユーザが指定した頻度以上に現れるラベルの構造を見つける方法を提案した．我々の研究はラベルがない場合でも利用でき，頻度の指定も不要である．

本論文で提案する方法は簡約度分析[11),12)] の改良版であり，高階ランク分析[13)] を詳細化したものである．

本論文の構成は次のとおりである．2 章では基本概念と方法を説明する．3 章では，実際のウェブサイトを用いた実験結果を示し，その結果により我々の方法の評価を行う．最後に，4 章で結論を述べる．

2. ウェブ構造の高階ランク分析

本章では，我々の分析で用いる基本概念を説明する．2.1 節でウェブを有向グラフで表す方法，2.2 節では高階ランクと呼ぶ，各頂点の周りの構造を表す指標を導入する．高階ランクは，2.3 節で説明する AFA とよばれる非有基的集合論を基にしたものである．2.4 節で AFA を求めるアルゴリズムの基礎となる最も粗い安定分割について説明する．高階ランクを用いて，不規則な構造を持つ頂点を発見する高階ランク分析は 2.5 節で紹介する．2.6 節では，以前提案した簡約度分析[11),12)] を紹介し，2.7 節で，簡約度分析と比較して高階ランク分析が優れていることを示す．

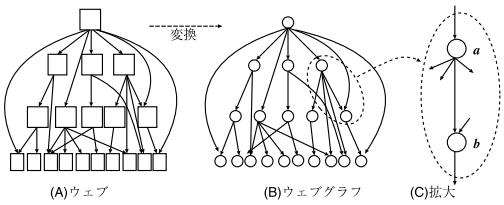

図 2 ウェブからウェブグラフへの変換
Fig. 2 Web and Web graph.

2.1 ウェブのグラフ構造

本論文ではウェブのリンク構造にのみ注目し，ページ内のその他の情報（文字，記号，画像など）は無視する．これにより，ウェブを，ページが頂点で，リンクが弧である有向グラフと見なすことができる（**図 2**）．ここで，頂点の集合，弧の集合は有限集合とする．このグラフを**ウェブグラフ**と呼ぶ．我々の分析は，ウェブをウェブグラフに変換したものに対して行う．

本論文では以下のグラフ理論の用語を用いる．グラフに対して，N を頂点の集合，$R \subseteq N \times N$ を弧の集合とする．頂点 a から頂点 b への弧 $(a,b) \in R$ は $a \to b$ と表し，b を a の**子**，a を b の**親**と呼ぶ．たとえば，図 2 においては，頂点 a は頂点 b の親で，頂点 b は頂点 a の子である．子のない頂点を**葉**と呼ぶ．

2.2 高階ランク

リンクをたどって，そこから先がない葉と，たど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を提案する．

頂点 a の**ランク**を $\mathrm{rank}(a)$ と表し，$\mathrm{rank}(a)$ を次のように定める．子孫に葉を含む頂点 a の $\mathrm{rank}(a)$ は，頂点 a から葉までの最短の有向パスの長さとする．子孫に葉を持たない頂点 a に対しては $\mathrm{rank}(a) = \infty$ とする．

例として，ランクを頂点の周りに書いたものを**図 3** に示す．図 3 (A) のように，葉のランクは 0，葉を子として持つ頂点のランクは 1 となる．子をたどって葉まで到達できない頂点のランクは ∞ となる．ランク 1 の頂点とランク ∞ の頂点を子に持つ頂点のランクは 2 となる．このようにして，すべての頂点に対してランクが定まる．

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，そのほかにランク 1 やランク 2 の頂点を子に持つことができる．そこでランクを組み合わせた k 次ランクを導入して，さらに頂点を分類することを考える．

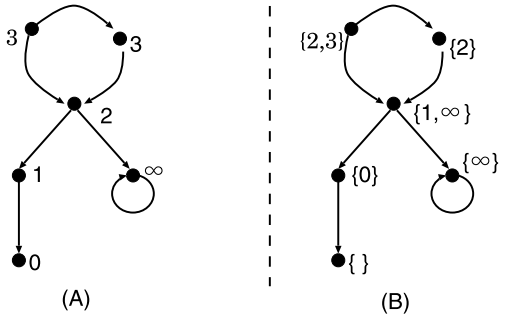

図 3 高階ランク：ウェブグラフの rank_0(A) と rank_1(B)
Fig. 3 Higher-order rank: rank_0(A) and rank_1(B) on a Web graph.

0 次ランクは上で定義したランクで，1 次ランクは子の 0 次のランクの集合とする．たとえば，ランクが 0 と 1 の頂点を子に持つ頂点の 1 次ランクは $\{0,1\}$，ランクが 1 と 2 の頂点を子に持つ頂点の 1 次ランクは $\{1,2\}$ となる．図 3 (B) のランク 2 の頂点の 1 次ランクは $\{1,\infty\}$ となる．2 次ランクについても同様に考える．たとえば，ある頂点が 1 次ランクとして $\{1,0\}$ と $\{2,1\}$ を持つ子を持っていたら，その頂点の 2 次ランクは $\{\{1,0\},\{2,1\}\}$ とする．一般に，k 次ランクを次のように帰納的に定義する．$k>0$ のとき，頂点 $n\in N$ に対して，$\mathrm{rank}_k(n)=\{\mathrm{rank}_{k-1}(m)\mid (\,n\to m\,)\in R\}$，$k=0$ のとき，$\mathrm{rank}_k(n)=\mathrm{rank}(n)$ と定義する．これにより，k 次ランクが $k\geq 0$ に対して定義される．k 次ランク（$k\geq 0$）を総称して**高階ランク**と呼ぶ．

同じ k 次ランクを持つ頂点の集合を，数学的用語を用いて k **次ランクの同値類**と呼ぶ．ここで重要な高階ランクの性質に次のものがある．

命題 任意の k 次ランクの同値類は $k+1$ 次ランクのいくつかの同値類の和集合となる．

証明 紙面の制限のために省略する[14]．

いい換えれば，k 次ランクから $k+1$ 次ランクへ至る過程で，各同値類はそのまま残るか，それより小さい同値類に分割される．この分割過程は十分大きい k で終わり，その k に対する k 次ランクの同値類は高階ランクによって定義された最も細かい同値類となる．このときの k を**ランク最大次数**と呼ぶ．

もう 1 つ重要な高階ランクの性質として次があげられる．

命題 高階ランクによる最も細かい同値類が反基礎の公理（anti-foundation axiom: AFA）[15),16)] に基づいた非有基的集合論（2.3 節）の集合と一致する．

証明 紙面の制限のために省略する[14]．

この命題により，高階ランク分析はウェブグラフよりサイズが小さい AFA 構造に適用すればよいことが保証される．

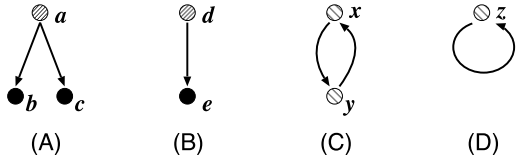

図 4 集合の標準的なグラフ表示
Fig. 4 Sets in standard graphical notation.

2.3 反基礎の公理：AFA

集合論を用いてウェブのリンク構造を分析することによって，同じ構造を持つページを同一と見なして扱うことができるようになる．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一般的な集合論を採用することができない．それで，我々は循環を許す非有基的集合論を用いることにした．我々が採用した非有基的集合論は反基礎の公理（AFA）[15),16)] に基づいた集合論である．本論文では，非有基的集合論といえば AFA に基づく集合論を指すこととする．

集合の所属関係は有向グラフによって表すことができる．このとき，集合は頂点によって，$b\in a$ の所属関係は弧 $a\to b$ によって表現される．このグラフを**所属関係グラフ**と呼ぶ．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．この構造のことを **AFA 構造**と呼ぶ．

図 4 (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 4 (B) の集合 d，e はそれぞれ $d=\{e\}=\{b\}=a$，$e=\emptyset=b=c$ となり，図 4 (C) は，AFA の定義により $x=y$ となる．図 4 (D) の z は，図 4 (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．

ウェブグラフよりも AFA 構造の方が小さく簡単で分析に向いているので，本論文では，ウェブグラフを AFA 構造に変換したものに対して，高階ランクによる分析を適用する．

2.4 最も粗い安定分割

N の部分集合の族 $\mathcal{F}$ を，次の条件を満たすとき N 上の**分割**と呼ぶ．

(1) $\mathcal{F}$ の任意の元 A，B が $A\cap B=\emptyset$ または $A=B$ を満たす．
(2) $\mathcal{F}$ の元の和集合が N である．

分割 $\mathcal{F}$ は分割 $\mathcal{G}$ の任意の要素が $\mathcal{F}$ の要素の部分集合であるとき，$\mathcal{G}$ より**粗い分割**と呼ぶ．

有向グラフ $G=(N,R)$ に関して，$\mathcal{F}$ の任意の元 A，B が，$B \subseteq \{b \mid (a \to b) \in R, a \in A\}$ または，$B \cap \{b \mid (a \to b) \in R, a \in A\} = \emptyset$ を満たすとき，$\mathcal{F}$ を G **に関して安定な分割**と呼ぶ．

命題 任意の有向グラフに関して，最も粗い安定分割が存在する．

証明 紙面の制限のために省略する[14]．

最も粗い安定分割によって誘導された同値関係は AFA による頂点の同値関係と一致する．最も粗い安定分割は頂点数が n，弧数が m のとき $O(m \log n)$ で計算できることが知られている[17]．したがって，上の命題から AFA 構造は $O(m \log n)$ で計算できることになる．

2.5 高階ランク分析

ウェブ構造の不規則性を見つけるために，ランクの同値類が分割される過程に着目する．このアイデアを**図 5** を用いて説明する．図 5 (A) のウェブをグラフに変換すると，図 5 (B) のような AFA 構造となる．

0 次ランクの同値類は $\{a\}, \{b,c,d,e,f\}, \{g\}, \{h\}$ で，そのうち $\{a\}$ と $\{g\}$ と $\{h\}$ が単一集合☆である．0 次ランクの単一集合の同値類の頂点は，その頂点から葉への有向パスに関して同じ構造の頂点が存在しないので，a，g，h を不規則な構造と考える．

次に，1 次ランクの同値類は，$\{a\}, \{c\}, \{b,d,e,f\}, \{g\}, \{h\}$ となる．0 次ランクの同値類の $\{b,c,d,e,f\}$ は，$\{b,d,e,f\}$（1 次ランクは $\{2\}$）と $\{c\}$（1 次ランクは $\{2,\infty\}$）に分割される．よって，$\{b,d,e,f\}$ と $\{c\}$ は，0 次ランクより細かい 1 次ランクで見ると異なる構造を持っているが，$\{c\}$ の方が要素数が小さいので，$\{b,d,e,f\}$ ではなく $\{c\}$ が不規則な構造と考える．

この観測結果より，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な構造と関係があると仮定する．この仮定に基づき，k 次で初めて同値類が単一集合となるときの単一集合の頂点を k **次の不規則な頂点**と定義する．

k 次の不規則な頂点は以下のように求められる．

(1) ウェブサイトを AFA 構造に変換する．
(2) $k=0$ から始めて k を増加させながら，k 次ランクの同値類がすべて単一集合になるまで，以下を繰り返す．
  (a) 各頂点の k 次ランクを計算し，k 次ランクの同値類を求める．
  (b) k 次で初めて生成された単一集合の同値類の頂点を k 次の不規則な頂点とする．

☆ 要素数が 1 の集合を単一集合と呼ぶ．

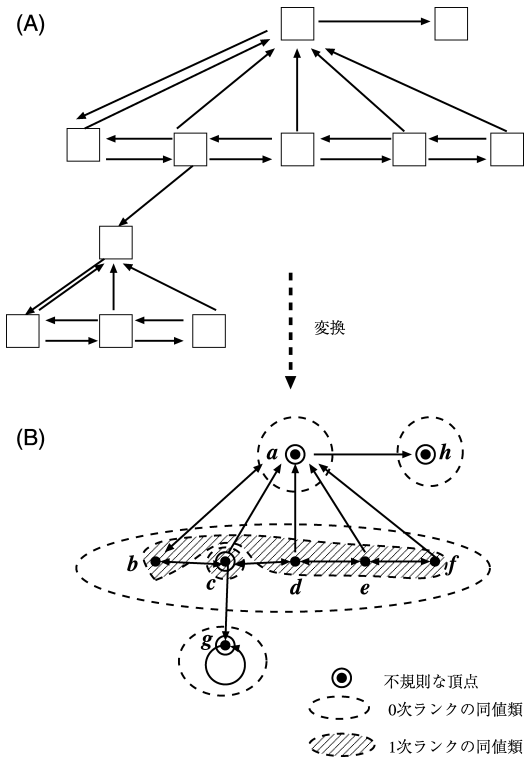

図 5 高階ランク分析
Fig. 5 Higher-order rank analysis.

たとえば，図 5 の場合，頂点 g は 0 次の不規則な頂点，頂点 c は 1 次の不規則な頂点となる．

比較的低次の不規則な頂点を見つけることで不規則な構造を発見する方法を**高階ランク分析**と呼ぶ．

なお，次節で述べる簡約度分析は適用できる対象が階層構造と直列構造に限定されるが，高階ランク分析は任意のグラフに適用できる．

2.6 簡約度分析

高階ランク分析との比較のために簡約度分析[11),12)]を紹介する．

Lynch ら[18] によると，ウェブには直列構造，階層構造，クモの巣構造の 3 つの構造が存在する．3 番目のクモの巣構造には構造上の制限がほとんどないので対象からはずし，**図 6** の直列構造と**図 7** の階層構造を分析の対象とした．ここで，直列構造に関しては，直列構造の AFA 構造もまた直列構造となり，不規則な構造は簡単に見つかる．それで，より分析が難しいものとしてトップページを加えた直列構造を扱うことにした．

図 6 と図 7 に示すように，不規則な構造を持たないウェブは単純な AFA 構造に変換される．逆にいえ

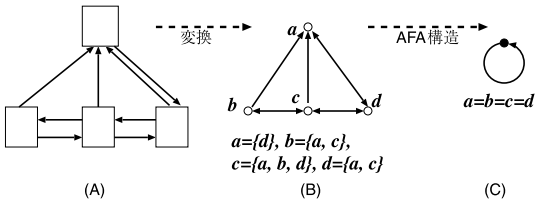

図 6 直列構造のウェブ，ウェブグラフ，AFA 構造
Fig. 6 Linear structure (representation in Web, Web graph, and AFA structure).

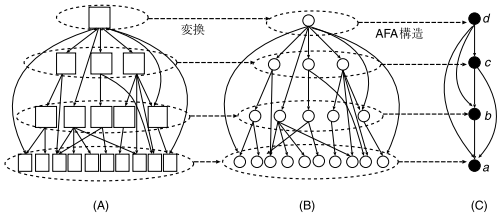

図 7 階層構造のウェブ，ウェブグラフ，AFA 構造
Fig. 7 Hierarchical structure (representation in Web, Web graph, and AFA structure).

ば，不規則な構造が AFA 構造を複雑にしているといえる．誤った弧を削除することによって，AFA 構造の一部が同一となり簡約化された AFA 構造を得ることができる．そこで，簡約度分析では，削除することによって AFA 構造がより単純になる弧を**不規則な弧**と定義し，それを不規則な構造と見なす．たとえば，**図 8** (C) は図 8 (D) より単純なので，それぞれに対応する図 8 (A) と図 8 (B) の差である鎖線の弧を不規則な構造として検出する．

複数の弧が検出された場合は，構造に応じて以下の方法で弧を選択する．

階層構造手法：**図 9** (A) の場合，弧を削除することによって同一化される頂点の集合は $\{\{a,d\}\}$, $\{\{a,d\},\{c,e\}\}$, ... となる．この集合の包含関係 $\subseteq$ に関して，極大の集合を生成する弧を選択する．図 9 (A) の例では，弧 $e \to d$ が選択される．

直列構造手法：図 9 (B) において，同一化される頂点の集合は $\{\{g,h\}\}$, $\{\{g,h,i\}\}$, $\{\{g,h,i,j\}\}$, $\{\{g,h,i,j,k\}\}$, ... となる．この集合の要素間の包含関係 $\subseteq$ に関して，極大の集合を生成する弧を選択する．図 9 (B) の例では，弧 $k \to l$ が選択される．

このようにして選択された候補から，不要な弧を選択して削除する．

2.7 高階ランク分析と簡約度分析の比較

この節では，高階ランク分析と簡約度分析を比較する．

高階ランク分析と簡約度分析はページの内容ではなくリンク構造の分析に焦点をあてている点で同じであるが，高階ランク分析は不規則構造としてページを，簡約度分析では不規則構造としてリンクを発見する点で異なる．また，簡約度分析は対象を直列構造と階層構造に限定しているが，高階ランク分析は対象となる構造を限定していないので優れているといえる．

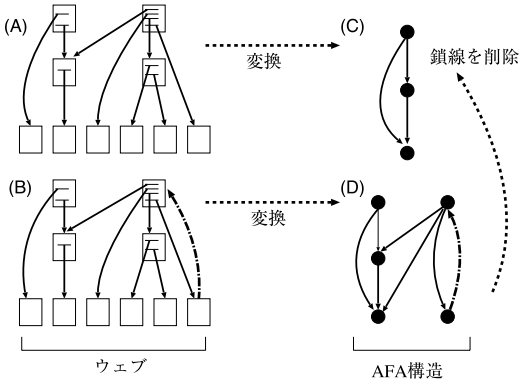

図 8 不規則な弧とその AFA 構造への影響
Fig. 8 An irregular link and its effect on an AFA structure.

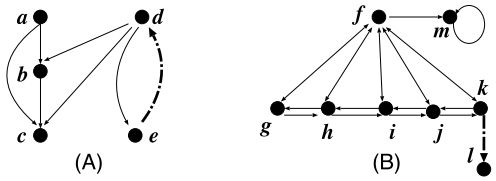

図 9 階層構造手法 (A) と直列構造手法 (B)
Fig. 9 Hierarchical structure scheme (A) and linear sequence scheme (B).

次に，**図 10** の構造を用いて検出能力を比較する．高階ランク分析をこの例に適用すると，図 10 (A) に見られるように，a，c，g，h，i が不規則な頂点として見つかる．図 10 (B) に示すように，簡約度分析の直列構造手法を適用すると，弧 $a \to h$ のみが発見される．高階ランク分析で発見された頂点 a は簡約度分析で発見された弧 $a \to h$ の起点 a として見つかるが，高階ランク分析によって見つかった頂点 c と i は簡約度分析では発見することができなかった．これは，頂点 c（とその弧）が複数の不規則な構造（$c \to i$ と $c \to g$）を持つため，弧を 1 つ削除する簡約度分析では見つけることができないからである．一般に，実際のウェブサイトでは 1 つのウェブページおよび関連するリンクは不規則な構造を複数持つことが多く，それらは，簡約度分析では見つけることができない．高階ランク分析ではそのような不規則な構造も見つけることができる．

簡約度分析と高階ランク分析の計算量を大まかに比較する．ウェブグラフの頂点数を n，弧の数を m とする．安定分割は $O(m \log n)$ で計算できる（2.4 節）．簡約度分析の計算量は，すべての弧を 1 つずつ削除した後に，毎回安定分割を計算するので，$O(m \cdot m \log n)$ となる．これに対し，高階ランク分析の計算量は，子の数の平均を $d = m/n$ とすると，全頂点に対する k 次ランクの計算量は $O(dn) = O(m)$ であるので，k 次まで計算すると $O(km)$ となる．このとき，弧数 m に対して，k のランク最大次数は一般に小さい．我々の実験では弧数とランク最大次数の関係は**表 1** のようになった．したがって，高階ランク分析の計算量 $O(km)$ は簡約度分析の計算量 $O(m^2 \log n)$ に比べて小さい．

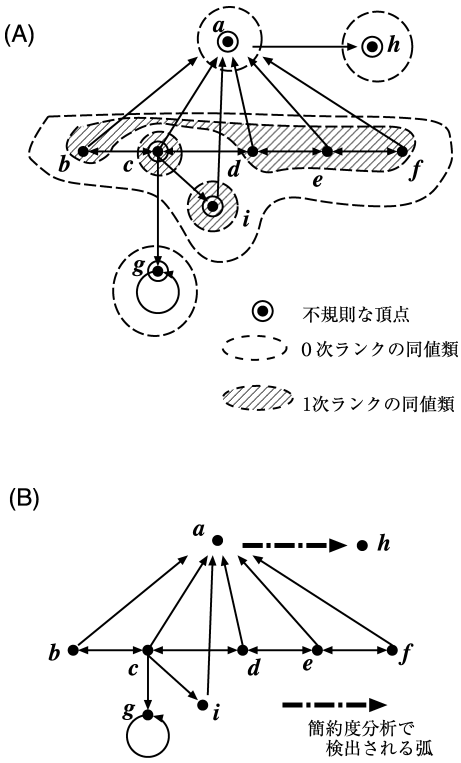

図 10 高階ランク分析 (A) と簡約度分析 (B) の比較
Fig. 10 Comparison of higher-order rank analysis (A) and reduction analysis (B).

表 1 AFA 構造の弧数とランク最大次数
Table 1 The number of arcs and the maximum order of rank.

サイト名	AFA 構造の弧数	ランク最大次数
Google	1,161	13
Yahoo	2,563	7
外務省	105,901	39
首相官邸	18,583	26

上記より，高階ランク分析の方が簡約度分析より次の点で優れている．

- 対象となるサイトの構造に制限がない．
- 検出能力が高い．特に，不規則な構造に関係した弧が複数ある場合，関係する頂点を発見できることがある．
- 計算量が小さい．

3. 適 用 実 験

この章では，4 つのウェブサイトに高階ランク分析を適用した実験を紹介する．

3.1 デ ー タ

本実験では，Google（Google Japan ☆1），Yahoo（Yahoo! Japan ☆2），外務省（外務省公式ページ☆3），首相官邸（首相官邸公式ページ☆4）について分析を行った．Google は，広く使用されているサーチエンジンの 1 つであり，Yahoo は，ディレクトリサービスである．外務省，首相官邸は，日本の政府機関の公式サイトである．

実験ではこれらのサイトから 2005 年 3 月 20 日前後に最大深さ 10 までダウンロードしたデータを用いた．**表 2** に基本属性を示す．すべてのウェブサイトはリンクを持たないページに対応する不規則な頂点を必ず持つが，以下の例ではその不規則な頂点は省略した．

3.2 Google Japan

この例では，典型的な構造と異なる構造を持つページが見つかった．

このサイトには，0 次ランクの同値類が 3 個，1 次ランクの同値類が 5 個あり，これらは不規則な頂点を持たなかった．2 次ランクの同値類は 12 個で，不規則な頂点が 2 個あった．

- 2 次ランクの不規則な頂点のうちの 1 つはリンクを持たない 6 枚のページに対応する．このサイトのほとんどのページがサイドメニューを持つが

表 2 実験に使用したウェブサイトの基本属性
Table 2 Basic figures of Web sites for experiments.

サイト名	ウェブ		AFA 構造	
	ページ数	リンク数	頂点数	弧数
Google	1,630	5,684	136	1,161
Yahoo	12,798	14,076	50	2,563
外務省	36,409	299,064	8,280	105,901
首相官邸	43,449	100,388	3,276	18,583

☆1 http://www.google.co.jp/
☆2 http://www.yahoo.co.jp/
☆3 http://www.mofa.go.jp/
☆4 http://www.kantei.go.jp/

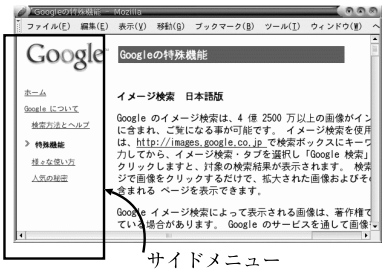

図 11 Google Japan のサイドメニューのある典型的なページ
Fig. 11 Typical page with a side menu at Google Japan.

図 12 Google Japan のサイドメニューのないページ
Fig. 12 A page without a side menu at Google Japan.

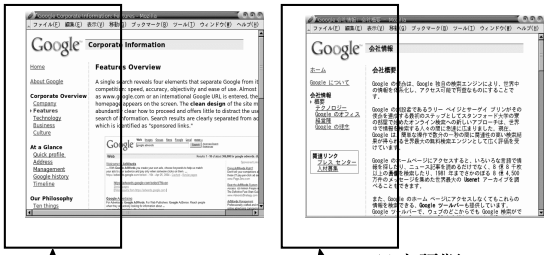

図 13 日本語版と英語版でサイドメニューの構造が異なる例
Fig. 13 The side menu of Japanese pages which is different from that of English pages.

(**図 11**)，これら 6 ページは Google Japan のトップページへのリンクのみで，サイドメニューを持たなかった（**図 12**）.

- もう 1 つの不規則な頂点は 4 枚のページと対応する．これらは英語のページと日本語のページとの間で相異なるリンク構造があるページだった（**図 13**）.

次に，3 次ランクでは 32 個の同値類と 15 個の不規則な頂点が見つかった．これらの不規則な頂点は，サイトマップや問合せのためのページ，サブディレクトリのトップページ，など様々な種類の不規則な構造を持つページに一致した．サイトマップ（**図 14**）はすべてのカテゴリのインデックスにリンクがあるため，問合せのページは問合せのための URL やメールアドレスが列挙されているため，他の典型的なページに比べてリンク数が多かった．また，新機能の宣伝のようなサブディレクトリのトップページは他と違った独自の構造を持っていたため検出された．

図 14 Google Japan のサイトマップ
Fig. 14 Site map at Google Japan.

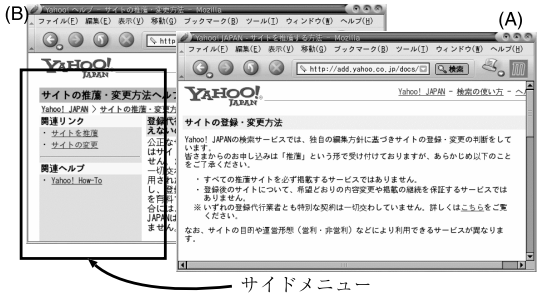

図 15 Yahoo! JAPAN のサイドメニューのない例外的なページ (A) とサイドメニューのある典型的ページ (B)
Fig. 15 An exceptional page without a side menu (A) and a standard one with a side menu (B) on Yahoo! JAPAN.

今回検出された不規則な構造の中でも，たとえばサイトマップのように，特殊な役割を持つことが周知されているために構造が不規則でも読者が混乱しないと考えられるものがある．しかし，この場合でも，読者の意図と異なり，予期せずに適した説明のないサイトマップにたどりついたら読者は容易に混乱するであろう．

3.3 Yahoo! Japan

高階ランク分析がディレクトリサービスのカテゴリのインデックスページを見つける例を示す．

このサイトでは，2 個の 0 次ランクの同値類と 3 個の 1 次ランクの同値類が見つかり，1 次ランクの不規則な頂点が 1 つ見つかった．

1 次ランクの不規則な頂点は Yahoo! Japan のトップページで，他のページよりも多くのページへリンクを持つことから見つかった．さらに，5 個の 2 次ランクの同値類と 1 つの不規則な頂点が見つかった．**図 15** (A) で見られるように，2 次ランクの不規則な頂点に一致するページは他のページ（図 15 (B)）と異

図 16 外務省のトップメニューを持つ典型的なページ
Fig. 16 Standard page with a top menu at Mofa.

図 18 外務省のサブカテゴリのトップページ
Fig. 18 Top page of a subcategory at Mofa.

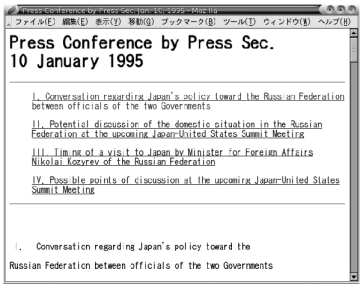

図 17 外務省の古い資料の一部
Fig. 17 Part of an old paper at Mofa.

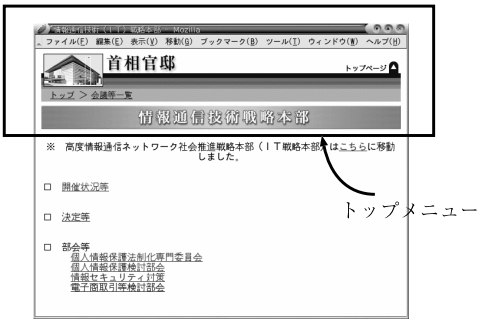

図 19 トップメニューを持つ首相官邸の典型的ページ
Fig. 19 A standard Kantei page with a top menu.

なりサイドメニューがなかった．

次に，3 次ランクの同値類が 10 個あり，3 個の不規則な頂点が見つかった．3 次ランクの不規則な頂点に一致するすべてのページがカテゴリのインデックスページだった．これはカテゴリごとに異なる構造を持つことを意味する．

3.4 外 務 省

この例では，不規則な構造から古いページが見つかった．

外務省のほとんどのページには，**図 16** に示すように，トップメニューを持っており，0 次ランク同値類は 7 個，1 次ランクの同値類は 31 個で，2 個の 0 次ランクの不規則な頂点，4 個の 1 次ランクの不規則な頂点が見つかった．

- 2 個の 0 次ランクの不規則な頂点は 2000 年以前に作成された古い 190 枚のページに対応する．**図 17** にあるように，これらのページは自分自身にのみリンクを持ち，他のページと異なる構造を持っていた．
- 4 個の 1 次ランクの不規則な頂点はそれぞれ 4 枚のページに対応した．それらは，他のページが持つトップメニューを持たなかった．これらの 1 つは 1996 年のサブカテゴリのトップページだった（**図 18**）．

これらのページはこのサイトのウェブ構造が確立される前に作成されたものだと思われる．

3.5 首 相 官 邸

この例では，高階ランク分析により古い資料とカテゴリのトップページが見つかった．

このウェブサイトでは，0 次ランク同値類は 55 個で，25 個の 0 次ランクの不規則な頂点が見つかった．

- 1 つの不規則な頂点はトップメニューを持たない 200 枚のページに対応した．これらのページの構造はトップメニューを持つ典型的なページ（**図 19**）とは異なっていた．
- 24 個の不規則な頂点はそれぞれ 24 枚のページに対応した．これらはナビゲーションのためのリンクを持たない単純な直列構造のページの一部であった（**図 20**）．

これらは以前の首相時代の官邸の古い資料であった．

次に，1 次ランクでは 99 個の同値類と 16 個の不規則な頂点が見つかった．16 個の頂点は 5,043 枚のページに対応する，それらには日本国憲法のページ（**図 21** (A)），子供向けのカテゴリのトップページ（図 21 (B)），そして，情報保護法委員会のインデックスページ（図 21 (C)）などがあり，構造の点で，サイト全体とは独立した方針で書かれているように見ら

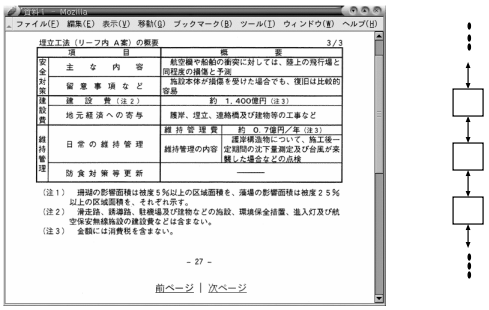

図 20 首相官邸の直列構造の古い資料の一部
Fig. 20 Part of a sequential paper at Kantei.

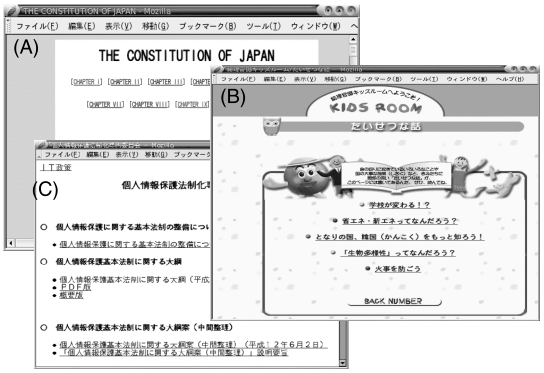

図 21 日本国憲法 (A)，子供への情報のトップページ (B)，小委員会のトップページ (C)
Fig. 21 The page for the constitution of Japan (A), the top page for kids (B), and the index page of a subcommittee meeting (C).

れる．

4. 結　　論

本研究では，ウェブ構造の分析のために，高階ランクに基づく高階ランク分析を提案した．そして，実際のウェブサイトに高階ランク分析を適用し，不規則な構造として異質なページを見つけることに成功した．

今後，この方法を，不規則な構造を修正する案を提示できるように拡張する予定である．また，ハイパーテキストの構造の特徴抽出にも応用する予定である．

参 考 文 献

1) Botafogo, R.A. and Shneiderman, B.: Identifying aggregates in hypertext structures, *3th ACM Conference on Hypertext and Hypermedia*, pp.63–74 (1991).
2) Botafogo, R.A., Rivlin, E. and Shneiderman, B.: Structural Analysis of Hypertexts: Identifying Hierarchies and Useful Metrics, *ACM Trans. Inf. Syst.*, Vol.10, No.2, pp.142–180 (1992).
3) Nielsen, J.: *Designing Web Usability: The Practice of Simplicity*, New Riders Publishing, ISBN 1-56205-810-X (1999).
4) Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A. and Wiener, J.: Graph structure in the web, *Proc. 9th International World Wide Web Conference*, pp.247–256 (2000).
5) McEneaney, J.E.: Graphic and numerical methods to assess navigation in hypertext, *Int. J. Hum.-Comput. Stud.*, Vol.55, No.5, pp.761–786 (2001).
6) McEneaney, J.E.: Visualizing and Assessing Navigation in Hypertext, *10th ACM Conference on Hypertext and Hypermedia*, pp.61–70 (1999).
7) Amitay, E., Carmel, D., Darlow, A., Lempel, R. and So, A.: The Connectivity Sonar: Detecting Site Functionality by Structural Patterns, *14th ACM Conference on Hypertext and Hypermedia*, pp.38–47 (2003).
8) Chen, C.: Structuring and Visualising the WWW by Generalised Similarity Analysis, *8th ACM Conference on Hypertext and Hypermedia*, pp.177–186 (1997).
9) Weiss, R., Ve'lez, B. and Sheldon, M.A.: HyPursuit: a hierarchical network search engine that exploits content-link hypertext clustering, *7th ACM Conference on Hypertext and Hypermedia*, pp.180–193 (1996).
10) Wang, K. and Liu, H.: Discovering Typical Structures of Documents: A Road Map Approach, *SIGIR'98*, pp.146–154 (1998).
11) Horie, I. and Yamaguchi, K.: Structural Analysis for Web Documentation Using the Non-Well-Founded Set, *15th ACM Conference on Hypertext and Hypermedia*, pp.42–43 (2004).
12) Horie, I. and Yamaguchi, K.: Structural Analysis for Web Documentation by the Non-Well-Founded Set, *International Conference on Cybernetics and Information Technologies, Systems and Applications* (*CITSA*), pp.210–215 (2004).
13) Horie, I., Yamaguchi, K. and Kashiwabara, K.: Higher-Order Rank Analysis for Web Structure, *16th ACM Conference on Hypertext and Hypermedia*, pp.98–106 (2005).
14) Horie, I., Kashiwabara, K. and Yamaguchi, K.: Higher-order rank for directed graphs, submitted.
15) Devlin, K.: *The Joy of Sets: Fundamentals of Contemporary Set Theory*, Springer Verlag (1993).

16) Aczel, P.: Non-well-founded Sets: CSLI Lecture Notes Number 14, Stanford (1988).
17) Paige, R. and Tarjan, R.E.: Three partition refinement algorithms, *SIAM J. Computer*, Vol.16, No.6, pp.973–989 (1987).
18) Lynch, P.J. and Horton, S.: *Web Style Guide: Basic Design Principles for Creating Web Sites*, Yale Univ Pr. (2002).

(平成 17 年 10 月 6 日受付)
(平成 18 年 6 月 1 日採録)

堀江 郁美

1970 年生．1994 年津田塾大学数学科卒業．2004 年東京大学大学院広域システム科学系博士課程満期退学．2004 年津田塾大学情報科学科助手．半構造データ，特に Web 構造の分析に興味がある．日本ソフトウェア科学会，ACM 各会員．

山口 和紀（正会員）

1956 年生．1979 年東京大学理学部数学科卒業．1981 年東京大学理学部助手．1985 年理学博士（東京大学）．1989 年筑波大学電子情報工学系講師．1992 年東京大学教養学部助教授．1999 年東京大学情報基盤センター教授．コンピュータのためのモデリング全般に興味を持つ．ACM 会員．

柏原 賢二

1968 年生．1990 年東京工業大学数学科卒業．1995 年同大学大学院システム科学専攻博士課程満期退学．1995 年より東京大学広域システム科学系助手．理学博士．組合せ論，特に集合族や多面体の理論を研究．