研究論文

Web ページ群の構造解析とグループ化

Structural analysis and grouping of Web pages

小島 秀一
東京大学大学院工学系研究科
Shuichi KOJIMA
Graduate School of Engineering, University of Tokyo
高須 淳宏
国立情報学研究所
Atsuhiro TAKASU
National Institute of Informatics
安達 淳
国立情報学研究所
Jun ADACHI
National Institute of Informatics

要旨

Web 上に散在する情報を扱い易くするための手段として，サイト上のページをグループ化するという方法を提案する．意味的に関連した文書をひとまとまりにすることにより，サイトの全体像をユーザへ提示することなどが可能となる．従来の文書の自動分類などでは文書間の類似度を利用して処理が行われているが，本手法ではページ間のリンク構造に着目してサイト内のページ集合を Web グラフとみなし，強連結成分をグループとして抽出することを試みている．またグループは階層的な構造をしているので，その階層構造を抽出するために強連結成分の分割を行っている．

ABSTRACT

In order to easily cope with scattered information on the Web, we propose a method of grouping Web pages on a site. This method allows us to decompose the whole structure of a site by considering semantically related pages as one virtual document. In this paper, we describe the proposed grouping method of Web pages based on the link structure between pages without using similarity between documents which is utilized by traditional text categorization. We consider a Web page set as a Web graph and try to extract strongly connected components as groups. Next, because groups comprise hierarchy structure, we divide a strongly connected component to extract the hierarchy structure.

[キーワード]

グループ化，Web グラフ，階層構造，強連結成分

[Keywords]

grouping, Web graph, hierarchy structure, strongly connected component

1　はじめに

Web 上には大量の情報が氾濫しており，そこから効率よく情報を得るための工夫が多く試みられている．情報を扱いやすくするための方法として，ページを単独ではなく関連するページ群をひとまとまりのグループとして扱う方法がある．例えば現在のサーチエンジンの様に検索結果をただ羅列するのではなく，ヒットした結果を分類（クラスタリング）して提示することに

より，ユーザはより早く目的の情報に辿り着くことができる．グループ化の利点はこのように必ずしも情報のひとまとまりではないページをある基準によりまとめることによって，ユーザが大量の情報を閲覧することを容易にすることである．

またその他の利点として，大量の情報を予め整理しておくことにより，後で分析などを行う場合にその処理の負担を軽減できるという点がある．本研究では特にサイト内のページ群を前処理としてグループ化し，その結果をナビゲーションシステムや検索システムなどへ応用することを目指している．本稿ではその第一歩としてグループ化ができるかどうか，またそのグループが一つの電子文書と見なせるかどうかなどの点について考察する．

グループとは，Web における存在位置をも考慮に入れた，意味的な関連を持つページ群であるとする．たとえば，Yahoo!の掲示板のサイトにおけるあるトピックに属するページ群は一つのグループと見なせる．また複数のトピックをまとめたカテゴリに含まれるページ群も一つのグループと見なせる．このようにグループによりまとめられる集合の単位には様々な粒度が考えられ，階層的な構造を成していると考えられる．

従来このような意味的に関連するページ群を見つけ出す手法としてはテキストの類似性を用いた方法が一般的であった．しかし Web 文書を対象とした場合，ハイパーリンクが意味の繋がりを表しており，そのような文書間の関係を利用した方がグループ化には適していると考えられる．本稿では主にこのリンク構造を利用したグループ化の手法について提案し，その評価を行う．

本稿の構成は，まず 2 章で関連研究について述べたあと，3 章においてグループ化の目的とグループの定義について述べる．次に 4 章と 5 章でグループのリンク構造についての考察とグループ化手法の提案を行い，6 章において評価方法の検討と実際の評価結果について述べる．そして 7 章で今後の展望について述べた後，8 章でまとめを述べる．

2　関連研究

Web ページ群をグループ化する研究として，風間らの研究[1]がある．ここでは同一ディレクトリ内の文書を一つのグループとみなし，そのページ群のインディックスページをファイル名や他ページグループからの被リンク数などにより決定している．しかし同一ディレクトリの文書をグループとみなすと，その内部に存在するグループを抽出できず，また多くのディレクトリに跨がるグループなども見つけることができない．なお，後述のように本研究でもディレクトリ情報は用いるが，あくまでリンク構造による解析が基本であり，両者を組み合わせることによる効果的なグループ化を目指している．

また Web グラフを解析することにより，関連するページ群を発見する手法として Trawling[2]がある．この手法はあるトピックに関連するページ群は，少なくとも一つの完全 2 部グラフを含むという仮定に基づいている．またこの Trawling の手法を，Web 全体ではなく特定のジャンルに絞って関連するページ群を発見する手法が村田により提案されている[3]．しかしこれらの手法が対象とするのは Web 全体，もしくはその部分集合であり，このページ群はコミュニティーに相当するものであって，サイト内のグループとは異なる．

Web グラフの解析によりあるトピックに関する有用なページを探し出す手法としては HITS[4]がある．この手法はトピックに深く関連した内容を含む Authority ページと，多くの Authority ページへのリンクを持つ Hub ページというものを定義し，それらの相互関係を利用して各ページの Authority 値と Hub 値を求めることで重要ページを探し出すものである．この手法が対象とする Web グラフはあるトピックに関連したページ群であり，そのページ間のリンクのほとんどは意味的なリンクであると考えられる．つまり各ページからのリンクは，そのページとの関連があって重要度の高いページへのリンクが多いと考えられる．しかし本研究のようにサイト内の全てのページ群を対象とする場合には，サイト構造を構成するための機能的リンクが多数を占めるため，対象とする Web グラフの性質が異なる．

一方 Amento らはあるトピックに関連するページ群を集める際に，Web ページ単独ではなく Web サイトを基本単位として扱うことを試みている[5]．Web ページ単独では分析などを行うのに扱いにくいので，複数のページをまとめて扱おうという発想は同じであるが，対象としているのはあるトピックに関連したページ群であり，また分類の単位はサイト単位であって，本手法のようにサイト内を細かく分類するということは行っていない．

本研究ではサイト内のページ群を対象とし，リンク分析をベースとしてグループ化することを試みている．

3 サイト内のページのグループ化

3.1 グループ化の目的

本稿で論じるグループとは，少数もしくは一人の作成者により作られた，意味的に関連するページ群のことを指す．従って別々のサイトに存在するページにより構成されるコミュニティーとは概念が異なる．グループは単に同じ内容を持つページ群ではない．表面的には別の内容を含んでいる場合でも，ある概念や目的のために意図的に構成されたページ群が存在し，それらは他と切り離してひとまとまりのグループとして見ることができる．またグループは階層的な構造をしている場合もある．例えばある大学のサイトには学部ごとのページ群が存在し，その下に学科ごとのページ群が存在する．さらに学科のページ群の下には，講義に関するものや研究室紹介のページ群が存在するだろう．またあるメーカーのサイトには製品カタログのページ群が存在し，その下に製品のマニュアルのページ群が存在するという例を挙げることができる．

このようにサイト内にはグループとして扱える単位が存在するのが普通であり，グループ化することには以下のような利点が考えられる．

- Web 検索システムにおいて，従来検索の対象となる単位は一つのページであったが，これをグループ単位で行うことで検索性能を向上させることができる可能性がある．またグループを 1 つの文書として扱うことにより，この文書単位の探索，分類等の処理が可能となる．
- Web 検索システムが検索結果をユーザに提示する際に，ページ単位ではなくグループ単位で提示を行うことにより結果の閲覧や把握が容易になる．
- グループを用いることにより自動的にサイトマップを作成することができ，Web ナビゲーションシステムなどに応用できる．

なお本稿ではクラスタリングとグループ化という用語を異なる概念を表すものとして扱うことにする．すなわち，クラスタリングは文書をその内容のみで分類していくものであり，グループ化は文書間の関係も考慮して分割していくものとする．クラスタリングでは全く同じ文書が存在した場合，それらは同じクラスタに分類されるが，グループ化ではその文書が存在する場所により別々にグループ化され得る．例えばあるソフトウェアの違うバージョンについてのマニュアルがそれぞれ存在した場合，個々のマニュアルに全く同じ内容の概要説明のページが存在したとしても，それらは

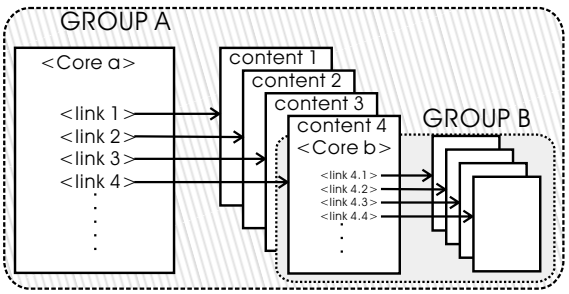

図 1: グループの定義 (1)

別々のグループに分けられるべきである．グループ化とはこのように各ページの Web 空間における存在位置をも考慮したページのまとまりを探し出す処理である．

3.2 グループの定義

ここでグループの定義を具体的にまとめておく．これはグループ化の結果を評価する際の基準となるものである．グループは以下のような特徴を持つ複数のページからなる集合であるとする．

1. 中心となるページ（これをコアページと呼ぶ）があって，そのページのテーマ（トピック）がグループ全体のテーマとなっている．
2. コアページからグループのメンバのみを経由したパスが存在し，かつテーマに合致しているページ（コンテンツ）がグループのメンバである．
3. ドキュメントの補足説明のためにリンクが張られたページもグループのメンバとする．
4. より狭いテーマでページが複数まとまっている場合に，それらをこのグループの下層のグループとする．
5. 各ページはグループ構造の各階層において，いずれか一つのグループに属する．また各グループのページ群は，必ずその上位階層のグループにも属する．

図 1 の例では，コアページ a がグループ A の中心となりページ群を一つのまとまりとして束ねる役割をしている．このコアページから参照されているコンテンツ 4 のページはさらにそれ以下のページを束ねるコアページであり，これらのページ群がグループ B を形成する．

一方図 2 のようにコアページの中でリンクがトピックごとに分類されている場合，それぞれの参照先のページ群はある意味上のまとまりを成していると考えられる．しかしこの様な観点からページをグループ化する

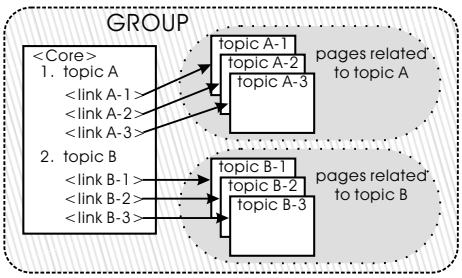

図 2: グループの定義 (2)

と，考えられるグループの構造が非常に多様になってしまい，評価も困難となる．したがって上の定義の 1, 2 によりこれらのページ群は全体で一つのグループであり，それ以下の階層グループは存在しないものとする．

4 グループ化のための基本的アプローチ

4.1 グループ構造についての検討

グループとは前節で述べたようにページのコンテンツとリンク構造の両面から定義されるものであるが，本論文ではコンテンツを見ずにグループ化を試みる．本章ではリンク構造を利用したグループ化の実現のために，グループのリンク構造について考察を行い，典型的なグループの構造と強連結成分との間に深い関わりがあることを述べる．

サイト内のリンク構造は複雑であるが，グループを構成するページ群には典型的な構造パターンがあり，複雑なグループの構造もこれらのパターンの組み合わせが基本となっていると考えられる．サイト全体の構造は，それらの典型的パターンの組み合わせと，それらを結ぶ不規則なリンクにより成り立っているものであると考える．

4.1.1 グループの基本構造パターン

グループのリンク構造の基本パターンとして図 3 のようなモデルを挙げる．

図 3 の (a) の様な構造は，グループの構成要素として最も基本的なものである．このパターンは各コンテンツへのインデックスを保有するコアページを中心とし，そこから各コンテンツページへのリンクにより結合されている．点線で示した各コンテンツページからコアページへのリンクは無い場合もある．また各コンテンツページは，さらに上の階層のノードへのリンクを保有する場合も多い．またコアページは他のグループへのリンクを多く保有していたり，また他のグループから多く参照されている傾向がある．

図 3 の (b) に示す構造は，任意のページ間がお互いにリンクを保有するパターンであり，完全グラフを構成している．このようなパターンの例としては，ニュースを提供するサイトで，各記事のページが関連記事へのリンクをすべて保有する場合などが挙げられる．

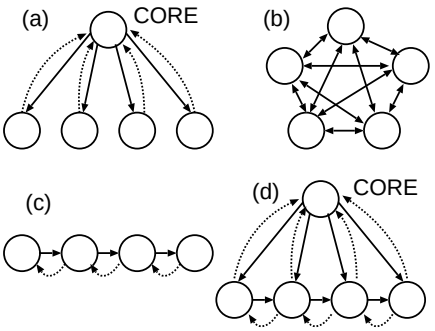

図 3: グループの基本構造パターン

図 3 の (c) はページがリスト状に繋がったものである．このパターンは，例えばスライドのようなページ群である．この図の点線は無い場合もある．

図 3 の (d) は，(a) と (c) の構造の両方の性質を持つものである．これはリスト状に繋がった各ページへのインデックスを保有するページが存在するパターンで，例えばマニュアルや写真集などのページ群で使われる．

これらのパターンに共通していることは，各ページ間に有向閉路が存在する場合が多いということである．つまりグループ内の各ページはリンクを辿ることにより互いに到達可能である場合が多い．有向グラフにおいて，このような各頂点が互いに到達可能である集合を強連結成分と呼ぶ．従ってグループを構成するページ群は多くの場合，強連結成分を構成すると言うことができる．以上の考察より，グループ化の手法として強連結成分に着目し，強連結成分の抽出を基本的なアプローチとする．

4.1.2 グループの階層的グラフ構造

前節のモデルは非常に単純なものであるが，どのような大きさのグループも，これらのモデルを組み合わせたものが基本的な構造となっていると考えられる．とくに注目する構造が図 3(a) の構造である．この構造が強連結成分として抽出できるのは,「目次」などのコアページからのリンクと，各コンテンツページからの「戻る」などのアンカーにより表される逆リンクなどにより，コアページを経由した有効閉路が多く存在すると予想できるからである．

そしてこのパターンにおける各コンテンツページをコアページへ置き換えることで階層的なグループが構

成される．これは各コアページへのリンクを持つさらに上位のコアページが存在して，より大きなグループを構成しているものである．つまり階層的な構造を成すグループの構造は，各階層のグループごとにページを一つにまとめるコアページが存在していて，そのコアページ同士が階層的にリンクをしているものと考えることが出来る．多くの場合，この階層構造もまた強連結成分を成す．階層的なグループ化には，この構造に着目したアプローチを行う．

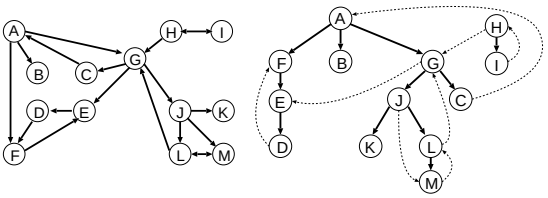

図 4: 深さ優先探索

4.2 強連結成分の抽出

ここで強連結成分を抽出するアルゴリズムについて説明する．このアルゴリズムは，あるノードを起点として深さ優先探索を行うことを基本としている．あるページに複数の子孫が存在する場合には，HTML ファイル内のリンクの出現順序の順に探索を行うこととする．ページ内に同一の URL を指すリンクが複数存在する場合には，一番始めに出現したリンクの出現位置のみを考慮する．深さ優先探索により得られる連結成分は深さ優先探索木と呼ばれ，連結されない複数の木をまとめて深さ優先探索森と呼ぶ．図 4(a) に対して深さ優先探索を行うと，図 4(b) において実線で定義される深さ優先探索森が得られる．(但し A から探索を行った場合には H，I は探索出来ない.）この図において，実線は探索の際にそのノードを初めて訪問した時に利用した辺に対応し，点線はそれの指す頂点が既に訪問済であった辺に対応する．点線の辺には 3 種類あり，上の頂点 (祖先）を指す上向辺，下の頂点 (子孫）を指す下向辺，それ以外の交差辺に分類される．

アルゴリズムの要点は，再帰的な探索の各ステップにおいて，訪問中の頂点 x が次の条件を満たすかどうかをテストすることである．

1. 子孫，上向辺を持たない.
2. x を指す上向辺のある子孫をもち，x より上の頂点を指す上向辺を持つ子孫を持たない.

これは訪問中の頂点 x が，強連結成分をなす集合の一番上の親であるかどうかを判定するためのものである．この条件を満たす時，x と x の子孫から上の 2 つの条件を満たす頂点と子孫を除いたものが強連結成分をなす．このテストは再帰的に行われ，各ステップで条件を満たす成分は切り離されていき，その成分が切り離されたグラフに関してさらにテストが繰り返されていく．

図 4(b) の例では，頂点 B と K は条件 1 を満たし，それら単独で強連結成分をなす．また頂点 F，H，A は条件 2 を満たし，それぞれ {F,E,D}，{H,I}，{A,G,J,L,M,C} の強連結成分の親となっている．

条件 2 を満たす頂点 x の子孫のうち，条件 1，2 を満たす頂点もしくはその子孫ではない頂点 y が，x と同じ強連結成分をなすことは次のように説明できる．x から y へは探索木を下降して辿ることの出来る有向道が存在する．従って y から x への有向道が存在することが言えれば良い．y の性質により，y もしくはその子孫に y より上で x 以下の頂点 z への上向辺を持つノードが存在する．従って y から z への有向道が存在する．z が x と異なる場合は y と同じ性質を有するので，同じように下降，上昇をすることにより x に達することができる．よって y からも x への有向道が存在するので，x と y は同じ強連結成分をなす.

5 階層的グループ化のアプローチ

5.1 強連結成分の分割

Web サイトのグラフに強連結成分の抽出アルゴリズムを適用すると大小様々な強連結成分が得られる．最小のものはページ単独で強連結となっている孤立点であり，大きいものでは数千ページ以上にもなる．ユーザへグループを提示する場合などを考えると，数千ページの塊では扱いにくい．このような巨大なグループは多くの場合，全く関係のないグループが，その相互に張られた数本のリンクのために連結された構造であったり，または幅広いトピックを扱ったページからのリンクが複雑に入り組むことにより，一つの巨大な強連結成分をなしていると考えられる．またグループは階層構造になっているので，上位階層のグループの中からより小さな粒度の下層のグループを抽出する必要がある．ここでは強連結成分の中から，それぞれのグループの切り分けと，グループの中の階層構造の抽出方法について述べる．

グループの階層構造は 4.1.2 節で考察した様に，各階層グループのコアページがリンクで連結した構造が基本になっていると考えられる．したがって，一つの塊

となっているグループを，その下の階層のグループに分割するために有効な手段は，上位階層のコアページへのリンクを切断することである．これにより上位のコアページが強連結成分から切り離される．下位階層のグループはそのグループのコアページを中心にまとまっているので，強連結成分として抽出できることが期待できる．

5.2 コアページの推定

コアページに着目したリンクの切断を行うためには，まずコアページを特定する必要がある．ここではコアページを推定するための基準をいくつか述べる．

(a) 探索順位を利用した方法 サイトのトップページをスタートノードとして探索をした場合に，各グループのコアページはそのグループ内で最も早く探索される可能性が高いという考えに基づく手法である．分割対象の強連結成分の中で，探索木のトップノードのページをコアページとして見なす．

(b) ページの in-degree，out-degree に着目した方法 コアページは各コンテンツ，もしくは下位階層のグループのコアページへのリンクを多く持ち，また各グループのページから参照されている数も多いという仮定に基づく手法である．in-degree（他ページから参照されている数）のみ，out-degree（ページが持つリンク数）のみ，もしくは in-degree と out-degree の和いずれかを用いて，その値が最大のものを強連結成分内のページから捜し出し，それをコアページとする．

(c) ディレクトリを単位としたリンクに着目した方法 これはそれぞれのページから，どれほどの種類のディレクトリへのリンクが存在するかを基準とする手法である．ディレクトリはサイト制作者により意図的にまとめられたページの入れ物であり，何らかの意味的なまとまりがあると考えられる．そこでこのディレクトリを単位とし，多くのディレクトリへリンクを張っているページは，多くのコンテンツへのリンクが存在するということであるので，上位の階層のコアであると見なせる．

(d) ディレクトリの階層構造に着目した方法 各ページがどの程度コアページらしいかを表すために，コア値というものを導入する．より上位階層のグループのコアページの方が，下層のグループのコアページよりもコア値が高いものであると定義する．つまりサイトのトップページのコア値が最大になるものとする．この手法はディレクトリの階層構造を利用し，コア値を計算することによりコアページを推定する方法であり，次のような考えを前提としている．

- より多くのコアページへ参照しているページのコア値はより高くなる．
- ディレクトリ構造における子孫へのリンクは，より詳しいコンテンツへのリンクである．
- コアページから参照されるコンテンツページや下位階層のコアページは，同ディレクトリ，もしくはディレクトリ階層における子孫に存在する．
- ディレクトリの階層が上位のページの方がコア値が高い可能性が高い．

これらの考えに基づき，

$$C(p) = \sum_i \left(D(q_i)/T + C(q_i)\right)^k \qquad (1)$$

によりコア値を求めることにする．ただし $C(p)$ はページ p のコア値，$D(q_i)$ はページ q_i から参照されているページの内，ページ q_i と同ディレクトリもしくは下層のディレクトリに存在するページの数，ページ q_i はページ p の下層のディレクトリに存在するもので，ページ p から参照されているページとする．

ディレクトリ構造の末端に存在するページのコア値はパラメータとして与える．T は目次ページがもつ標準的なリンク数を設定する．T 以下のリンクを持つページは，その上位のページのコア値への貢献度が低くなる．k を 1 より大きく設定した場合，参照しているページ群の中でコア値が大きいページや子孫や同ディレクトリのページへのリンクが多いページの影響が大きくなる．この計算は強連結成分のみに対して行うのではなく，サイトに存在する全ページを対象にして計算を行う．またディレクトリ階層の末端のページのコア値を 0 より大きい値に設定した場合には，その与えた値以上のコア値を持つページをコアページとする．この条件を満たすページは限られており，分割の際にこの条件を満たすコアページが見つからない場合があり得るが，その場合には（b）または（c）の方法でコアページを決定する．

以上，コアページの推定法として (a) から (d) までの手法を述べたが，以下の評価においてそれぞれの有効性の比較を行う．

5.3 リンクの切断方針

コアページが分かれば，コアページへのリンクを切断することによりそのグループの一つ下のグループを抽出することができるはずである．しかし実際にはグループの中のコアページ以外のページが他のグループとリンクで結ばれていて，そのリンクを経由することによりグループ間に有向閉路が存在して強連結成分が分割できない場合がある．そこでコアページの存在するディレクトリ以下のページ群への，それ以外のページ群からのリンクを切断するという方法をとる．順番としてはまず分割対象の強連結成分のグラフからこのディレクトリに基づくリンクの切断を行い，切断後のグラフに対し強連結成分の抽出を行う．これによりこのコアページがまとめるグループが抽出できる．そしてコアページへの他のページからのリンクを切断したグラフに対して強連結成分の抽出を行い，その下の階層のグループを抽出する．

5.4 階層的グループの抽出手順とグループ構造の再編成

以上の切断方法を用いてグループを抽出する手順は以下のようになる（図 5）．

1. サイトのトップページから深さ優先探索を行い，各ページにから参照されている URL を抽出してページごとのリンクリストを作成する．
2. サイトのトップページをスタートノードとして強連結成分の抽出を行う．
3. ある特定の大きさ以上の強連結成分の集合の中からコアページを推定し，そのページの存在するディレクトリ以下のページ群への外部からのリンクを切断する．2 ページ以下のグループ集合はそれ以上分割しても意味がないので 3 以上のページが対象となる．強連結成分抽出の探索ははじめはコアページから開始するが，リンクの切断により探索できないページが現れる可能性があるので，その場合は分割対象のページ群がすべて探索できるまで，探索されなかった任意の点から探索を繰り返す．
4. ステップ 3 により抽出されたグループのうち，コアページを含むグループについては各ページからコアページへのリンクを切断した後，コアページより強連結成分抽出の探索を開始する．このリンクの切断は強連結成分を成すページ群に対して行うので，このリンクの切断を行っても必ずコアページからこのページ群の任意のページへのパスが存在する．

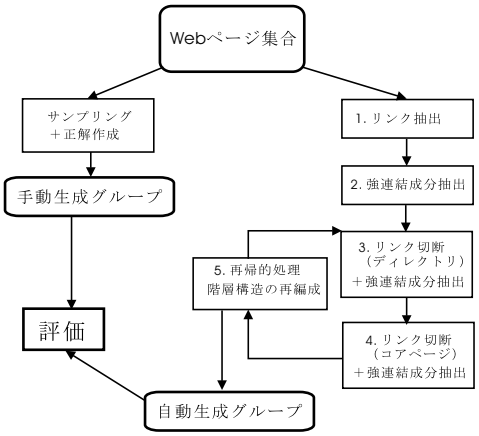

図 5: 処理の流れ

5. ステップ 3，4 を再帰的に繰り返して全てのグループが特定の大きさ以下になるようにする．

以上がグループの抽出の手順であるが，実際のグループの階層構造はこの再帰的手順で得られる通りにはなっていない場合が多い．またあまり深い階層構造にするとユーザへの提示する場合などに問題となる．そこでここでは階層構造に次のような条件を設けることにし，その条件を満たすようにグループの階層構造を再編する．

階層構造の条件 下の階層のグループの大きさは，上の階層のグループの大きさの半分以下になるようにする．ただし下のグループが孤立点のみで構成される場合にはその条件は適用しないこととする．

6 実験と評価方法の検討

6.1 実験と結果の考察

同じホスト名を持つ URL のページを同じサイトに属すものと定義し，特定のサイトに限定してグループ化の実験を行った．実験の条件として，手動で与えたトップページからのパスが存在する*.html，*.htm ファイルのみを解析の対象とした．

まずサイトのグラフ構造を解析して強連結成分がどのように抽出されるかを調べた．表 1 に，それぞれのサイトにおける解析対象のページ数，抽出されたグループ数，孤立点の数，グループ中に含まれる平均ページ数，グループ化率，巨大グループによるページ占有率を示す．これらのグループはサイトの持つグラフそのものを解析した結果得られた強連結成分のことを指しており，リンク切断による階層的なグループ化は行っていない．

ここでのグループ数にはページ単独で強連結成分と

表 1: サイト別の強連結成分の抽出結果

site	#page	#group	#isolated node	average size	grouping rate(%)	large group's share(%)
www.kantei.go.jp	15883	1775	1531	8.95	90.4	68.9
www.waseda.ac.jp	10023	1909	1803	5.25	82.0	69.5
www.toshiba.co.jp	8930	408	366	21.9	95.9	91.6
www.u-tokyo.ac.jp	575	102	101	5.64	82.4	82.4
www.nii.ac.jp	33415	1434	1355	23.3	95.9	91.8

図 6: グループ化の出力例 (1)

図 7: グループ化の出力例 (2)

して抽出された孤立点も含んでいる．グループ化率とはどれくらいのページが孤立点としてではなく，複数のページの集合として抽出されたかを表すものであるとし，次式で計算する．

$$(\text{グループ化率}) = \frac{(\text{全ページ数}) - (\text{孤立点の数})}{(\text{全ページ数})} \times 100 \quad (2)$$

巨大グループのページ占有率は，複数のページにより強連結成分を構成するグループの中で，大きさが上位 1%に入るグループが含むページが全体のページの何%を占めるかを示す値である．これはどれだけ少数のグループがページを占有しているかを見るためのものである．

グループ化率の値を平均すると 89.3%となり，大部分のページが複数ページによる強連結成分をなしていることが分かり，グループ化の手法として強連結成分を採用することは有効であることが分かる．

また巨大グループの占有率は平均で 80.8%であり，ごく少数のグループが大部分のページを占有していることが分かる．これは実際には別々のグループが連結して少数の強連結成分の塊となっていることを表している．このことからも強連結成分を分割することが必須であることが分かる．

6.2 グループ化の出力例

ここではグループ化の結果がどのようなものであるかを紹介する．首相官邸のサイト（http://www.kantei.go.jp/）について行ったグループ化の結果を図 6，図 7 に示す．図 6 の結果は，サイトの Web グラフに対し強連結成分を抽出した結果の一部である．抽出された結果はグループの大きさによりソートしている．また図に表示してある URL はそのグループのコアページの URL を表す．URL の横に括弧で囲ってある数字は各グループのページ数を表す．

図 7 は図 6 のグループを表すフォルダを開いてグループの階層構造の例を示したものである．この図により初めは大きな強連結成分の塊としか抽出できなかったものが階層的に分割されている様子が分かる．

実際にこのグループ化がどれほど適切になされているかについては次節以降で検討する．

6.3 グループ化の評価における問題点

グループ化の評価には，従来の情報検索や文書の自動分類の評価と比べていくつかの問題点がある．従来の情報検索は，基本的には与えられた文書集合の中から検索質問に関連する文書を一次元的なランキングを付けて出力するものである．その評価方法としては，検索すべきものをどれだけ漏れなく検索できたかを表す再現率と，検索したものの中に検索すべきものがどれ

だけ含まれているかを表す精度を用いる方法が一般的である．検索結果が一次元的であるので，精度対再現率グラフを書くことができ，その曲線により性能の良し悪しが判断できる．一方グループ内の文書集合はランキングされているわけではないので，そのようなグラフは描けない．またグループ化では文書集合の大きさ自体が評価すべき対象であり，精度対再現率グラフにおいては，ひとつの点だけが出力されるものであるとみることができる．たがそもそもグループ化の場合には，出力された結果の中のどの文書集合を評価対象とするのかが明確でない．

文書の自動分類では，分類するカテゴリがあらかじめ決まっており，評価すべき文書集合が明確であるため，情報検索と同じように精度と再現率の評価が可能である．

ここでグループ化において評価しなければならない点をまとめる．

1. グループ間の境界

   意味的にまとまりのある文書集合が，どの程度うまく抽出できているかを以下の 2 点の観点から見る必要がある．

   (a) グループの構成要素の精度

   グループ内の文書がどの程度関連した文書の集まりであるかという点に関する評価である．関係の無い文書が紛れている場合には精度が落ちる．またいくつものグループが連結されたままの状態では精度は低くなる．

   (b) グループの構成要素の再現率

   同じグループに分類されるべき文書がどれだけひとつのグループにまとめられているかという点に関する評価である．グループの粒度が必要以上に小さくなってしまった場合には再現率は低くなる．

2. グループの階層構造

   文書の自動分類などとは違い，グループ化では文書の集合を階層的に抽出するのでその階層構造が適切であるかどうかも評価する必要がある．最小単位のグループは，それよりも広いトピックでひとまとまりにできる場合がほとんどであり，そのような上の階層のグループが抽出できていない場合には評価を低くすべきである．逆に大きな単位のグループしか抽出できない場合も同様である．

3. グループのコアページ

   Web 文書のグループには，それらをまとめる中心的役割を果たすコアページが存在するのが普通であり，ユーザにグループの提示などを行う場合にはそのようなページをグループの代表として提示すべきである．したがってこのコアページの推定が出来ているかも評価する必要がある．

6.4 評価方法についての検討

グループ化を評価するにあたり，前節で述べた評価すべき点の 1，2 は密接に関係していて切り離して評価することが難しい．評価するグループの粒度を固定して評価するということも考えられるが，グループ化された文書集合の大きさは様々であり，そのように階層構造を非階層構造に投影することは難しい．例えば最下層のグループを評価の対象にしようとした場合，もしグループ化された 30 ページのうち，5 ページだけがその下の階層のグループとしてまとまっているとすると，その他の 25 ページは孤立点として扱わなければならない．また閾値を導入して，その値以下で最大の大きさのグループを評価の対象とすることで，ある程度の粒度の統制をとることができるが，その閾値の設定をどうするかが問題となる．

そこでここでは，各階層のすべてのグループを評価の対象とする．以下，具体的な方法を述べる．まずグループ化を手動で行い，それを正解グループとする．実際にはサイト内のページを全て手動でグループ化することは不可能なので，その一部分をサンプリングしてそれを正解グループとする．そしてその正解グループの各階層のグループすべてに対し，それぞれ自動グループ化の全グループと比較を行う．この比較は，自動グループ化のグループそれぞれが，精度と再現率の両方の観点からどれだけ正解グループに近いかを調べるもので，

$$F(G_i, g_j) = \frac{1}{\alpha \frac{1}{P} + (1-\alpha)\frac{1}{R}} \tag{3}$$

$$P = \frac{x_{i,j}}{X_j} \quad , \quad R = \frac{x_{i,j}}{x_{i,0}} \tag{4}$$

により行う．ここで G_i は正解グループ，g_j は自動生成された評価対象グループ，$F(G_i, g_j)$ は G_i に対する g_j のスコア，α はパラメータ $(0 < \alpha < 1)$，P は精度，R は再現率 $x_{i,0}$ は G_i に含まれるページ数，$x_{i,j}$ は G_i のうち評価対象グループに含まれるページ数，X_j は 評価対象グループのページ数を表すものとする．

これは情報検索の評価における F 尺度といわれる評価基準である[6]．この式では精度，再現率ともに 1 の場合に最高値の 1 となり，最低値は 0 となる．この式は精度，再現率のどちらかがよいだけでは高いスコアが得られず，どちらも高い場合においてのみスコアが高くなるものなので，グループ間の境界がはっきりしているかどうかの判定基準になり得ると考えられる．またパラメータにより精度と再現率のどちらを重視するかを設定できることも，この式を選んだ理由である．パラメータ α が大きいほど再現率を重視していることを意味し，$\alpha = 1/2$ とすると再現率と精度を同等に扱うことになる．

そしてある正解グループに関し，どれほど自動でグループ化が出来ているかをしめすスコアを，自動グループ化の全グループについて行った比較で得られたスコアの最大値とする．ただし最大値をスコアとすると，自動グループ化が正解付近のグループをたくさん出力している場合には有利となる可能性がある．例えばある階層のグループの大きさと，その上下の階層のグループの大きさが 1 つしか異ならない場合などである．だがこのように必要以上に階層が深いものは，ユーザに提示した場合などにも見にくく，適切な階層構造をしているとは言えない．したがってこのような構造をしたグループ化には低いスコアを与える必要があるが，ここではグループ化に階層構造の制限を加えることにより，この問題が起らないようにしている．具体的には 5.4 のような条件をグループの階層構造に課しているので，このような問題はないと考えられる．

各正解グループについての最終的なスコアは，正解グループのページ数に比例させて各スコアを足し合わせたものとし，次式で計算する．

$$TotalScore = \sum_i (\frac{x_{i,0}}{x_{all}}) max_j F(G_i, g_j) \qquad (5)$$

ただし $x_{all} = \sum_i x_{i,0}$ とする．

6.5 正解作成について

前節で述べた評価を行うためには，自動グループ化の結果と比較するための正解グループが必要となる．そこで手動によってグループを作成し，それを正解グループとする．サイト内のページをすべて手動でグループ化できれば，それが正解としてもっとも相応しいが，首相官邸のサイトの例では 1 万 5 千ページ以上ものページが含まれており，手動で完全な正解を作成する事はほぼ不可能である．含まれるページ数が比較的少ないサイトを選ぶということも可能であるが，グループ化の目的はむしろ首相官邸のサイトのような膨大な量のページを含むサイトのグループ構造を見い出すことにあるので，評価としてはあまり相応しくない．したがってサイト内のある一部のグループを選びだして評価することになる．グループは階層構造になっているが，一部のグループを選ぶということはこの階層構造を木構造と見た場合の部分木を取り出すということである．この取り出す部分木は，その部分木以下の構造が完全な状態で抽出する必要がある．途中の枝が刈られているような状態では，その部分を含むグループの網羅性が完全ではなくなってしまう．次にどれぐらいの大きさの部分木を選び，それをいくつ用意するかが問題となる．今回は約 20 から 150 程度の大きさの部分木を 10 個選んで正解を作成した．各部分木のグループは階層構造をしているので，評価対象となるグループの数は部分木の数よりも多くなる．ここでは首相官邸のサイ

表 2: グループ化正解例 (1)

ID	Group	#page	#group
1	小泉総理の動き	161	17
2	首相官邸バーチャルツアー	18	3
3	首相公選制を考える懇談会	17	5
4	情報セキュリティ対策	60	10
5	ミレニアム・ゲノム・プロジェクト	36	5
6	中央省庁改革 概要資料	19	3
7	小泉内閣総理大臣演説等	60	2
8	小泉内閣閣僚名簿	21	1
9	お答えします	29	4
10	総理大臣官邸整備	19	3

表 3: グループ化正解例 (2)

ID	Group	#page
1	小泉総理の動き	161
1-1	ＡＳＥＡＮ＋３首脳会議	7
1-2	平成 13 年 4 月 26 日〜5 月 31 日	16
1-3	平成 13 年 6 月 1 日〜6 月 30 日	23
1-3-1	日・米首脳会談	6
1-4	平成 13 年 7 月 1 日〜7 月 31 日	26
1-4-1	Ｇ７首脳声明	6
1-4-2	Ｇ８首脳会合	5
1-5	平成 13 年 8 月 1 日〜8 月 31 日	16
1-6	平成 13 年 9 月 1 日〜9 月 30 日	33
1-6-1	テロ対策関係閣僚会議	3
1-6-2	ニューヨーク訪問	4
1-6-3	三宅島視察	4
1-7	平成 13 年 10 月 1 日〜10 月 31 日	29
1-7-1	緊急テロ対策本部（第１回）	3
1-7-2	ＡＰＥＣ首脳会議	5
1-7-3	ＡＰＥＣ参加国との首脳会談	7

表 4: 各リンク切断方法に対する評価結果

group ID	search	in-dgree	out-degree	in + out	directory link	directory's hierarchy
1	0.698	0.589	0.604	0.590	0.418	0.698
2	0.907	0.929	0.929	0.864	0.857	0.929
3	0.915	0.906	0.929	0.911	0.906	0.929
4	0.759	0.759	0.387	0.759	0.301	0.759
5	0.975	0.722	0.745	0.722	1.00	0.745
6	0.832	0.892	0.892	0.876	0.835	0.892
7	0.0995	0.818	0.391	0.387	0.821	0.133
8	0.0909	0.0909	0.0909	0.0909	0.0909	0.0909
9	0.655	0.345	0.651	0.354	0.678	0.356
10	0.985	0.889	0.889	0.889	0.972	0.987
Total	0.700	0.663	0.616	0.623	0.581	0.666

トを評価の対象に選び，正解を作成した．作成した正解例を表 2，表 3 に示す．表 2 は正解に選んだ部分木全体のグループであり，グループのページ数と階層的なグループの数を示している．表 3 はそのうちの一つのグループについての階層構造を示したもので，各グループについてのページ数を示している．

6.6 コアページ推定に関する評価

サイトマップに存在するリンクは，各グループの頂点となるものである場合が多いと考えられる．したがってコアページの推定方法により得たコアページの候補リストとサイトマップに存在するリンクのリストを比較することにより，コアページの推定方法の妥当性を調べる．

まずサイト内のリンクから各グループのトップとなるページを手動で抽出する．このグループ化におけるコアページの推定に関するスコアを，A を手動のコアページ数，コアリストを上位 A 位以内のコアページの候補として次のように定義する．

$$Score = \sum_i s_i / A \qquad (6)$$

$$s_i = \begin{cases} 1 & \text{コアリストにページ } i \text{ が存在する場合} \\ 0 & \text{コアリストにページ } i \text{ が存在しない場合} \end{cases}$$

6.7 評価結果

首相官邸のサイトについて行ったグループ化を，6.4 節で述べた評価方法を用いて評価した結果を表 4 に示す．評価に用いた式 (3) の α の値は $1/2$ とし，精度と再現率を同じ重みで評価した．ここで比較しているのは各コアページの推定方法であり，探索順位，in-degree, out-degree，in-degree と out-degree の和，ディレクトリを単位としたリンク数，そしてディレクトリの階層構造を利用した各手法によるグループ化のスコアを示した．ディレクトリの階層構造を利用した手法で用いる式 (1) では，末端に存在するコアページの初期値を 0.1 に設定し，T， k をそれぞれ 10，2 とした．またこの式によるコアページが見つからない場合には in-degree と out-degree の和が最大のページをコアページとした．トータルのスコアでは探索順位を利用した方法が最も優れているという結果になった．ただし正解グループによっては他の手法が優れているものもあり，常にこ

表 5: グループ 1 に関する評価例

ID	search		
	x	X	score
1	156	172	0.937
1-1	7	7	1.00
1-2	13	172	0.138
1-3	23	6	0.414
1-3-1	6	6	1.00
1-4	6	6	0.475
1-4-1	6	6	1.00
1-4-2	5	5	1.00
1-5	16	172	0.170
1-6	33	172	0.312
1-6-1	3	3	1.00
1-6-2	4	4	1.00
1-6-3	4	4	1.00
1-7	7	7	0.389
1-7-1	3	3	1.00
1-7-2	5	5	1.00
1-7-3	7	7	1.00
total			0.698

表 6: コアページの推定に関する評価

in-dgree	out-degree	in + out	directory link	directory's hierarchy
0.33	0	0.19	0.095	0.29

の方法が有効であるとは言えない．特にこの探索順位を利用する方法は，トップページにおけるリンクの出現位置なども影響するので，ロバストな方法ではないと思われる．今後様々なサイトにおける評価を行うことで，様々なサイトの構造に対応できる方法であるかどうかを検証する必要がある．

グループ 8 に対する評価はどの手法も同じスコアになっているが，これはこのページ群が強連結成分を成していないため，すべてのページが孤立点として抽出されてしまったためである．このような強連結成分を成さないページ群に対する抽出に関しては今後の課題である．

また正解グループ 1 に対する詳しい評価結果を，探索順位を優先した方法によるグループ化について表 5 に示した．この表には表 3 に対応する各正解グループに対して，スコアが最大となる自動グループに含まれる正解ページ数 x とそのグループの大きさ X，そしてスコアが示してある．この結果を見ると，この正解ページ群全体から成る大きさが 161 であるグループ 1 に対するものとして，自動グループ化では 172 ページのグループを抽出し，そのうち 156 ページが正解グループに含まれるページである．これは精度 0.97，再現率 0.91 でありどちらも十分な値であると言えるので，スコアが 0.937 というのはグループ化が適切に行われたことを示すものであると言える．また正解グループの大きさが 10 以下のグループに対しては，自動グループ化においても全く同じページ群を抽出しており，スコアが 1.00 となっている．一方でこれらの中間に位置する大きさが 20 前後のグループに関しては，X の大きさが 172 であったり 10 以下の大きさであったりすることから分かるように，全く抽出が出来ていない．これはこのグループの各ページにこのグループのコアページへのリンクが存在しないために，このページ集合が強連結成分を成さないことが原因である．これらのグループに関しては，スコアはいずれも 0.5 以下となっている．この例では正解グループの最上位の階層のグループと末端のグループの抽出ができ，中間の階層のグループが抽出できていない例であるが，その中間グループが抽出できないことに対するマイナス点が，0.698 という全体のスコアに反映されていると考えることができる．このような不適切な階層的グループ化がスコアにどう反映されるかは，正解グループの階層構造によっても異なってくるが，グループ化が適切に行われているかの一応の目安にはなり得ると考えられる．なお正解グループ全てに対するトータルのスコアが探索順位を利用したもので 0.700 となったが，このスコアは表 5 の例とほぼ同じ値であるので，ある程度のグループ化には成功していると考えられる．

次にサイトマップを利用してコアページに関する評価を行った結果を表 6 に示す．表には各手法によるコアリストの推定に関する精度を示している．サイトマップに存在するサイト内へのリンクの中で，グループの代表ページになり得ると判断した 21 個のリンクを用いて評価を行った．結果は in-degree による指標が最も優れていて，ディレクトリ階層を用いたものがそれに続いている．ディレクトリ階層を用いた手法では in-degree の情報は全く使っていないので，両者の情報を組み合わせることにより，コアページの推定方法を改善できると考えられる．

7 今後の展望

本論文で提案したグループ化手法は，Web グラフにおける強連結成分に着目したものであり，ある程度の有効性は示すことができた．今後の展望として，グループ化手法自体の改善とグループ化を応用したシステムへの応用の点から課題点をまとめる．

7.1 強連結成分以外のグループ統合指標

強連結成分の抽出を行った実験結果において，グループ化率の平均が約 90%であることからも分かるように，残りの約 10%のページは孤立点となっていてグループとして抽出することができない．リンク分析だけでこれらのページ群をグループ化することは困難だと思われるが，いくつかのヒューリスティクスを使うことによりある程度のグループ化は可能であると考えられる．例えば参照されているページが一つしか存在しない場合にはそれらは同じグループであり，それ以上分割できない単位であると見なせる．このようなヒューリスティクスなどを使い，複数ページによる強連結成分を成さないページ群のグループ化を今後試みる．

7.2 グループのメタ情報の抽出

グループ化した結果を利用したナビゲーションシステムなどへの応用のために，グループの内容などを表すメタ情報を抽出する必要がある．考えられる最も単純な方法は，グループのコアページのタイトルをそのグループの内容を表すものとしてそのまま利用する方法である．しかしタイトルがグループの内容を適切に表していない場合や，もう少し詳しい内容を知りたいという場合もあり，ページ群のテキスト解析なども必要であると考えられる．

7.3 検索システムへの応用

さらに次の段階の展望としては，検索システムへの応用がある．索引付けをグループ単位で行うことにより複数のページに跨がる文書を一つの文書として扱うことができ，内容が多岐に渡るような検索質問に対する検索性能の向上が期待できる．得られるグループは階層構造をしているので，このようなシステムを実現するためにはグループの粒度をどのように扱うかを検討する必要がある．例えば粒度を固定してある大きさ以下のグループのみを利用するか，もしくは何種類かの粒度を設定してその粒度ごとに索引付けを行うということが考えられる．複数の粒度を設定してそれぞれに索引を作る場合には，検索処理の段階でどのようなランキングを行うかが課題となる．

8 おわりに

本論文では，Web サイト内におけるグループのリンク構造について考察し，サイト内のグループと Web グラフにおける強連結成分との間に密接な関係があるという仮説を立てた．その考えを基に，リンク情報をベースとしたグループ化の手法についての提案を行った．そしてグループ化の評価についての検討を行い，評価を行った．

今後は評価に用いるための正解を増やし，評価の妥当性をさらに検証する必要がある．その上でグループ化アルゴリズムの改善を試み，7 章で述べたような課題について取り組んでいく．

文献

[1] 風間一洋; 原田昌紀; 佐藤進也, 「サーチエンジンの検索結果のマルチレベル・グルーピング」, 第 2 回インターネットテクノロジーワークショップ, 1999.

[2] Ravi Kumar; Prabhakar Raghavan; Sridhar Rajagopalan; Andrew Tomkins, “Trawling the web for emerging cyber-communities”, *8th International World Wide Web Conference*, 1999.

[3] 村田剛志, 「参照の共起性に基づく Web コミュニティーの発見」, 人工知能学会論文誌, Vol.16, No.3, 2001.

[4] Jon M. Kleinberg, “Authoritative Sources in a Hyperlinked Environment”, *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms*, 1998.

[5] Loren Terveen; Will Hill; Brian Amento, “Constructing, Organizing, and Visualizing Collections of Topically Related Web Resources” *ACM Transactions on Computer-Human Interaction*, Vol.6, No.1, 1999.

[6] 徳永健伸, 「情報検索と言語処理」, 東京大学出版会, 1999.