パラメータ化された連結性に基づく Web ページのグループ化

Grouping Web Pages Based on Parameterized Connectivity

正田 備也♥ 高須 淳宏♦ 安達 淳♦

Tomonari MASADA Atsuhiro TAKASU Jun ADACHI

WWW 上の情報が増大するにつれ，テキスト情報のみに基づく Web 検索は非現実的となっている．近年の研究は，HTML 文書のリンク情報を利用した効果的な Web 検索技術を提案している．Web 検索は主にページのランク付けとページのグループ化とから成る．本論文において，我々はリンク情報のみに基づく新たなグループ化手法を提案する．個別の Web ページはしばしば乏しい情報しか含まないため，Web ページをグループ化し，テキストに基づく検索のための新たな論理的単位を生成すれば，Web 検索の改善が期待できる．我々のグループ化手法を用いれば，一つの閾値パラメータを調整することで，ページ・グループの粒度を制御することができる．さらに，我々の手法は，グラフ理論的観点からは強連結成分分解の一般化とみなすことができる．本論文は，50 万の現実の Web ページについて行われた予備実験の結果も含む．

As the amount of information on WWW grows, Web search based only on textual information becomes unrealistic. Recent researches provide efficient Web search techniques utilizing hyperlink information of HTML documents. Web search mainly consists of page ranking and page grouping. In this paper, we present a new grouping method based only on hyperlink information. Since individual Web pages often contain poor textual information, by grouping Web pages and generating new logical units for text-based retrievals, improvement of Web search can be expected. With our grouping method, the granularity of page groups can be controlled by adjusting one threshold parameter. Moreover, from the graph-theoretic viewpoint, our method can be interpreted as a generalization of decomposition into strongly connected components. This papar also includes the results of preliminary experiments with half a million real Web pages.

1. はじめに

1.1 研究の目的

本研究は，リンク情報のみに基づいてWebページのグループ化を実現する手法を提案する．まず，Webページのグループ化には少なくとも三つの目的がある．第一に，個々のページではなく，ページの集合を，テキスト情報に基づく検索のための新たな検索単位とすること．これは，個別のページがしばしば貧弱なテキスト情報しか含んでいないためである．第二に，検索の前処理としてページ間の関連性を明示化しておくこと．なぜなら，[1]にも指摘されているように，情報検索において再現率を向上させるには，類似した文書を予め束ねておくことが有効だからである．第三に，大量の文書を閲覧する際の利便性を上げること．なぜなら，一次元的なリストとして出力される検索結果は見づらく，ユーザ・インターフェイスの観点からも望ましくないからである[2].

次に，Webページのグループ化を，特にリンク情報のみに基づいて行うことの目的は，少なくとも三つある．第一に，テキスト情報を含まないページもグループ化の対象とすること．なぜなら，画像や動画しか含まないページは実際かなり多いからである．第二に，グループ化のアルゴリズムにスケーラビリティを確保すること．Web検索が困難なのは，もっぱら文書数が莫大だからである．よって，形態素解析のようなコストの高い処理を前提とするテキスト処理は，グループ化に要する時間を増加させる．第三に，他の領域への応用を容易にすること．なぜなら，リンク情報だけを使うアルゴリズムは，Webだけでなく，有向グラフ構造を抽出できるすべての領域に適用可能だからである．

1.2 本研究の特色

本研究は，以下の 4 点を特色とする．これら 4 点を兼備した研究は，従来なかったと思われる．

(1) リンク情報のみに基づいて，Web ページのグループ化を実現するアルゴリズムを提案していること．

(2) 提案されているアルゴリズムによるグループ化が，リンク構造を有向グラフと見た場合の強連結成分分解の細分化になっていること．

(3) 一つの閾値パラメータの値を調節することで，得られるグループの粒度を制御できること．

(4) リンク構造上での Web ページ間の類似性の尺度として，三角不等式を満たす距離概念を導入していること．

要するに，本研究の提案するグループ化は，閾値パラメータに依存して粒度が変化する，強連結成分分解の細分化である．そこで本研究では，このグループ化によって得られるグループを，パラメータ化された連結成分(parameterized connected component)と呼び，PCC と略記する．

1.3 既存研究との比較

得られるグループの粒度を制御しつつ文書のグループ化を行う試みは[1]に見られる．だが，[3]に指摘のあるように，グループ化にテキスト情報を利用しているため，相当な処理時間を要する．Webページの数の莫大さに鑑みれば，テキスト情報からの特徴量の抽出が完了していることを前提とするグループ化手法は望ましくない．[4]では，リンク構造を利用して検索単位を個々のページよりも大きくとる手法が提案されている．しかし，リンクでつながれた文書間の類似性評価にやはりテキスト情報を利用しているため大規模なページ集合には適用できない．テキスト情報によらずにWebページをグループ化するための発見的手法として，URLを手がかりとするものがすでに提案されている[2]．しかし，URLを手がかりとするためには，様々なサイトの内部構造とURLの階層構造との対応関係を経験的に調査する必要がある．また，このような調査の結果が十分な一般性を持つとは限らない．

そこで，リンク構造を利用することが考えられる．Webペ

♥ 学生会員 東京大学大学院情報理工学系研究科 masada@nii.ac.jp

♦ 正会員 国立情報学研究所 {takasu,adachi}@nii.ac.jp

ージのリンク構造は有向グラフとみなすことができる．[5]は，この有向グラフの隣接行列をAとして，$A^T A$ないし$A A^T$の固有ベクトルのエントリの符号を利用するグループ化手法を提案している．しかし，その効果を疑問視する研究がある[6].

ところで，周知のように，有向グラフ上では強連結成分分解によって頂点をグループ化することができる．しかもこの作業は頂点数と有向枝の数の線形オーダの計算量で実行可能である[7]. だが，Webページの集合について強連結成分分解を行うと，あまりにも多くのページを含むグループが構成されてしまうことが知られている[8][9]．そこで，本研究では，強連結成分をさらに細分化するかたちでWebページのグループ化を実現するアルゴリズムを提案する．さらに，このアルゴリズムを使えば，一つの閾値パラメータを調節することで，得られるグループの大きさを制御できる．

本論文の構成は以下のとおりである．第2章では，リンク構造上でのWebページ間の近さを定量的に表現するための新しい概念を導入し，さらに，実装を考慮してこれらの概念を再定義する．第3章では，この概念を利用したグループ化のアルゴリズムを提示する．第4章では，予備実験の結果を示し，第5章で結論と今後の課題を述べる．

2. 概念の定義

本研究は，ある Web ページから別の Web ページへの移行のしやすさを表わす尺度として，**ドリフト**という概念を提案する．そして，このドリフトに基づいて，リンク構造上での Web ページ間の近さを定量的にあらわす概念として**相互リンク距離**を定義する．さらに，これらの概念を計算量縮減の観点から再定義することを試みる．

2.1 ドリフト

WWWのリンク構造は，Webページを頂点，ハイパーリンクを有向枝と見なすことによって, 有向グラフ$G=(V,E)$と解される．以下頂点数をnとする．なお，頂点から自分自身に対して張られている有向枝は無視し，2頂点間の同じ向きの複数の有向枝は一つとみなすことにする. 有向グラフG において，頂点$i \in V$ から出て行く有向枝の数(out-degree)をd^+_i, 頂点i へと入って行く有向枝の数(in-degree)をd^-_iと書くことにする．頂点i から頂点j への歩道(walk)とは，頂点の列 $i,i_1,...,i_p = j$ および有向枝の列$(i,i_1), (i_1,i_2), \ldots, (i_{p-1},j)$で，頂点や有向枝が重複してもよいもののことをいう．有向路(path)とは，相異なる頂点からなる歩道のことである．有向グラフは，任意の頂点 $i,j \in V$ について，i からj への有向路と，j からi への有向路が存在するとき, 強連結(strongly connected)と呼ばれる．強連結成分分解とは，与えられた有向グラフを極大な強連結成分へと分解することをいう．さて，mを下式を満たす非負の整数とする．

$$m > \min\left(\max_{i\in V} d_i^+, \max_{i\in V} d_i^-, \max_{i\in V} \sqrt{d_i^+ d_i^-}\right)$$

最後の値は，[10]において与えられている，有向グラフの隣接行列のスペクトル半径の上界である．そこで，実数r を $r = 1/m$と定め，このr を使って行列B を

$$B \equiv \sum_{l=1}^{\infty} (rA)^l$$

と定義する．ここでA は有向グラフG の隣接行列とする．r の決め方より，B の定義式である和は収束する[11]. また，行列B は，$n \times n$の単位行列をIとして$B = (I - rA)^{-1} - I$ という式によって求めることができる．なお，行列B の(i,j)エントリb_{ij} は，あらゆる長さの歩道の本数を，歩道の長さが増大するにつれて指数関数的に減少する重み付けによって加え合わせたものになっている．なぜなら，A^l の(i,j)エントリ$a^{(l)}_{ij}$は，頂点i からj への長さl の歩道の総数に等しいからである．そこで，本研究では値b_{ij} を**ドリフト(drift)**と呼び，頂点i から頂点j への移行のしやすさの定量的評価に用いる．

2.2 相互リンク距離

上述のドリフトに基づき，Webページ間の近さを

$$d(i,j) \equiv -\log_m b_{ij} - \log_m b_{ji}$$

と定義する．このように定義された近さは三角不等式を満たす. このことは, 頂点i からj へ至るすべての歩道の集合が，頂点i から第三の頂点k を経由して頂点j に至るすべての歩道の集合をその部分集合とするという事実より，明らかである．そこで，本研究ではこの$d(i,j)$を頂点iとjとの**相互リンク距離(mutual-link distance, ML-distance)**と呼ぶことにする．なお，頂点i から頂点j への歩道が存在しないか，頂点j から頂点i への歩道が存在しない場合は，相互リンク距離$d(i,j)$は無限大と定めることにする．

2.3 パラメータ化された連結成分

上記の相互リンク距離に基づいて，パラメータ化された連結成分(PCC)を以下のように定義する．閾値パラメータが τ のとき，頂点集合がPCCをなすとは，属する任意の2頂点間の相互リンク距離が2τ以下であることをいう．本研究では，与えられた閾値パラメータ τに応じて，有向グラフの頂点集合を，互いに交わらないPCCへと分解するアルゴリズムを提案する．

2.4 実践的観点からの再定義

上記のドリフト，および相互リンク距離の定義は，実装に配慮したものではない．なぜなら，行列Bを式$B = (I - rA)^{-1} - I$ によって計算するには$O(n^3)$程度の計算量を要するためである．そこで，ドリフトの定義においては頂点iと頂点jとの間の最短有向路の長さ$SDPL(i,j)$の寄与が支配的である点に注目し，ドリフトを$r^{SDPL(i,j)}$と，また相互リンク距離を

$$d_{PRACTICAL}(i,j) \equiv SDPL(i,j) + SDPL(j,i)$$

と定義し直す．これにより，近さの評価の精密さを犠牲にする代わりに，実装上の有利さを得る．PCCもこれにしたがって再定義される．なお，再定義された相互リンク距離も，三角不等式を満たす.

3. アルゴリズム

本研究の提案するパラメータ化された連結成分(PCC)は，強連結成分概念の一般化とみなすことができる．実際，次の段落で見るように，閾値パラメータ τをある値以上にとると，アルゴリズムは強連結成分分解を与える．また，PCCへの分解を求めるアルゴリズム自体も，強連結成分分解を求めるアルゴリズムに類似している．

ところで，ある頂点v について，それを含む極大な強連結成分は，v から到達可能な頂点の集合と，v へと到達可能な頂点の集合との交わりであることが知られている[7]. 本研究の提案するアルゴリズムも，類似の事実を利用している．ただし，v から到達可能な頂点の集合と，v へと到達可能な頂点の集合との交わり全体を求めるのではない．この交わりに属する頂点のうち，v との相互リンク距離が τ以下のものだけを，頂点v を含む連結成分の要素として認める．こうすれ

ば，頂点v との間に双方向に有向路をもち，しかもv からの相互リンク距離が τ以下の頂点だけを選ぶことができる．なお，2頂点間の有向路の長さは高々n -1なので， τが2(n -1)以上のとき，このアルゴリズムは強連結成分分解を与える．

PCC構成のためのアルゴリズムを以下に示す．アルゴリズムに与えられる閾値パラメータは τとする．

1．すべての頂点を未処理としてマークする．
2．未処理の頂点vを一つ任意に選び，これを処理済とマークする．
3．v から，リンクを順方向に辿って，幅優先探索を探索木の深さが τになるまで進める．
4．v から，今度はリンクを逆向きにたどって，幅優先探索を木の深さが τになるまで進める．
5．３と４の両方の幅優先探索で訪問された頂点のうち，３でのvからの深さと４でのv からの深さの和が τ以下のものだけを，頂点vと同じPCCに属する頂点として登録する．同時に，これらの頂点を処理済とマークする．
6．頂点がすべて処理済とマークされるまで，２から５を繰り返す．

このアルゴリズムについて，以下の事実が成立する．

1．上記のアルゴリズムが与える任意のグループについて，それに属する任意の２頂点の相互リンク距離は高々2 τ．
2．上記のアルゴリズムにおいて，頂点vを中心に構成されたグループに属さない頂点はすべて，vとの相互リンク距離が τより大きい．

証明は，相互リンク距離が三角不等式を満たすことに基づいて行われる．これらの事実は，今回提案したアルゴリズムが，相互リンク距離の観点から，性質の良いグループだけを生成することを保障している．

なお，ここで提案したアルゴリズムは，あらかじめすべての頂点対について相互リンク距離を求めているのではない．すべての頂点対について相互リンク距離を保存するには，$O(n^2)$の空間計算量を要する．すると，nが大のとき，全データをメモリ上に保持することが困難となり，アルゴリズムのスケーラビリティが失われる．なぜなら，データをメモリではなくディスク上に保存すると，アクセス時間のために実行時間が増大するからである．したがって，今回提案したアルゴリズムのように，必要最小限のデータをメモリ上に保持する工夫が必要である．

4．予備実験

予備実験では，あるページからのクローリングによって取得した490118件のWebページ集合を用いた．実験環境は，Sun Blade 1000 (CPU: UltraSPARC-III 750MHz, 900MHz, Memory: 8192Mbyte)である．実験結果を示す前に，このサンプル集合の性質を，次数の分布を調査することによって明らかにしておく．各ページに入ってくるリンク数，つまりin-degreeの分布と，各ページから出ていくリンク数，つまりout-degreeの分布とは，それぞれ図1，図2のようになっている．いずれもpower lawと呼ばれる分布に従っている．つまり，次数xと，その次数を持つページの個数yとの間に，ある定数 α， βに対して

$$\log y = \alpha - \beta \log x$$

という関係が成立している．この調査が，[8]における大規模な調査の結果と合致していることより，予備実験で使用したサンプル集合におけるリンク構造が，Web全体の傾向から大きく逸脱していないと推測できる．

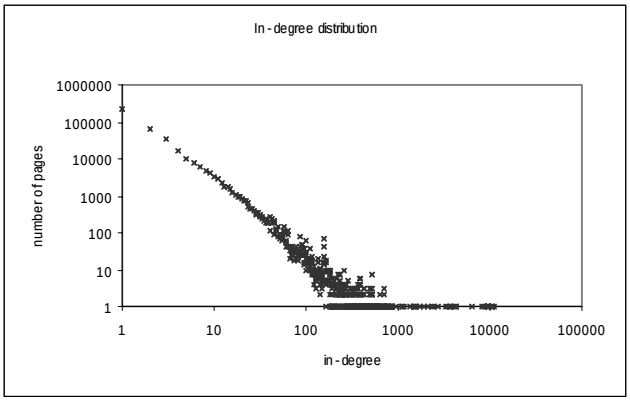

図1：今回使用したページ集合でのin-degreeの分布
横軸が次数．縦軸がその次数をもつページの数．

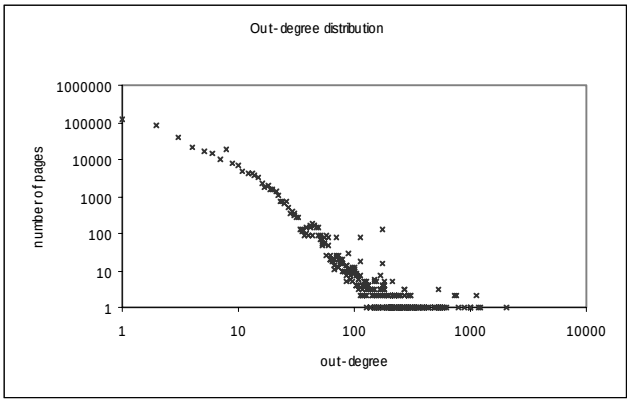

図2：今回使用したページ集合でのout-degreeの分布
横軸が次数．縦軸がその次数をもつページの数．

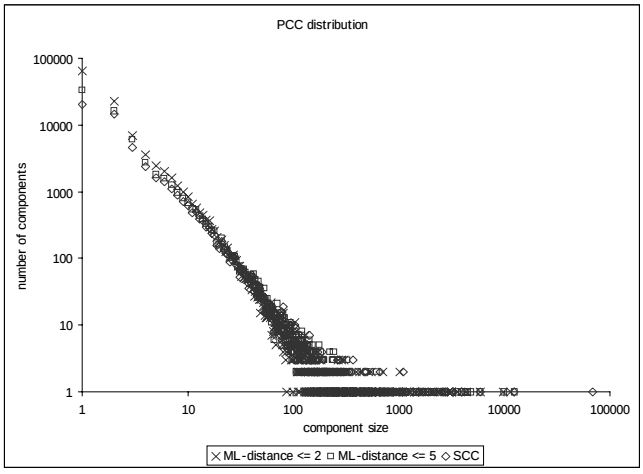

図3：PCCの大きさの分布
横軸がPCCの大きさ．縦軸がその大きさをもつPCCの数．閾値パラメータ τが2, 5の場合，そして強連結成分分解の場合を同時にプロットしてある．パラメータを減少させると，グループの粒度が低下している．

次に，第3章で提案したアルゴリズムを適用し，PCCへの分解を求めた実験結果を図3に示す．グラフより，強連結成分分解の場合, 閾値パラメータ τが5の場合, 2の場合の順に，大きいPCCの個数が減り，かつ，小さいPCCの個数が増加している．つまり，閾値パラメータの調節によって，グループの粒度が制御されていることが分かる．

最後に，規模最大のグループの大きさが，閾値パラメータに応じてどのように変化するかを表1に示す．

threshold parameter	Largest group size
2	11417
5	12488
10	24866
15	58077
30	69179
(SCC)	69180

表1：最大PCCの大きさの変化
閾値パラメータを変化させると，最も大きいグループの大きさがどのように変化するかを示す．最下行は強連結成分分解での最大グループの大きさ．パラメータを減らすと，最大グループの大きさが小さくなる．

閾値パラメータの減少にともなって最大グループの大きさは小さくなっており，やはりグループの粒度が制御されていると観察される．

5. まとめと今後の課題

本研究は，ドリフトおよび相互リンク距離という新たな概念を導入することによって，閾値パラメータに応じてグループの粒度を変化させつつ Web ページのグループ化を行う手法を提案することに成功した．また，このグループ化の手法は，強連結成分分解の一般化とみなすことができる．実際，閾値パラメータをある値以上に設定すると，強連結成分分解そのものが与えられる．

ところで，ここ数年，有向グラフに関する理論的考察が大きく前進している．具体的には，古典的なランダム・グラフの理論が，次の２点において拡張されている．(1)無向ランダム・グラフについて得られた成果の，有向ランダム・グラフの解析への応用. (2)枝が一様な確率に従って生起するという仮定を除去し，各頂点での次数が任意の値をとると仮定することによるランダム・グラフの一般化．

具体的には，Web グラフのいわゆる「蝶ネクタイ」構造[8]が，次数の分布が特定の条件を満たすときに必然的に帰結する構造であることが，証明されるなどしている[12]．今回提案した手法によって得られるグループの大きさの分布は，第 4 章の結果を見る限り，強連結成分の大きさの分布と同様，ほぼ power law に従っている．そこで，このことがリンク構造に内在する何らかのグラフ理論的性質からの形式的な帰結であるのか否かを明らかにすることが，今後の課題である．

[謝辞]

本研究は，文部科学省科学研究費補助金特定領域研究「情報学」（課題番号 13224087）の助成のもとに行われた．

[文献]

[1] Cutting, D. R., Pedersen, J. O., Karger, D., and Tukey, J. W.: “Scatter/Gather: A cluster-based approach to browsing large document collections”, Proceedings of the Fifteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp.318-329 (1992).

[2] Loren Terveen, L., Hill, W., and Amento, B.: "Constructing, organizing, and visualizing collections of topically related Web resources", ACM Transactions on Computer-Human Interaction, Vol. 6, No.1, pp.67-94 (1999).

[3] Hearst, M. A. and Pedersen, O.: "Reexamining the cluster hypothesis: Scatter/Gather on retrieval results", In Research and Development in Information Retrieval, pp.76-84 (1996).

[4] Tajima, K., Mizuuchi, Y., Kitagawa, M., and Tanaka, K.: “Cut as a querying unit for WWW, netnews, and e-mail”, Proceedings of ACM Hypertext '98, pp.235-244 (1998).

[5] Kleinberg, J. M.: “Authoritative sources in a hyperlinked environment”, Journal of the ACM, Vol. 46, No. 5, pp.604-632 (1999).

[6] Borodin, A., Roberts, G. O., Rosenthal, J. S., and Tsaparas, P.: “Finding authorities and hubs from link structures on the World Wide Web”, Proceedings of 10th International WWW Conference, pp.415-429 (2001).

[7] Nuutila, E. and Soisalon-Soininen, E.: ”On finding the strongly connected components in a directed graph”, Information Processing Letters, Vol. 49, pp.9-14 (1993).

[8] Broder, A. Z., Kumar, S. R., Maghoul, F. , Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A., and Wiener, J.: "Graph structure in the Web", Proceedings of 9th WWW Conference, pp.309-320 (2000).

[9] 小島 秀一, 高須 淳宏, 安達 淳: "Web ページ群の構造解析とグループ化", NII Journal, Vol.4, pp.23-35 (2002).

[10] Kwapisz, J.: ”On the spectral radius of a directed graph", Journal of Graph Theory, Vol. 23, No. 4, pp.405-411 (1996).

[11] Bapat, R. B. and Raghavan, T. E. S.: "Nonnegative Matrices and Applications", Encyclopedia of Mathematics and Its Applications, Vol. 64, Cambridge University Press (1997).

[12] Cooper, C. and Frieze, A.: “The size of the largest strongly connected component of a random digraph with a given degree sequence”, pre-print., http://www.math.cmu.edu/~af1p/papers.html (2002).

正田 備也 Tomonari MASADA
東京大学大学院情報理工学系研究科博士課程在学中．1995 年東京大学大学院理学系研究科情報科学専攻博士前期課程修了．情報検索の研究に従事．情報処理学会学生会員．

高須 淳宏 Atsuhiro TAKASU
1989 年東京大学大学院工学系研究科博士課程修了．工学博士．同年，学術情報センター研究開発部助手．1993 年より同センター助教授．データベースシステム，文書画像理解，機械学習の研究に従事．電子情報通信学会，情報処理学会，人工知能学会，ACM，IEEE 各会員．

安達 淳 Jun ADACHI
1981 年東京大学大学院工学系研究科博士課程修了．工学博士．東京大学大型計算機センター，文部省学術情報センターを経て現在国立情報学研究所教授．東京大学大学院情報理工学研究科教授を併任．データベースシステム，分散処理システム，情報検索，電子図書館システム等の開発研究に従事．電子情報通信学会，情報処理学会，IEEE，ACM 各会員．