D-12　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‡

† 東京大学 情報理工学系研究科, ‡ 国立情報学研究所

Tomonari Masada†, Atsuhiro Takasu‡, Jun Adachi‡

†Graduate School of Information Science and Technology, The University of Tokyo, ‡The National Institute of Informatics

1　はじめに

研究の目的：本研究では、リンク情報のみにもとづいてWebページのグループ化を行う手法を提案する。その目的は、Web上での高速な検索の実現にある。一般に情報検索のためのアルゴリズムは、(a) 文書から抽出される特徴量と、(b) 文書数とに依存する計算量を持つ。(a) については、各文書の特徴量として、最も標準的には、単語の出現頻度を基に計算されるtf·idfと呼ばれる値をエントリとする高次元のベクトルが求められるため、このベクトルの次元を削減するということが行われる。次元削減のためには、特異値分解法 SVD[BDO94] やランダム・プロジェクション [BM01] などを利用することができる。しかし、本研究が直接かかわるのは (b) である。すなわち、複数のWebページを束ねることでそれを単独の文書とみなし、これによって見かけ上の文書数を減らして検索を高速化しようとするのである。具体的には、ユーザが与えた質問を、個々のWebページではなく、Webページのグループと比較対照し、類似度が高いと判定された一つないし数個のグループを、"荒い検索" の結果として、次の "より細かい検索" へ引き渡す。このように、情報検索を多段的に構成するという試みは、すでに [CPKT92] に見られる。だが、[CPKT92] では文書のグループ化にテキスト情報を利用している。そのため、処理に非常に時間のかかることが [HP96] で指摘されている。Webページの数の莫大さに鑑みれば、形態素解析などによる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であろう。

グループ化のための手法：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ある [THA99]。しかし、その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り、このような調査の結果が満足できる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。リンク構造は一つの有向グラフとみなすことができるので、**強連結性**の概念を利用するというアイディアも捨てがたい。だが、Webページの集合について強連結成分分解を行うと、テキスト内容上も一つにまとまっていると期待するにはあまりにも多くのページを含むようなグループが構成されてしまうことが知られている [BKM+00][小島 02]。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。本研究の特徴は、次の3点にまとめられる。(1) 提案されているアルゴリズムによるグループ化が、強連結成分分解の細分化になっている点。(2) 提案されているアルゴリズムに、得られるグループの大きさを、一つの閾値パラメータの増減によって制御できる仕組みが備わっている点。(3) グループの粒度の制御を可能にするための理論上の道具として、リンク構造上でのページ間の近さを定量的に表す概念を提案している点。

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。

準備：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、一つの有向グラフ $G=(V,E)$ と理解される。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することにし、同じ2つの頂点間にある同じ向きの複数の有向枝は一つとみなすことにする。グラフ G において、頂点 $i \in V$ から出て行く有向枝の数を d_i^+、頂点 $i \in V$ へと入って行く有向枝の数を 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 $(i, i_1), (i_1, i_2), ..., (i_{p-1}, j)$ 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 $i, j \in V$ 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。

ドリフト：m を、下式を満たす非負の整数とする。

$$m \geq \min\left\{\max_{i\in V} d_i^+, \max_{i\in V} d_i^-, \max_{i\in V}\sqrt{d_i^+ d_i^-}\right\} \qquad (1)$$

最後の値は、[Kwa96] 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 [BR97]。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。ところで、行列 B の第 (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ともなっ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^k の第 (i,j) エントリ $a_{ij}^{(k)}$ は、頂点 i から頂点 j への長さ k 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。

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 d を

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

と定義する。この量は、有向グラフ上での2頂点間のいわば "親密さ" の度合いを表している。さらに、上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 から頂点 j へ至るすべての有向路の集合は、頂点 i から第三の頂点 k を経由して頂点 j に至るすべての有向路をその部分集合として含む、という事実に基づいて証明される。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**と呼ぶことにする。なお、頂点 i から頂点 j への有向路が存在しないか、頂点 j から頂点 i への有向路が存在しない場合は、相互リンク距離 $d(i,j) = \infty$ と定めることにする。

3　アルゴリズム

本研究では、上記の相互リンク距離という距離概念を利用し、有向グラフ上で、強連結成分をさらに細分化するような頂点の分解を与えるアルゴリズムを提案する。さらに、このアルゴリズムは、一つのパラメータを増減させることで、結果として得られる頂点のグループの大きさを調整できるような仕組みを備えている。そこで、こうして得られる頂点のグループを、**パラメータ化された連結成分 (PCC: parameterized connected component)**

と呼ぶことにする。PCC への分解を得るためのアルゴリズムを以下に示す。

for each $i \in V$ **do**
  Do breadth first search from i
  and Compute drift to every other page;
$C := \{1\}$;
for each $i \in V \setminus C$ **do**
  if $d(i,j) \geq \tau$ holds for all $j \in C$ **then**
    $C := C \cup \{i\}$;
for each $i \in V \setminus C$ **do**
begin
  Find $j \in C$ nearest to i;
  $PCC(j) := PCC(j) \cup \{i\}$;
end.

最初の for ループでは、任意の Web ページから、他のすべての Web ページへのドリフトを求めている。ドリフトの算出には、$B = (I - rA)^{-1} - I$ という式によって全ドリフトを数値計算的に一挙に求めるという手法がある。しかし、逆行列の計算に必要な計算量に鑑みれば、むしろ、ドリフトの値 b_{ij} が結果的にページ i から j への最短有向路長によって支配されることを利用し、ドリフトの計算を、各ページから他の全ページへの最短有向路長の算出によって置き換えることが考えられる。そのためには、各ページから幅優先探索を行うことになるが、さらにここで、幅優先探索をある深さで打ち切ることによって、それ以上遠いページについては有向路がないものとみなし、一層近似を推し進めることで実行時間の短縮を図るという手法も考えられる。今回の予備実験に際しての実装では、この最後の手法を採用した。

次に、第二の for ループでは、結果として得られるグループの中心となる Web ページだけを枚挙している。まず、適当な Web ページをアド・ホックに中心ページと定め、以下、すでにグループの中心ページとして登録されている頂点のすべてから、閾値パラメータ τ 以上に離れているものだけを、新たな中心ページとして登録する。$\tau = \infty$ のとき、アルゴリズムの与えるグループ化は、強連結成分分解に一致する。そして、τ を徐々に減少させることで、強連結成分分解よりも段階的にグループの粒度が細かくなっていくようなグループ化が実現される。最後に、第三の for ループにおいて、残った頂点をそれに最も近い中心の配下にあるグループの構成員として登録していく。なお、$PCC(i)$ とは頂点 i を中心とする Web ページのグループである。こうして得られるグループについては、それに属する任意の 2 頂点 i, j の距離 $d(i,j)$ が $d(i,j) \leq 2\tau$ を満たす、という事実を証明できる（証明略）。

4 予備実験

今回の予備実験では、特定の Web ページからのクローリングによって得られた 59129 件の Web ページを対象とした。すべての処理は Sun Blade 1000(CPU UltraSPARC-III 750MHz, 900MHz。メモリ 8192M バイト。) 上で行っている。値 $r = 1/m$ は式 1 によって求めた。今回の実験では $m = 567$ とした。図 1 は、閾値パラメータを増減させることによって、得られるグループの大きさがどのように変化するかを示している。横軸はグループの大きさ、縦軸はグループの個数である。閾値パラメータの値を小さくすることによって、全体的にグループのサイズが小さくなっていくのが分かる。特に、次の 2 点に注意されたい。(1) 最も大きなグループのサイズが、閾値 5 の場合と、閾値 10 や 20 の場合とでは明らかに前者の方が小さいこと。(2) サイズが 50 から 100 の中間的な粒度のグループについては、閾値 5 の場合の方が、閾値 10 や 20 の場合に比べて、その個数が多いこと。

5 おわりに

今後の課題：本研究は、同じリンク解析ではあっても、PageRank[BP98] や Hub/Authority[Kle99] などとは異なり、Web ページのランキングではなく、グループ化の手法を提案した。しかし、今後は、これら既存のリンク解析手法との比較を可能にするような評価基準を提案し、それぞれの手法がリンク構造のどのような特徴を抽出しており、また、それぞれの手法にどのような得手・不得手があるのかを、[BRRT01] のような先例に倣いつつ、より形式的に議論する必要があると思われる。

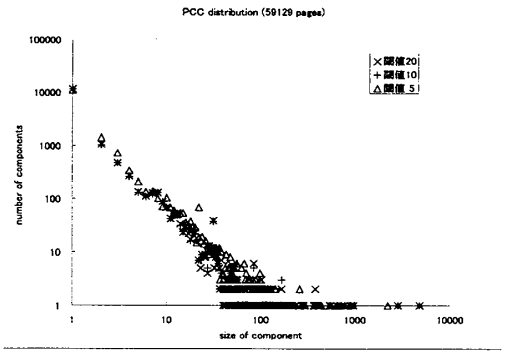

図 1: 閾値が 5, 10, 20 それぞれの場合のグループの大きさの分布

参考文献

[BDO94] Michael W. Berry, Susan T. Dumais, and Gavin W. O'Brien. Using linear algebra for intelligent information retrieval. Technical Report UT-CS-94-270, Department of Computer Science, University of Tennessee, 1994.

[BKM+00] A. Z. Broder, S. R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. Graph structure in the web. In *Proceedings of 9th WWW Conference*, pp. 309–320, 2000.

[BM01] Ella Bingham and Heikki Mannila. Random projection in dimensionality reduction: applications to image and text data. In *Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2001), August 26-29, 2001, San Francisco, CA, USA*, pp. 245–250, 2001.

[BP98] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, Vol. 30, No. 1–7, pp. 107–117, 1998.

[BR97] R. B. Bapat and T. E. S. Raghavan. *Nonnegative Matrices and Applications*, Vol. 64 of *Encyclopedia of Mathematics and Its Applications*. Cambridge University Press, 1997.

[BRRT01] Alan Borodin, Gareth O. Roberts, Jeffrey S. Rosenthal, and Panayiotis Tsaparas. Finding authorities and hubs from link structures on the world wide web. In *10th International WWW Conference*, pp. 415–429, 2001.

[CPKT92] Douglass R. Cutting, Jan O. Pedersen, David Karger, and John W. Tukey. Scatter/gather: A cluster-based approach to browsing large document collections. In *Proceedings of the Fifteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 318–329, 1992.

[HP96] Marti A. Hearst and Jan O. Pedersen. Reexamining the cluster hypothesis: Scatter/gather on retrieval results. In *Research and Development in Information Retrieval*, pp. 76–84, 1996.

[Kle99] Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, Vol. 46, No. 5, pp. 604–632, 1999.

[Kwa96] Jaroslaw Kwapisz. On the spectral radius of a directed graph. *Journal of Graph Theory*, Vol. Vol. 23, No. No. 4, pp. 405–411, 1996.

[THA99] Loren Terveen, Will Hill, and Brian Amento. Constructing, organizing, and visualizing collections of topically related Web resources. *ACM Transactions on Computer-Human Interaction*, Vol. 6, No. 1, pp. 67–94, 1999.

[小島 02] 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.