社団法人 情報処理学会　研究報告
IPSJ SIG Technical Report

2003－MPS－46　(7)
2003／9／18

WWW 上の効率的なハブ探索法の提案と実装

松久保　潤[†]　林　幸雄[†]

本論文では，Web 上で高い入次数をもつ有益なページをできるだけ多く収集するために，Web クローラが発見した未探索ページの入次数に従って，探索順を適宜決める手法を提案する．従来法では，ページの探索順を決めるために Web 全体のリンク構造を用いるのに対し，提案手法では，局所的なリンク構造のみを用いて探索順を決める．さらに，Java で実装した Web クローラで探索を行い，提案手法が幅優先探索よりも効率的に有益なページを発見できることを確かめる．実験結果から，高い入次数をもつページを優先的に探索することによって，提案手法はコミュニティの核となるハブを経由してページを収集することが示唆される．

Effective searching method for hubs on WWW and the implementation

Jun Matsukubo[†] and Yukio Hayashi[†]

In this paper, we propose an effective searching method for high in-degree Web pages. On the proposed method, the searching order is decided adaptively by the in-degrees of the pages found during search. In conventional methods, the searching order is decided with the link structure of the entire Web, while in the proposed method, only with the local link structure. Then, we implement the Web crawler written with Java. By the saerch experiments with the crawler, we confirm that the proposed method can find hubs more effectively than breadth-first-search can. From the result, it is suggested that the proposed method may collect pages through hubs, which are the cores of Web communities.

1. はじめに

これまで，検索精度を上げられるよう，Web ページ収集のカバー率を上げる試みが続けられてきた．しかしながら，最近の報告 [1] では，Web ページの総数が推計 70 億であるのに対し，全サーチエンジンの推定カバー率は全体の約 40%とされている．そのため，半分以上の Web ページに現在のサーチエンジンから到達できないことになる．さらに，Web ページの数は現在も急速に増えているので，全てのページをカバーするのは現実に困難であり，できるだけ有益なページを優先する探索法の開発が望まれる．例えば，PageRank[2] では，被参照数 (入次数) の高いページほど多くの関心・興味を集めているので，有益であるとしている．本報告でも同様に，高い入次数をもつページほど有益であると考える．以下，高い入次数及び出次数をもつページをそれぞれオーソリティ及びハブと呼ぶ．

オーソリティを効率的に収集するために，Cho ら [3] はページに割り当てた重要度に従って探索順を決める手法を提案した．Cho らが定義した重要度 (出次数，入次数，PageRank) を正確に求めるには，**Web 全体のリンク構造を予め把握しておかなければならない**が，現実の Web では非常に困難である．そのため，この点を改善するために，Web 全体のリンク構造を用いないで探索順を決める修正法が考えられる．

そこで，本報告では，オーソリティを効率的に収集するため，**Web クローラが発見した未探索ページの入次数に従って，探索順を適宜決める**手法を提案する．探索が進むと，被参照リンクが見つかることで入次数が増えるページがあるため，提案手法でのページの探索順は適応的に変化する．さらに，Java を用いて Web クローラを実装し，幅優先探索による結果との比較から，探索されたページの結合分布，及びオーソリティ収集の効率性について議論する．

2. Web ページの重要度の評価法

Web ページの重要度の評価としては，様々なものが考えられる．Cho ら [3] は，オーソリティを効率的に収集するために，出次数 (他のページへの参照リンク数)，入次数 (他のページからの被参照リンク数)，及

† 北陸先端科学技術大学院大学 知識科学研究科

び PageRank[2] の値を重要度の指標として探索に用いる手法を提案した．ただし，この手法は Web 全体のリンク構造が既知であることを前提としている．

Cho らは出次数，入次数，及び PageRank によるリンク構造を用いた重要度の評価に「ページ p から q へのリンクは p の作成者が q に対して関心・興味をもっていることを示す」という仮定を用いている．入次数による評価は，上述の仮定のもとで多くの関心・興味を集めているページに高い重要度を割り当てることに対応する．PageRank はリンク構造からページの人気度を求める手法で，高い入次数をもつ頂点に高い重要度を割り当てる傾向がある．

この手法は Web 全体のリンク構造が既知であるとして探索順を決めているが，Web ページのリンク構造は巨大なだけでなく，**急速に成長・変化を続けている**ので，Web 全体のリンク構造を把握することは現実には難しい．そのため，出次数，入次数，及び PageRank を正確に求めることは困難である．この点を改善するために，Web 全体のリンク構造を用いずに探索順を決める修正法が考えられる．

次章では，**Web 全体のリンク構造ではなく，Web クローラが発見した未探索ページの入次数に従って，探索順を適宜決める**手法を提案する．

3. 入次数優先探索

本章では，まず 3.1 節で本報告の提案手法について示す．次に，3.2 節で Java によるクローラの実装法を説明し，3.3 節で探索実験の結果について考察する．

3.1 入次数優先探索

以下，本提案手法を入次数優先探索 (In-degree First Search; IFS) と呼ぶ．前章で説明した手法では，Web 全体のリンク構造を用いて探索順を決めるのに対し，IFS では，探索中に発見された未探索ページの入次数に従って，探索順を適宜決める．IFS では，まず HTML のダウンロード・構文解析を行い，他のページを参照している URL を抽出して未探索リストに格納する次に，未探索リストの中で最大の入次数をもつ URL に対応するページを次のステップで探索されるページとして選ぶ．以下，その手順を示す．

(1) 探索を開始する URL を p とする．
p を探索済みリスト U_{s} に追加する．
(2) p に対応する HTML から重複しないように URL を取り出し，一時 URL リスト U_{t} に格納する．
(3) U_{t} の先頭の URL を p_{t} とする．U_{t} から p_{t} を削除する．
(4) (a) U_{s} に p_{t} が含まれている場合，(5) へ進む．
(b) U_{s} に p_{t} が含まれていない場合:
(i) 重複チェック用 URL リスト U_{dup} に p_{t} が含まれている場合:
p_{t} の入次数 $D_{\mathrm{in}}(p_{\mathrm{t}})$ を 1 増やす．
(ii) U_{dup} に p_{t} が含まれていない場合:
探索キュー U_{s} に p_{t} を追加する．
U_{dup} に p_{t} を追加する．
p_{t} の入次数 $D_{\mathrm{in}}(p_{\mathrm{t}})$ を 1 にする．
(5) U_{t} が空でなければ，(3) に戻る．
(6) $D_{\mathrm{in}}(p)$ の最大の要素に対応する $p{\in}U_{\mathrm{s}}$(最大の入次数をもつ URL) を次の p とする．
D_{in} から p を削除する．
(7) (2) に戻る．

次節では，Java で実装した Web クローラについて説明する．

3.2 Java による Web クローラの実装

本節では，探索実験で用いた Web クローラの動作について述べる．

まず，HTML を解析するとき，他のページへのリンクとして，ページ間をつなぐリンクに対応する <A> タグの HREF 属性，ページに埋め込まれた HTML を指す <FRAME SRC> タグ，及び異なるページへのリダイレクションを示す <META> タグで指定されている URL を抽出した．

次に，上記の 3 つのタグから Web クローラが抽出した URL が，HTML 及び HTML に機能を拡張したページ (asp;Active Server Pages, jsp;Java Server Pages, php;Hypertext Preprocesser, cfm;Cold Fusion MX) を参照する場合のみ，リンクとして扱った．ただし，URL がクエリ “?...” を含む場合，要求側の状態に応じてクエリに対応するページが変化する可能性があるため，リンクとして扱わなかった．また cgi は無限にページを生成するプログラムを含むので，この実験ではリンクから除外した．http，https 以外のプロトコルを用いる URL は Web ページへのリンクではないので，実験では無視した．

次節で，Java で実装したクローラを用いて行った探索実験の結果を示し，探索されたページの結合分布とオーソリティの収集効率について議論する．

3.3 探索の実験結果の考察

入次数優先探索 (In-degree First Search; IFS) 及び幅優先探索 (Breadth First Search; BFS) を実装したクローラが探索したページの結合分布，オーソリティ及びハブ (高い出次数をもつページ) の収集効率を比較する．実験では IFS と BFS のどちらも同じページ

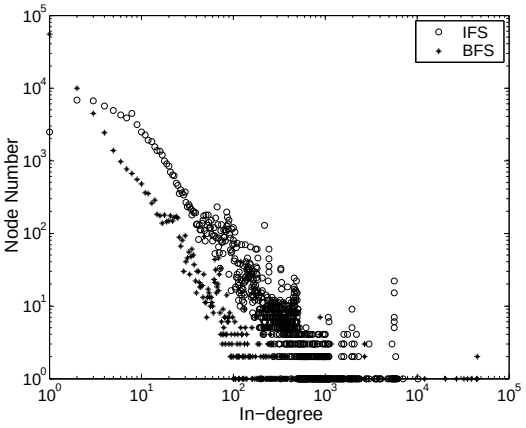

図 1 入次数に対する頂点数の分布

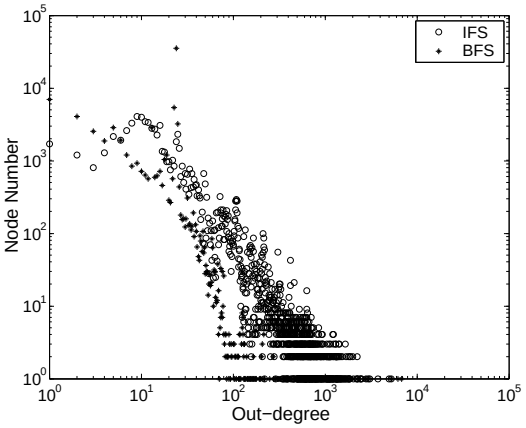

図 2 出次数に対する頂点数の分布

から探索を開始，探索ページ数が 10 万となったところで終了した．また，経路が互いに離れていると思われる 5 つのページから探索実験を行った．その結果，5 回のうち 4 回の探索実験において，IFS を用いた場合の方がオーソリティを多く収集できた．以下に示す図は http://www.jaist.ac.jp/から探索した結果である．

まず，図 1,2 は入次数及び出次数に対するページ数の分布を示している．図 1,2 から，IFS(o 印) 及び BFS(*印) を用いて探索したページの結合分布は，入次数及び出次数のどちらに対しても同等のべき係数をもっているが，IFS を用いて探索したページの結合分布を示すグラフは BFS と比べて右にシフトしていることが分かる．これは，IFS を用いて探索したページは BFS の場合と比べて多くの平均辺数をもち，IFS を用いて探索したリンク構造が密になることを示す．

次に，ある閾値以上の入次数及び出次数をもつ頂点をオーソリティ及びハブ (高い出次数をもつ頂点) とし，Web クローラが訪問したページ数を v と表記し，v に対するオーソリティ及びハブの累積獲得数を調べた．IFS 及び BFS によるオーソリティ (ハブ) の累積獲得数をそれぞれ a_{IFS} 及び a_{BFS} (h_{IFS} 及び h_{BFS}) と表記し，オーソリティ (ハブ) を決める閾値を,「IFS で探索したリンク構造から入次数 (出次数) の降順に 0.1%のページを選択し，その中に含まれるページがもつ最小の入次数 (出次数)」とした．

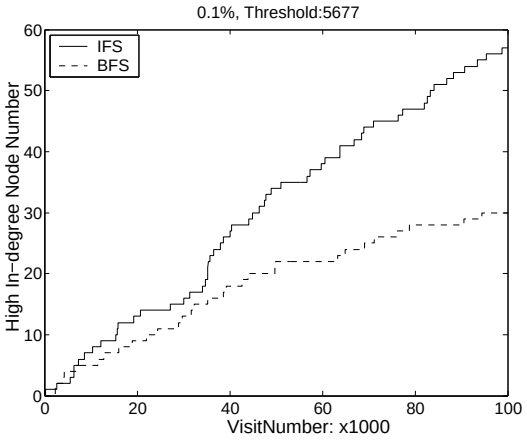

図 3 Web クローラが訪問した頂点数 v に対するハブの累積数 a_{IFS} 及び a_{BFS}

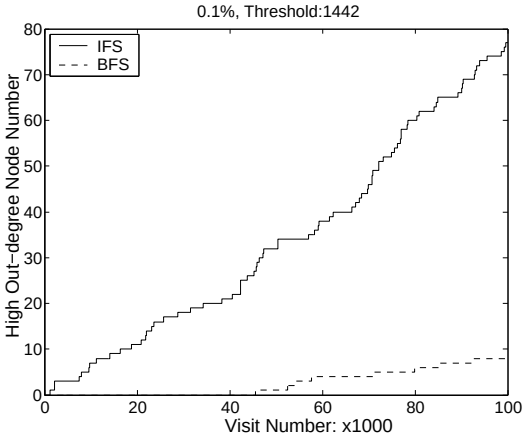

図 4 Web クローラが訪問した頂点数 v に対するハブの累積数 h_{IFS} 及び h_{BFS}

図 3 では，オーソリティの累積獲得数を比較している．探索の初期では，a_{BFS}(実線) と a_{IFS}(破線) に大きな差はないが，Web クローラが訪問した頂点数 v が増加すると，a_{IFS} がほぼ線形に増え続けているのに対し，a_{BFS} の増え方は緩やかになっており，v の増加に伴ない，その差が広がっている．10 万ページの探索が終わったところで，その差は約 2 倍になっている．

図 4 では，ハブの累積獲得数を比較している．h_{IFS}(実線)，h_{BFS}(破線) のどちらも，ほぼ線形に増加しているように見えるが，h_{BFS} の傾きは h_{IFS} に比

べて緩やかであり，探索が進むのに伴い，その差は広がり続けている．10 万ページの探索が終わったところで，その差は約 10 倍になっている．この結果は，IFS ではクローラが探索した未探索ページのの入次数のみを用いてページの探索順を決めているにも関わらず，多くの関心・興味を集めているオーソリティだけでなく，多くのページへのリンクをもつハブも効率的に探索していることを示している．

ここで，上記の結果について考察する．Pennock ら [4] は特定のトピックを扱うページを収集したとき，次数に対するページ数の分布が対数正規分布に従うことを示している．また，Web 上では特定のトピックを扱うページ同士が密に結合して Web コミュニティと呼ばれるリンク構造を構成する [5]．図 1,2 が示す IFS で探索したページの結合分布は Pennock らの結果と類似しており，IFS がコミュニティを抽出しながら，Web 上を探索することを示唆している．

次に，IFS が効率的にオーソリティ及びハブを探索する原因について考察する．近年，頂点の次数に対する最近傍の平均次数の分布が Scale-Free ネットワークの種類によって異なることが明らかにされている [6)~8)]．Newman ら [6),7)] は Web ページのような社会的ネットワークでは，頂点の次数に対する最近傍の平均次数が，正の相関をもつことを解析的に示している．また，Vázquez[8)] は実データを用いて，Web 上で頂点の次数に対する近傍の平均次数が，高い次数をもつ頂点に対して正の相関をもつことを示している．これは，オーソリティ及びハブ同士の結合頻度が高いことに対応し，社会的ネットワークで，高い次数をもつ頂点がコミュニティ間を接続する原因となり得る．そのため，高い入次数をもつ頂点の優先的な探索によって，IFS はコミュニティの核となるオーソリティ及びハブを経由しながら Web 上を探索しているもの考えられる．

4. おわりに

従来の手法では，リンク構造からページの重要度を求め，ページの探索順を決定するには，Web 全体のリンク構造を把握しなければならなかった．

そこで，効率的にオーソリティを探索するために，クローラが発見した未探索ページの入次数に従い，ページの探索順を適宜決める入次数優先探索法を提案した．

Java で実装したクローラを用いた Web 上の探索実験の結果，ある程度大きな部分ネットワークを抽出したとき，入次数優先探索は幅優先探索と比べて，

(1)　探索したリンク構造の平均辺数が多くなる
(2)　オーソリティ及びハブを多く収集する

という傾向を示した．

Web 上では，特定のトピックを扱うページが Web コミュニティを構成し，その内部のページ同士が密に結合しているため，Web コミュニティの結合分布は Web 全体と比べて，平均辺数が多くなる [5]．探索結果を示す図 1,2 から，本報告で提案した手法は，この結果に対応している．

また，Web ページのリンク構造のような社会的ネットワークで，頂点の次数と最近傍の平均次数の間の正の相関と，オーソリティ及びハブ同士の高い結合頻度には関連性があると考えられる．すなわち，高い入次数をもつページの優先的な探索によって，コミュニティの核となるオーソリティ及びハブを経由しながら Web ページを収集しているものと示唆される．

今後の課題として，提案手法で探索されたページの次数に対する最近傍の平均次数が，他の社会的ネットワークと同様に正の相関をもつかどうか，また，提案手法で探索されたリンク構造がどのようなコミュニティを含むのかを調べることなどが挙げられる．

参 考 文 献

1) 山名 早人,「IT 社会を先導するインターネット」, 信学会誌, Vol.86, No.5, pp.304-310,(2003).
2) Page, L., Brin, S., Motwani, R., and Winograd., T., “The PageRank citation ranking: Bringing order to the web,” Stanford Digital Libraries Working Paper, (1998).
3) Cho, J., Garc a-Molina, H., and Page., L., “Efficient crawling through URL ordering,” Comp. Net. and ISDN Sys., Vol.30, pp.161-172,(1998).
4) Pennock, D., et al., “Winners don’t take all: Characterizing the competition for links on the web,” Proc. of the Nat. Acad. of Sci., Vol.99, No.8, pp.5207-5211,(2002).
5) J. M. Kleinberg, “Authoritative Sources in a Hyper-linked Environment,” Proc. of the ACM-SIAM Symp. on Discrete Algorithms,(1998).
6) Newman, M.E.J., “Mixing patterns in networks,” Phys. Rev. E, Vol.67, 026126,(2003).
7) Newman, M., and Park, J., “Why social networks are different from other types of networks,” cond-mat/0305612, (2003).
8) Vázquez, A., “Growing network with local rules: Preferential attachment, clustering hierarchy, and degree correlation,” Phys. Rev. E, Vol.67, 056104, (2003).