局所的なリンク構造に基づくコミュニティ構造の抽出法

松久保 潤[†] 林 幸雄[†]

Web 上には，類似する話題を扱う頁から成る密に結合したリンク構造がある．このようなリンク構造は Web コミュニティに対応するものだと考えられる．本論文では，コミュニティの核となる高い次数（リンク数）を持つ頁の効率的な収集を検討する．Web のリンク構造は急激に変化しており，全体のリンク構造を把握することが困難であるため，探索中に発見された頁のみから成る局所的なリンク構造上で，未探索かつ最大の入次数を持つ頁を最優先に収集する手法を提案する．提案手法を用いて，実際の Web 上で探索実験を行った結果，高い次数を持つ頁を優先的に経由することで，異なる話題を扱うコミュニティを横断することを見つけた．

An Extracting Method for Community Structure Based on Local Link Structure

Jun Matsukubo[†] and Yukio Hayashi[†]

Since the World Wide Web has link structures densely connected with similar topics, a set of the pages consists of a Web community. In particular, the pages with high in-degrees are some cores of topics. We investigate how to collect the pages, and we propose an on-line exploring method by using only local information, in-degree of pages. This method preferntially collect the pages with high in-degrees. In an experience, for several real data, we find traverses between Web communities of topics.

1. はじめに

近年，電力網，インターネットのルータの接続関係，Web，P2P のファイル共有システム，電子メールの送受信関係，論文の引用関係，論文の共著関係，俳優の共演関係，線虫の神経回路網，たんぱく質の相互作用など，現実の多くのネットワークのリンク構造がスケール・フリー（Scale-Free; SF）性を共通に持つことが明らかにされてきた[1]．SF 性は，次数（各頂点に接続されている辺数）に対する頂点数の分布がべき乗則に従うことで特徴付けられる．すなわち，高い次数を持つ（多くの辺が接続されている）頂点数が極少数であるのに対し，低い次数を持つ頂点が大多数を占める．

一方，SF 性を持つネットワークにおいて，コミュニティ構造が議論されてきた[2),3)]．たとえば，Web 上では，**関連する話題を扱う頁から成る密なリンク構造**が存在することが知られている．このようなリンク構造は **Web コミュニティ**[4)] をなすものと考えられる．Web コミュニティが現れる原因は，共通の興味・関心を持つ頁の作成者や利用者がつながりやすいという傾向によるものと考えられる．さらに，利用者のアクセスの立場からは，情報を得ることに関心があるので，企業と個人の頁を特に区別せず，関連する話題を扱う頁の集合で定義する．

コミュニティ構造とは，頂点どうしが密に結合して形成される局所的なリンク構造を指す．また，このような構造は，高い次数を持つ頂点を核として構成されていると考えられる．

Web コミュニティの構造解析は，Web 上の情報検索，Web の成長メカニズムの解明，地域情報空間の構造の解明[5)] などの応用につながる重要な課題である．本論文では，Web コミュニティの核となると考えられる高い次数を持つ頁を効率的に収集することを検討する．

関連する話題として，Web や GNUTELLA[6)] などの P2P ファイル共有システムで構成されるネットワークのリンク構造は急激に変化しているため，探索のためにネットワーク全体のリンク構造をつねに把握しておくことは困難である．そこで，SF ネットワーク上で，局所的なリンク構造のみから探索順を決める手法

† 北陸先端科学技術大学院大学知識科学研究科
School of Knowledge Science, Japan Advanced Institute of Science and Technology

が提案されている[7)~11)].

まず，Cho ら[7)] は Web 頁の効率的な再探索のために，リンク構造に基づいて各頁に対する探索の優先度を決める手法を提案した．実際の Web 上における探索実験の結果，Cho らの手法が幅優先探索やランダム・ウォークなどの従来法よりも効率的に頁を再探索できることが示された．

Cho らの手法では，探索中に発見されたすべての頂点を探索候補とするのに対し，以下の手法では，効率的な経路探索や頂点のカバーリングを目的として，探索された頂点の最近傍のみを探索候補とする．

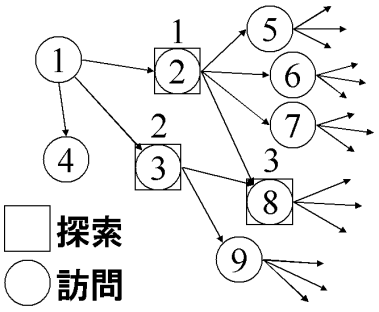
Tadic[8)] は Web 上で任意の 2 頂点間を結ぶ経路の効率的な探索のために，各ステップの探索候補として，最近傍をその入次数に比例した確率で選択する手法を提案した．Web を模倣した人工的なネットワーク上での探索実験を行った結果，Tadic の手法は，任意の 2 頁間を結ぶ経路をランダム・ウォークよりも効率的に探索できることを示した．

また，Adamic ら[9)]，Kim ら[10)] は GNUTELLA などの P2P ファイル共有システムで構成されるネットワーク上で，各ステップで探索した頂点の最近傍の中から，最大の次数を持つ頂点を最優先に探索する手法を提案した．人工的に生成した SF ネットワーク上での探索実験の結果，ランダム・ウォークよりも短い経路で任意の 2 頂点間を探索できることを示した．

他方，Ikeda ら[11)] は有限個の頂点で構成される SF ネットワーク上で，すべての頂点を探索するために必要なステップ数を理論的に解析した．彼らは n 個の頂点から成るネットワーク上をランダム・ウォークで探索したとき，ステップ数が $O(n^3)$ となるのに対し，隣接頂点の次数に逆比例した確率で，次の探索先を選択する手法を提案し，ステップ数が $O(n^2)$ となることを示した．

以上の研究では，あらかじめ**探索された頂点の最近傍の次数が既知であることを前提としている**．一方，膨大かつ逐次変化する Web 頁を探索するために上記の手法は適用できない．

本論文では，**探索中に発見される頂点のみで構成される局所的なリンク情報のみを用いて，最大の入次数を持つ頂点を最優先に探索する**手法を提案する．また，提案手法と幅優先探索法のそれぞれを用いて，高い次数を持つ頁の収集効率を比較する．実際の Web 上での探索実験の結果，提案手法を用いた探索で，高い次数を持つ頁を優先的に経由しながら，異なる話題を扱う Web コミュニティを横断するような挙動が見られた．



□外の数は探索（Web 頁のダウンロード）順，○内の数は訪問（URL の確認のみ）順を指す．

図 1 過去の履歴を活用する探索法
Fig. 1 A searching method employing the searching history.

表 1 局所的なリンク構造に基づく探索の従来法
Table 1 Conventional searching method based on local link structures.

	WWW	経路探索	カバーリング
(i)	Cho, et al. 本提案手法		
(ii)	Tadic	Adamic, et al., Kim, et al.	Ikeda, et al.

まず，2 章で，局所的なリンク構造から探索順を決める従来法について述べ，3 章で本論文の提案手法を説明する．次に，4 章で提案手法と幅優先探索を用いた探索実験の結果を比較する．5 章では結論を述べる．

2. SF ネットワーク上の効率的な探索の従来法

Web や GNUTELLA のようなファイル共有システムなどから構成されるネットワークのリンク構造は，大規模かつ逐次変化する．そのため，リンク構造全体の把握が困難である．本章では，効率的な Web 頁の再探索，任意の 2 頂点間の経路探索，および頂点のカバーリングなどのために，局所的なリンク構造のみから探索順を決める手法について説明する．これらの手法には，探索候補を (i) 過去の履歴を活用して決めるもの（**図 1**）と，(ii) 各ステップでその時探索された頂点の最近傍から選択するものがある．探索候補の選択法，および対象とするネットワークによる分類を**表 1** に示す．

まず，(i) の手法として Cho ら[7)] は，Web 上で頁の再探索を効率化するために，リンク構造に基づいて各頁に割り当てられた重要度順に頁を探索する手法を提案した．スタンフォード大学の Web サイト内で探索実験を行った結果，PageRank（Google での頁評価法）[12)] を重要度としたとき，ランダム・ウォークや幅

優先探索よりも高い入次数を持つ頁を効率的に探索できることが示された．

次に，経路探索や頂点のカバーリングに対する (ii) の手法として，Tadic[8]，Adamic ら[9]，および Kim ら[10] の手法がある．

Tadic は，Web 上での効率的な経路探索のために，探索候補を各ステップで探索された頁から参照される（リンクが張られている）最近傍が持つ入次数に比例する確率で選択する手法を提案した．Web を模倣した人工的なネットワーク上で，ランダム・ウォークとの経路探索の性能を比較した結果，Tadic の手法を用いた場合，ランダム・ウォークより少ないステップで，長い経路で連結している頁に到達できることが示された．

Adamic らは GNUTELLA のようなファイル共有システムなどから構成される P2P ネットワーク上での効率的な経路探索法を提案した．まず，彼らは次数分布がべき乗則に従う，P2P 関係に対応する人工的なネットワーク上で，低い次数を持つ頂点が高い次数を持つ頂点（ハブ）に結合している頻度が高いことを解析的に示した．この結果から，任意の 2 頂点がハブを介して連結している頻度が高いと考えられる．彼らは各ステップで探索した頂点の最近傍から次のステップの探索候補として，最大の次数を持つものを選択する手法を提案した．Tadic の手法と Adamic らの手法はどちらも (ii) に従っているが，Tadic の手法を用いた場合，探索候補が確率的に選択されるのに対し，Adamic らの手法を用いた場合，次数の降順に選択される．Adamic らは，経路探索の性能を人工的に生成したネットワーク上でランダム・ウォークと比較し，彼らの手法が任意の 2 頂点間をランダム・ウォークよりも短い経路で探索できることを示した．また，ランダム・ウォークより効率的に多くの頂点へのリンクを発見できることが示された．

次に，Kim らは Adamic らと同様の手法を用いて，人工的に生成したネットワーク上で，探索される経路長が SF ネットワーク上の任意の 2 頂点間を結ぶ最短経路長と同様に $O(\log n)$（n はネットワークの頂点数）となることを示した．

Densmore[13] は探索手法の違いによる経路探索の性能を比較するために，人工的に生成されたネットワーク上で実験を行った．その結果，Adamic らの手法を用いた場合，平均変数の増加にともない，幅優先探索やランダム・ウォークを用いた場合よりも急激に経路長が短くなることを示した．

上記の手法はいずれも，各ステップでそのとき探索した頂点の最近傍の次数が既知であることを前提としているが，大規模で逐次変化する Web 上での探索には適切でない．

次章では，実際の Web 上でコミュニティの核になると考えられる高い次数を持つ頂点を効率的に収集するために，探索中に発見された頁のみから成るリンク構造に基づいて探索順を決める手法を提案する．

3. 入次数優先探索

本章では，Web コミュニティの核になると考えられる高い次数を持つ頁の効率的な収集のために，**探索中に発見された頁から成る局所的なリンク構造上で，適宜，未探索かつ最大の入次数を持つ頁を最優先に探索する**手法を提案する．提案手法を用いた探索では，(i) に従う Cho 達の手法と同様に，発見されたすべての未探索頁を探索候補とする．また，探索が進むと被参照リンクが見つかり，入次数が増える頁があるので，探索候補の優先順位はメタ・ヒューリスティクス[14] のように適応的に変わる．以下，本提案手法を入次数優先探索（In-degree First Search; IFS）と呼ぶ．

IFS を用いた探索の概要は以下のとおりである．まず，探索候補となる HTML のダウンロード・構文解析を行い，他の頁を参照する URL（リンク）を抽出して未探索リストに格納する．続いて，未探索リスト内で最大の入次数を持つ URL に対応する頁を次の探索候補とする．以下，この処理を繰り返す．

次に，IFS を用いた探索の手順を示す．

Step 1. 探索 URL p に探索開始 URL を格納
　探索キュー U_{s} を空に設定
　重複チェック用リスト U_{dup} を空に設定
　各頁の入次数リスト D_{in} を空に設定
Step 2. p に対応する HTML から
　重複なしで抽出された URL を
　一時 URL リスト U_{t} に格納
Step 3. U_{t} の先頭の URL を p_{t} に格納
　U_{t} から p_{t} を削除
Step 4. (a) p_{t} が U_{dup} に含まれる場合：
　p_{t} の入次数 $D_{\mathrm{in}}(p_{\mathrm{t}})$ を 1 だけ増加
　(b) p_{t} が U_{dup} に含まれない場合：
　U_{s} に p_{t} を追加
　U_{dup} に p_{t} を追加
　p_{t} の入次数 $D_{\mathrm{in}}(p_{\mathrm{t}})$ を 1 に設定
Step 5. U_{t} が空でなければ，Step 3. から繰り返す
Step 6. $D_{\mathrm{in}}(p)$ の最大の要素に対応する
　$p \in U_{\mathrm{s}}$（最大の入次数を持つ URL）を
　p に格納
　U_{s} から p を削除 D_{in} から p を削除

表 2 実験で用いた探索開始 URL（http://を省略）
Table 2 Search-starting URLs used on the experiments ("http://" is dropped).

1	www.jaist.ac.jp/	北陸先端大
2	www.whitehouse.gov/	ホワイトハウス
3	www.tamuranaomi.com/	田村直美（個人運営サイト）
4	www.stanford.edu/	スタンフォード大学
5	www.parliament.uk/	イギリス王立図書館
6	www.ndl.go.jp/	国会図書館
7	www.kakaku.com/	価格ドットコム
8	www.msu.ru/	ロシアの新聞社サイト
9	www.foreignaffairsj.co.jp/	FOREIGN AFFAIRS JAPAN（米国外交専門誌）
10	www.pozaman.com/	ぽざまん（個人運営サイト）

Step 7. Step 2. から繰り返す

4. 高い次数を持つ頁の収集効率および探索されたリンク構造の比較

ここでは，IFS と幅優先探索（Breadth First Searching; BFS）のそれぞれを用いて，実際の Web 上で探索実験を行い，探索されるリンク構造の特徴および探索の振舞いを比較する．

BFS は，IFS とは異なる目的で提案されたが，BFS を用いた探索で，高い PageRank を持つ頁を効率的に収集できることが実験的に示されている[15]．

表 2 に示すように，Web 上で相互に離れていると思われる 10 種類の URL（BFS と IFS の両方で共通）から開始し，収集頁数は 10 万とした．いずれの URL から探索を開始した場合も同様の結果が得られたので，以下では北陸先端大からの探索結果について説明する．

比較項目は，探索されたリンク構造に対して，

(1) 収集頁数に対する In-Hub（高い入次数を持つ頁），および Out-Hub（高い出次数を持つ頁）の累積獲得数
(2) 頁の次数分布
(3) 頁の次数に関する結合相関

とした．また，探索の振舞いに対して，

(4) IFS および BFS のそれぞれを用いたときの In-および Out-Hub を含むドメイン間の移動の様子
(5) In-および Out-Hub を含むドメインの特性

とした．ここで，結合相関とは，任意の頁が持つ入次数に対するその頁から参照される頁の平均入次数である．

4.1 収集頁数に対する In-および Out-Hub の累積獲得数

IFS と BFS で収集した In-および Out-Hub の累積獲得数をそれぞれ**図 2**，**図 3** に示す．

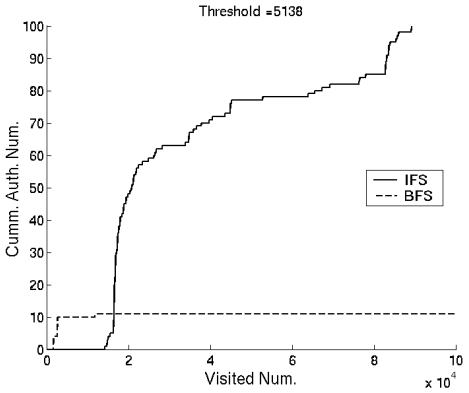


図 2 ステップ数に対する In-Hub の累積獲得数
Fig. 2 Accumulative number of In-Hub to step one.

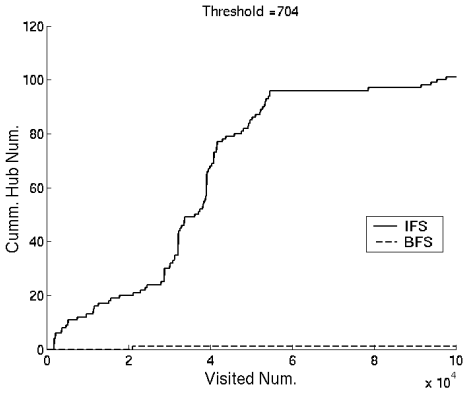


図 3 ステップ数に対する Out-Hub の累積獲得数
Fig. 3 Accumulative number of Out-Hub to step one.

ここでは，In-Hub（Out-Hub）を閾値以上の入（出）次数を持つ頁とした．閾値は，IFS および BFS のどちらの結果に対しても，IFS で探索された頁の降順で 100 番目の入（出）次数とした．In-および Out-Hub の閾値はそれぞれ 5138 および 704 となった．

図 2，図 3 において，BFS の結果では，実験の初期でのみ In-および Out-Hub が探索されたのに対し，IFS の結果では，中盤以降でも探索されたことが分かる．探索が終了した時点で，IFS を用いた場合の In-および Out-Hub の累積獲得数はそれぞれ，BFS を用いた場合の約 10 倍，および約 50 倍となった．

4.2 入次数および出次数に対する頁数の分布

入次数および出次数に対する頁数の分布を比較した結果を**図 4**，**図 5** に示す．それぞれの図の横軸は入次数および出次数を示しており，縦軸はどちらも頁数に対応している．どの軸も対数スケールで表されている．また，IFS（o 印）と BFS（*印）の実測値，べき乗則

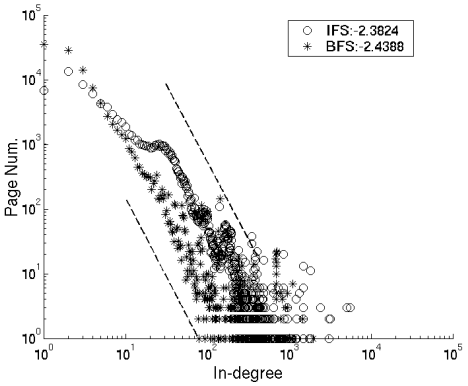


図 4 入次数に対する頁数の分布の比較
Fig. 4 Comparison of the distribution of pages to In-degree.

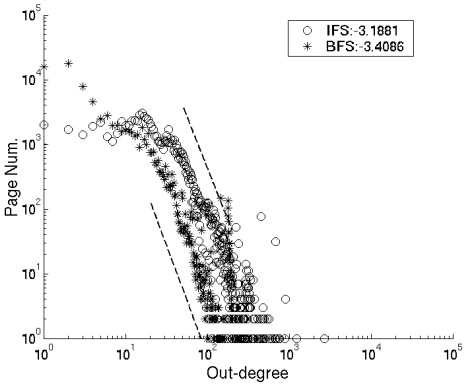


図 5 出次数に対する頁数の分布の比較
Fig. 5 Comparison of the distribution of pages to Out-degree.

に従う直線（破線）を描画している．

まず，IFS および BFS のどちらを用いた場合にも，分布の中央付近にべき乗則に従う部分がみられ，その直線の傾きを表す，べき指数の値は，IFS および BFS を用いた場合で，入次数に対して約 −2.38（図 4 上側の破線）および −2.44（図 4 下側の破線），出次数に対して約 −3.19（図 5 上側の破線）および −3.41（図 5 下側の破線）となった．

次に，IFS の結果で，中央付近の分布が BFS の結果よりも右にシフトしていることから，IFS を用いた場合の頁の平均辺数が BFS を用いた場合よりも多くなったことが分かる．この結果は，IFS が BFS と比べ，より密に接続された部分グラフを探索したことを示す．

また，分布の左側では，IFS が BFS よりも，べき乗則から大きく外れている．この部分の分布は，関連する話題を扱う頁の次数分布が対数正規分布に従うという Pennock たちの結果[16] に対応するものと考えられる．

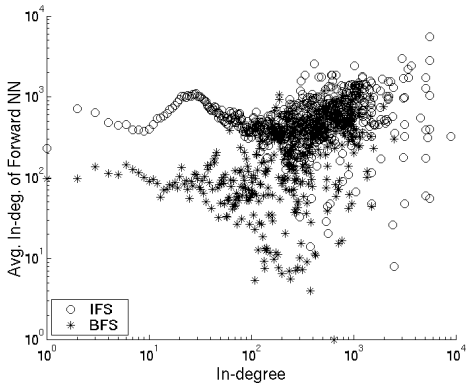


図 6 入次数に対する最近傍が持つ平均入次数の分布の比較
Fig. 6 Comparison of the distribution of the average In-degree of neighbors to one of a page.

4.3 頁の次数に関する結合相関

IFS および BFS のそれぞれを用いて探索されたリンク構造上の結合相関を比較する．

図 6 に，任意の頁の入次数 k_{in} に対するその頁からのリンク先頁（前方の最近傍）の平均入次数 $\langle k_{\mathrm{in}}^{\mathrm{NN}}\rangle(k_{\mathrm{in}})$ の分布を示す．

まず，IFS を用いて探索されたリンク構造の平均結合相関が全体的に高くなっている．また，BFS を用いたとき，分布が中央付近でばらついており，ほぼ無相関であるのに対し，IFS を用いたとき，100 以上の入次数に対して緩く右上がりになっており，弱い正の相関があることが分かる．

この結果から，IFS によって探索されたリンク構造では，BFS の結果より In-Hub 間の結合頻度が高いことが分かる．正の結合相関は Web 以外にも知人関係，俳優の競演関係，および論文の共著関係などの社会的ネットワーク上で，現れることが明らかにされている[2),3)]．そのため，正の結合相関はハブ間の結合頻度が高いことを示し，コミュニティの出現に関連するものと考えられる．

以上の結果から，IFS を用いた探索では，Web のリンク構造で結合相関が正である性質を利用し，コミュニティ構造の核となる In-および Out-Hub を優先的に経由しながら密なリンク構造を収集するものと考えられる．

4.4 In-および Out-Hub を含むドメイン間の移動の様子

次に，IFS および BFS を用いた場合の Web コミュニティ間の移動の様子を解析した．ここでは，IFS と BFS をそれぞれ用いて収集した In-および Out-Hub

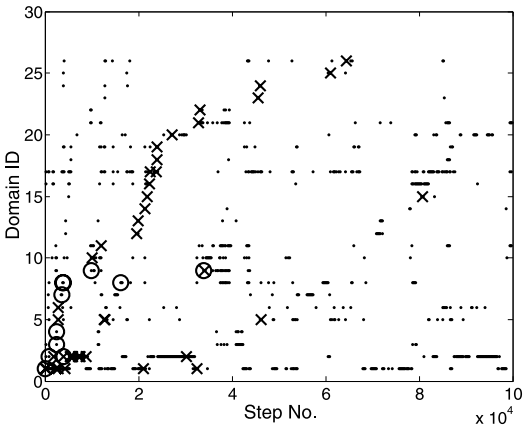


図 7 BFS を用いた場合のステップ数に対する In-および Out-Hub を含むドメイン ID の対応
Fig. 7 Correspondence of domain ID No. involving In- and Out-Hub to step No. with BFS.

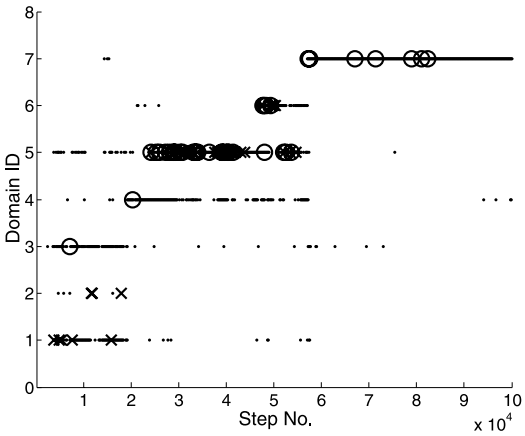


図 8 IFS を用いたステップ数に対する In-および Out-Hub を含むドメイン ID の対応 1
Fig. 8 Correspondence 1 of domain ID No. involving In- and Out-Hub to step No. with IFS.

を，入次数および出次数の降順でそれぞれ 50 個の頁として，探索ステップ数と In-および Out-Hub が含まれるドメインの対応を示す．

図 7，**図 8** では，横軸および縦軸をそれぞれ，探索頁数，および探索された In-および Out-Hub が属するドメインに割り当てた ID（**表 3** および**表 4** の ID に対応）とした．また，o 印，x 印および．印はそれぞれ，各 ID のドメインに属する In-Hub，Out-Hub，およびその他の頁が探索されたことを表す．

BFS を用いた図 7 では，小さな値の ID に対応する（初期に探索された）ドメインが，大きな値の ID に対応する（中盤以降に探索された）ドメインよりも重点的に探索されている．この結果は，BFS を用いた探索では，初期に探索したドメインへの探索頻度が高くなることを示す．また，Out-Hub が多くのドメイン数に分散している．さらに，各ドメインの探索順が不規則になっている．

これに対し，IFS を用いた図 8 では，In-および Out-Hub を含むドメインが BFS を用いた場合よりも少ない．すなわち，In-および Out-Hub が少数のドメインに密集している．また，In-および Out-Hub が収集された後，その頁を含むドメインが重点的に探索されている．さらに，In-および Out-Hub の探索にともない，それらの頁を含むドメインに重点的な探索が遷移する場合がある．図 8 では，20286 番，および 57125 番のステップで，4 番と 7 番のドメインにおいて In-Hub が収集され，これにともない，重点的な探索が 4 番と 7 番のドメインに遷移したと推測される．次節では，In-および Out-Hub を含むドメインで扱われる話題について解析する．

表 3 BFS を用いた場合の In-Hub および Out-Hub を含むドメイン（到達順）
Table 3 Domains including In- and Out-Hubs with BFS (in arriving order).

(1)	(2)	(3)	(4)	(5)	(6)
1	jaist.ac.jp	1	1	7	13,221
2	crl.go.jp	283		2	2,820
3	nii.ac.jp	289		2	3,622
4	mext.go.jp	293	11	15	16,402
5	yusei.go.jp	317	15		1,339
6	acs.org	784	1		1,898
7	affrc.go.jp	803		1	122
8	asahi.com	809	7		1,639
9	titech.ac.jp	811		4	224
10	nsknet.or.jp	1,558		1	1,035
11	yahoo.co.jp	2,368	14	2	2,679
12	biglobe.ne.jp	2,412		1	509
13	japanpost.jp	2,536	1		440
14	memphis.edu	3,612		1	21
15	shiga-med.ac.jp	3,957		1	80
16	jikei.ac.jp	3,982		1	69
17	keio.ac.jp	4,000		1	419
18	shizuoka.ac.jp	4,046		1	64
19	isis.ne.jp	4,363		1	882
20	apple.co.jp	4,399		1	435
21	mielparque.or.jp	5,089		2	414
22	joyjoy.com	5,577		1	138
23	jishin.go.jp	6,796		1	1,006
24	geocities.co.jp	9,824		1	353
25	cafeglobe.com	9,834		1	351
26	fujita-hu.ac.jp	12,752		1	20
27	dnp.co.jp	21,337		1	240

(1) ドメイン ID（図 7 に対応），(2) ドメイン名，
(3) 到達ステップ数，(4) In-Hub Num.（無記入は 0），
(5) Out-Hub Num.（無記入は 0），(6) 探索頁数

4.5 In-および Out-Hub を含むドメインの特性

IFS および BFS を用いたときの In-および Out-Hub

表 4 IFS を用いた場合のドメインの遷移と話題の対応 1（www.jaist.ac.jp: JAIST）
Table 4 Correspondence 1 of domain transition to topic with IFS.

(1)	(2)	(3)
(1〜20285)		
1: coverpages.org	2,394	SGML/XML の解説
2: zdmcirc.com	5	英文雑誌の予約購読
3: w3.org	1,779	Web 技術の広報
（その他）	(16,107)	
∗pcmag.com	2,228	PC および周辺機器
∗designer-info.com	869	CG 作品およびツール
（コミュニティのサイズ）	(20,285)	
(20286〜57124)		
4: hp.com	7,152	ヒューレット・パッカード
5: sun.com	16,591	サン・マイクロシステムズ
6: bea.com	3,717	基盤ソフトの開発
（その他）	(9,380)	
∗weblogic.com	1,648	Weblogic の解説
∗cisco.com	928	シスコ（ネットワーク機器）
（コミュニティのサイズ）	(36,839)	
(57125〜100000)		
7: eu.int	37,603	欧州連合
（その他）	(5,272)	
∗ecb.int	679	欧州中央銀行
∗cordis.lu	609	EU 研究開発
（コミュニティのサイズ）	(42,876)	

(1) ドメイン名，
nnn_1〜nnn_2 はコミュニティとする探索ステップ数，
各ドメイン名の先頭の数は図のドメイン ID に対応，
∗ は In-hub と Out-hub を含んでおらず，
かつ最多頁数を持つ上位 2 つのドメイン，
(2) 頁数，(3) ドメインの概要，表 5，表 6 でも同様．

が含まれるドメインの特性についての解析結果を表 3，表 4 に示す．これらの表には，北陸先端大（jaist.ac.jp）から探索を開始した結果を示している．

まず，表 3 の BFS による結果では，比較的近い経路上にあると思われる研究所（crl.go.jp，nii.ac.jp），官庁（mext.go.jp），学会（acs.org），他大学（titech.ac.jp，memphis.edu，shiga-med.ac.jp，jikei.ac.jp，keio.ac.jp，shizuoka.ac.jp）などのドメインが探索されている．一方，Web 上で北陸先端大と関連が薄いと思われる asahi.com と yahoo.co.jp で In-および Out-Hub が探索された原因は，これらのドメインが広く人気を集めているので，様々な頁から参照されているためだと考えられる．

次に，IFS を用いた場合の結果について，In-および Out-Hub の収集にともない，重点的な探索ドメインが遷移したステップで分割された頁の集合を解析する．図 7 の 4 番と 7 番のドメインで In-Hub が収集された 20286 番目，および 57125 番目のステップで，収集された頁の集合を 3 つに分割した．このとき，In-および Out-Hub を含むドメインは表 4 に示すように，Web の技術に関する話題を扱う 1〜3，ネットワーク技術に関する話題を扱う 4〜6，および EU に関連する 7 に分けられる．3 つ集合の各頁数に対し，In-および Out-Hub を含むドメインに属する頁数の割合を求めると，それぞれ約 42%，75%，88%となり，これらのドメインが各集合内の多くの頁を占めていることが分かる．

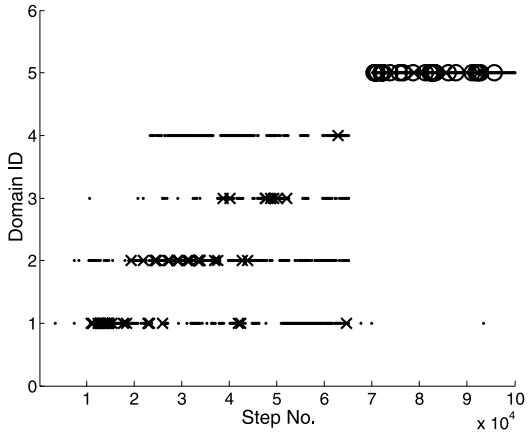


図 9 IFS を用いた場合のステップ数に対する In-および Out-Hub を含むドメイン ID の対応 2
Fig. 9 Correspondence 2 of Domain ID No. involving In- and Out-Hub to step No. with IFS.

一方，上記以外の多くの頁を含むドメインでも関連する話題が扱われていた（表 3 には各集合で多くの頁を含む上位 2 つのドメインを（その他）として併記）．また，他の実例として，www.tamuranaomi.com，および www.foreignaffairsj.co.jp の結果でも同様の傾向が見られたことをそれぞれ**図 9**，**表 5** および**図 10**，**表 6** に示す．

まず，図 9，表 5 では，5 番のドメインで In-Hub が収集された 70377 番目のステップで，重点的な探索ドメインが遷移したと思われる．このステップで頁の集合を分割すると，In-および Out-Hub が属するドメインは，ネットワークとソフトウェアに関する話題を扱う 1〜4 とコンピュータのハードに関する話題を扱う 5 に分けられる．2 つ集合の各頁数に対して In-および Out-Hub を含むドメインに属する頁数の割合を求めると，それぞれ約 64%，74%となった．

また，図 10，表 6 では，2 番と 9 番のドメインで In-Hub が収集された 9226 番目，51251 番目のステッ

表 5 IFS を用いた場合のドメインの遷移と話題の対応 2（www.tamuranaomi.com: 田村直美）
Table 5 Correspondence 2 of domain transition to topic with IFS.

(1)	(2)	(3)
(1〜70377)		
1: sun.com	15,054	サン・マイクロシステムズ
2: bea.com	19,725	基盤ソフトの開発
3: beasys.com	7,458	基盤ソフトの開発
4:weblogic.com	2,969	Weblogic の解説
（その他）	(25,171)	
∗hp.com	6,630	ヒューレット・パッカード
∗nortelnetworks.com	2,463	ネットワーク機器
（コミュニティのサイズ）	(70,377)	
(70378〜100000)		
5: pcworld.com	22,023	情報関連ニュース
（その他）	(7,600)	
∗amd.com	2,426	AMD（CPU 製造）
∗microsoft.com	1,308	マイクロソフト
(Total)	(29,623)	

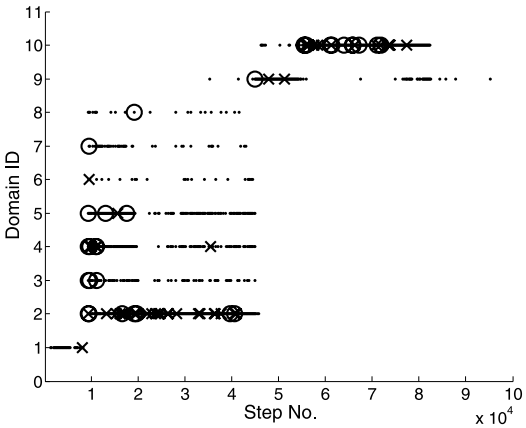


図 10 IFS を用いた場合のステップ数に対する In-および Out-Hub を含むドメイン ID の対応 3
Fig. 10 Correspondence 3 of Domain ID No. involving In- and Out-Hub to step No. with IFS.

プで，重点的な探索ドメインが遷移したと思われる．このステップで頁の集合を分割すると，In-および Out-Hub が属するドメインは，PHP に関する話題を扱う 1，建築物の設計に関する話題を扱う 2〜8 と米国政府に関する 9，10 に分けられる．3 つの集合の各頁数に対して In-および Out-Hub を含むドメインに属する頁数の割合を求めると，それぞれ約 38%，74%，96%となった．

上記のように，In-および Out-Hub の収集にともない，重点的な探索ドメインが遷移したステップで分割した頁の集合は，それぞれ関連する話題を扱う多くの頁を含むので，各集合が頁の内容からもコミュニティとなっているものと推測される．また，各集合で In-および Out-Hub を含むドメインが大半の頁を占めることから，コミュニティは In-および Out-Hub を核として構成されていると考えられる．さらに，各集合で異なる話題が扱われていることから，提案手法を用いた探索が，In-および Out-Hub を転換点として，異なる話題を扱うコミュニティを横断することが示唆される．

表 6 IFS を用いた場合のドメインの遷移と話題の対応 3（www.foreignaffairsj.co.jp: FOREIGN AFFAIRS JAPAN（ニュース・サイト））
Table 6 Correspondence 3 of domain transition to topic with IFS.

(1)	(2)	(3)
(1〜9225)		
1: php.net	3,496	php 公式
（その他）	(5,729)	
∗zend.com	2,662	php アプリケーション制作
∗gnu.org	829	Free Software Foundation
（コミュニティのサイズ）	(9,225)	
(9226〜51250)		
2: DesignCommunity.com	25,066	建築物ポータル
3: artifice.com	122	CAD ソフト制作会社
4: GreatBuildings.com	1,726	建築物ポータル
5: ArchitectureWeek.com	4,207	建築物ニュース
6: designcommunity.com	75	建築物ポータル
7: greatbuildings.com	79	建築物ポータル
8: cadoutpost.com	28	CAD 作品紹介
（その他）	(10,722)	
∗epa.gov	5,806	米国環境保護庁
∗cnn.com	149	CNN
（コミュニティのサイズ）	(42,025)	
(51251〜100000)		
9: epa.gov	2,235	米国環境保護庁
10: nih.gov	21,912	米国厚生省
（その他）	(1,905)	
∗eu.int	17,560	欧州連合
∗cdc.gov	4,774	米国防疫センター
（コミュニティのサイズ）	(48,750)	

5. おわりに

本論文では，コミュニティの核になると考えられる高い次数を持つ頁の効率的な収集法を検討とした．Web のリンク構造の急激な変化のため，全体のリンク構造をつねに把握しておくことは困難であるので，探索中に発見された頂点のみから成る局所的なリンク構造上で，適宜，未探索かつ最大の入次数を持つ頂点を最優先に探索する手法を提案した．

提案手法と幅優先探索のそれぞれを用いて，実際の

Web 上で探索実験を行った結果，提案手法を用いた場合に，高い次数を持つ頁の累積獲得数が多くなった．すなわち，提案手法を用いた探索で，効率的に高い次数を持つ頁を収集できた．

次に，探索されたリンク構造の特徴を解析した．その結果，提案手法で探索されたリンク構造では，幅優先探索を用いた場合と比較して，次の傾向が見られた．まず，平均辺数が高くなった．これは，提案手法を用いた探索が，Web の密な部分構造の頁を収集したことを示す．次に，ある頁の次数とその頁が参照している頁の次数の間に高い結合相関が現れた．すなわち，提案手法で探索されたリンク構造上で，高い次数を持つ頁間の結合頻度が高くなっていた．これが，提案手法を用いた探索で，効率的にコミュニティ核を収集できる原因の 1 つであると考えられる．

続いて，それぞれの手法を用いた探索におけるドメイン間の横断の様子を解析し，次の結果を得た．

幅優先探索を用いた場合，まず，探索開始頁から比較的近い経路上にあると思われるドメインが重点的に探索されていた．また，ドメインが探索される順番は不規則であった．さらに，探索開始頁のドメインと関連がない話題を扱うドメインも多く探索されていた．

一方，提案手法を用いた場合，まず，高い次数を持つ頁を含むドメインが，他のドメインよりも重点的に探索された．また，高い次数を持つ頁の探索にともない，重点的に探索されるドメインが遷移する場合があった．さらに，探索ドメインが遷移しない間，共通の話題に関連する頁を含むドメインが探索された．

以上の結果から，提案手法を用いた探索が，高い次数を持つ頁を転換点として，異なる話題を扱うコミュニティを横断することが示唆される．

最後に，これらの結果から，局所的なリンク構造に基づく探索法の有効性が，任意の 2 頂点間の経路探索やネットワーク全体の巡回の効率化などに対してだけでなく，Web 上におけるコミュニティ構造の抽出に対しても示されたものと考えられる．

謝辞 本研究の一部は文部科学省科学研究費補助金 13680404 の援助を受けている．

参 考 文 献

1) Barabasi, A.: *Linked: The New Science of Networks*, Perseus (2002).
2) Newman, M.E.J. and Park, J.: Why social networks are different from other types of networks, *Phys. Rev. E* (2003).
3) Vázquez, A.: Growing network with local rules: Preferential attachment, clustering hierarchy, and degree correlation, *Phys. Rev. E*, Vol.67 (2003). 05104.
4) Flake, G., Lawrence, S. and Giles, C.: Efficient Identification of Web Communities, *Proc. KDD-2000*, pp.150–160, ACM Press (2000).
5) 上林憲行：インターネット空間を包含した地域社会デザイン WEBLOGY と地域情報空間，情報処理学会研究報告情報メディア，Vol.2001, No.33 (2001).
6) Gnutella. http://www.gnutella.com/
7) Cho, J., Molina, G.H. and Page, L.: Efficient crawling through URL ordering, *Comp. Net. and ISDN Sys.*, Vol.30, pp.161–172 (1998).
8) Tadic, B.: Adaptive random walks on the class of Web Graphs, *Eur. Phys. J. B*, Vol.23, pp.221–228 (2001).
9) Adamic, L.A., Lukose, R.M., Puniyani, A.R. and Huberman, B.A.: Search in power-law networks, *Phys. Rev. E*, Vol.64, No.4 (2001). 046135.
10) Kim, B., Yoon, C.N., Han, S. and Jeong., H.: Path finding strategies in scale-free networks, *Phys. Rev. Lett.*, Vol.65 (2002). 027103.
11) Ikeda, S., Kubo, I., Okumoto, N. and Yamashita, M.: Impact of Local Topological Information on Random Walks on Finite Graphs, *LNCS*, Vol.2719, pp.1054–1067 (2003).
12) Page, L., Brin, S., Motwani, R. and Winograd, T.: The PageRank citation ranking: Bringing order to the web, *Stanford Digital Libraries Working Paper* (1998).
13) Densmore, O.: An Exploration of Power-Law Networks. http://backspaces.net/PLaw/
14) 柳浦睦憲：メタヒューリスティクスの世界，電子情報通信学会誌，Vol.86, No.6, pp.442–445 (2003).
15) Najork, M. and Wiener, J.L.: Breadth-First Crawling Yields High-Quality Pages, *Proc. 10th International World Wide Web Conference*, Hong Kong, pp.114–118, Elsevier Science (2001).
16) Pennock, D.M., Flake, G.W., Lawrence, S., Glover, E.J. and Giles, C.L.: Winners Don't Take All: Characterizing the Competition for Links on the Web, *Proc. National Academy of Sciences*, USA, Vol.99, No.8, pp.5207–5211 (2002).

(平成 16 年 3 月 24 日受付)
(平成 16 年 11 月 1 日採録)

松久保　潤

昭和 51 年生．平成 9 年北九州高等工業専門学校電子制御工学科卒業．平成 11 年同校専攻科制御専攻卒業．平成 13 年北陸先端科学技術大学院大学知識科学専攻前期課程修了．現在，同大学院後期課程在学中．電子情報通信学会学生会員．

林　　幸雄（正会員）

昭和 37 年生．昭和 60 年豊橋技術科学大学工学部電気電子工学科卒業．昭和 62 年同大学院修士課程修了．富士ゼロックス（株），（株）ATR 視聴覚機構研究所，（株）ATR 人間情報通信研究所等の研究員を経て，現在，北陸先端科学技術大学院大学知識科学研究科助教授．工学博士．数理工学，情報幾何学，力学系，Web 生態学，ニューラルネットの研究に従事．日本応用数理学会，電子情報通信学会，日本神経回路学会，INNS 等各会員．