動的ページの増加がもたらすWebリンク構造の変化

斉藤茂樹* 森口一郎**

いくつかのWebサイト群に対してクローリングを行い、リンク構造を解析した。静的ページではリンクの確率密度分布が過去の研究同様べき乗則に従っていることと、inリンクの緊密度が動的ページに比べ高いことがわかった。一方動的ページでは、リンク確率密度分布がべき乗則に従っておらず、inのクラスター係数が著しく低かった。また、リンク数が多くなるにつれ、特にinのクラスター係数が高くなることを確認した。これによって、inのリンク数が多いページ同士は緊密なリンク関係にあることがわかった。この事実から、静的、動的ページかによってクロール手法を変えることによって効率的なクロール手法の開発につながると思われる。

キーワード：クローリング、べき乗則、WWW、クラスター係数

Effects of increasing dynamic pages on the Web link structure

Shigeki SAITO and Ichirou MORIGUCHI

We performed crawling for several groups of Web sites and analyzed link structures. The probability density distribution of in-link showed the power-law which had been found in previous researches, and the closeness of in-link was found to be low in comparison with that of dynamic pages. On the other hand, in the dynamic pages, probability density distribution of links did not obey the power-law, and a clustering coefficient of in-link was remarkably low. Furthermore, we confirmed that particularly clustering coefficients of in-link increase as the number of the links increases. Therefore, it was found that the Web pages with many in-links have close relations mutually. These show a possible new crawling strategy that changes crawling methods by checking whether an Web page is static or dynamic.

Keyword：crawling, power-law, WWW, clustering coefficient

*東京情報大学総合情報学部 情報システム学科
Tokyo University of Information Sciences, Faculty of Informatics, Department of Information Systems
**東京情報大学総合情報学部 情報システム学科
Tokyo University of Information Sciences, Faculty of Informatics, Department of Information Systems

2009年12月7日受理

1．はじめに

今日情報を得る媒体としてWorld Wide Web（以降Webと略す）は重要な位置を占めるようになっている。このWebを使って情報を得るためには、URL（Uniform Resource Locator）をあらかじめ知っている必要がある。もし得たい情報を含むWebページのURLが不明であれば、検索エンジンを用いてURLを調べることが普通である。この検索エンジンはあらかじめWebを収集（クローリング）しておき、URLをインデックス化しているため、キーワードを使ってWebページを検索することができる。しかしWeb上のページ数は莫大であるため、効率的にクローリングを行っておく必要がある。また、目的のWebページを検索する際には、良質なWebページを優先的に検索結果として表示できなければ使いやすい検索エンジンとは言えない。Webページ同士はハイパーリンクで結びつけられているため、リンク構造を知っておく事は、効率のよいクローリング方法、重要なページを決定するアルゴリズムの開発に役立つと考えられる。

1999年、AlbertらがWebのリンク構造について研究し、初めてリンク数の確率密度分布（リンクの次数分布）がべき乗則に従うことを明らかにした［1］。またこれに続く他のWebリンク研究でも同様にべき乗則が現れると報告されており、原因としては、優先的選択が機能していると考えられている［2］。優先的選択とは、リンクを多く集めている人気のあるページをリンク先としてより高確率で選びやすくなる働きである。この研究結果に基づいて、リンク数の多いページを優先的に探索する効率的な手法も提案されている［3］［4］。しかし、これらの研究が行われた1999年～2002年頃はWebページ作成者が手動でコンテンツを作っている静的ページが多かったが、ここ数年はプログラムによってページを出力する動的ページが急増している。静的ページでは、作成者がそれぞれのコンテンツに対してあらかじめリンク先を選択しているが、動的ページはリンク先をプログラムが自動生成しているため、静的、動的ページとではリンク構造が異なると予想できる。

もしリンク構造が動的ページの増加によってここ数年で変わっているのであれば、参考文献［3］［4］のような情報探索手法が有効でなくなっている可能性がある。そのため、動的ページのリンク構造が静的ページと異なるのであれば、動的ページに対する情報探索手法はどのようにすべきか、対応策を検討するためのデータとする必要がある。また、静的ページのリンク構造が従来と同じであるか再確認も行った。さらに、リンク構造の調査にあたり、次数分布だけではWebページ間の結びつきは判断できないため、クラスター係数を用いた調査も行った。

本研究では、東京情報大学内の全Webページをクローリングし、URLから静的、動的ページを分類し、リンクの次数分布とクラスター係数を求めた。また、学外の数箇所を起点としてクローリングを行い、東京情報大学内のWebに特有の特徴があるかどうかのチェックも行った。その結果、静的ページの次数分布はべき乗則に従うことと、リンク数が多いページ間の結びつきが強いことを明らかにした。

2．方法

Webページにアクセスするためには、あらかじめURLを知っておかなければならない。未知のWebページを発見するには、既知のページにアクセスし、そのリンク先のURLを取得すればよい。つまり、あらかじめ起点となるWebページを決め、そのリンク先のURLを取得する。このようにして取得したURLのWebページに対しても同様にリンク先を取得する操作を繰り返すことで既知のWebページを増やしていく。この一連の操作をWebクローリングと呼ぶ。また、発見したURLは全てインデックス化しておくので、図１のようにあるホップ目のページから前のホップに対してリンクが

あるような場合でも、前のホップのページはすでに発見済みと判断でき、リンク情報は収集するが実際に2度アクセスする必要はない。同様に、もし②のページから⑤のページへのクローリングが先に行われ、③から⑤へのクローリングで再び⑤を発見した場合でも、⑤に2度アクセスする必要はない。

一般的にWebは情報を公開するために作られ、ページ作成者のトップページ等他のページから辿ることができる。一方、通常のクローリング手法ではoutリンクしかもたずinリンクを持たないページや、他のページ群からinリンクを持たないページ群（隠しページ群）はクローリングできない。これら隠しページはそのURLを知っている少数の者しかアクセスできない共有メモのようなものなので、リンク範囲が隠しページ間に限定されている。東京情報大学内にもそのような隠しページ群は存在すると思われるが、Webページ全体に比べ、そのページ数は非常に少ないと考えられる。さらに本研究ではWebページ全体の結合構造に着目しているので、このようなWeb全体から分離している隠しページは解析の対象外とした。

2.1 Webクローリングの方法

本研究でWebクローリングを行う際、以下の条件を設定した。

(1) Webページ以外のファイルの除外

Webページは画像、音声、実行ファイルなどWebページではないファイルに対してもリンクする事ができる。しかし本研究ではWebページ同士のリンクのつながりについて調査したいため、リンク先を抜き出す操作時に、これらリンク先を持ち得ないものについてはクローリング対象から除外した。

(2) エラーページの除外

Webサーバが見つからなかった場合とアクセスした際のHTTP応答コード400系列、500系列のエラーを返したWebページからはリンク先を検出できない。そのため、これらのエラーページについてもクローリング対象から除外した。

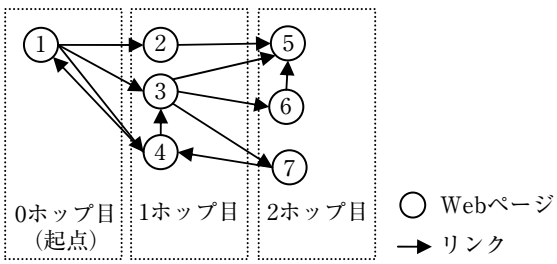

図1．Webクローリングの手順。Webページの数字はアクセスする順序を表す。

2.2 ページの分類方法

本研究では主に、Webページを静的ページと動的ページに分類している。静的ページとは、拡張子が.htmlとなるようなファイルであり、作成者がコンテンツを書いた後、ページ内容は変化しない。これに対して動的ページとは、.cgiや.phpのようにサーバ側で実行されたプログラムの結果によって変化するページである。本研究では、拡張子が.html、.htmのものを静的ページとし、.php、.pl、.cgiのような拡張子のものを動的ページとした。さらに、URL中に「?a=b」のようなクエリーストリングが付加されたものも、パラメータとして受け取り処理を行うため、動的ページとして扱う。また、分類はURL中の拡張子やクエリーストリングの有無で判定するので、たとえば、SSI（Server Side Includes）が使われ実質動的ページであっても、拡張子が.htmlであれば静的ページとして扱った。また拡張子が.htmlであってもクエリーストリングがあれば動的ページとした。

クエリーストリングはパラメータであり、これによってページ内容が変化することが多い。また、クエリーストリングがURL中に埋め込まれているため、この内容によってURLも異なるという特徴がある。

2.3 ページの分類毎の解析方法

WebページのURLから静的、動的ページを

分類し、静的ページ同士、動的ページ同士のリンクから、リンク生成メカニズムの違いについて調査を行った。

たとえば静的ページ同士のリンクについて調査する場合、クローリング結果から静的ページが持っている動的ページへのリンクは破棄し、静的ページ間のリンクのみを抽出した。同様に、動的ページ間のリンクを抽出した場合とで比較を行った。

3．解析

本研究ではWebページの持つリンク数（次数）の確率密度分布（次数分布）と、Webページ同士のリンクの緊密度を表すクラスター係数を求めた。

Webページの持つリンク数を k とし k 本のリンクを持つWebページが存在する割合をリンク数の確率密度分布$p(k)$とする。この$p(k)$は通常「次数分布」と呼ばれ、これがべき乗則、即ち$p(k) \propto k^{-\beta}$、に従えば過去の研究と一致し、リンク先の選び方に優先的選択が働いていると考えられる。

また、Webのリンクには方向性があるため、次数分布、クラスター係数それぞれをinリンク、outリンクに分けて求めた。ここで各Webページに対して他のWebページから入ってくるリンクをinリンクと呼び、各Webページから他のWebページへ向けて出て行くリンクをoutリンクと呼ぶ。（図２）

3.1　累積次数分布

次数分布がべき乗則に従う場合、次数分布を両対数プロットすると直線のグラフとなる。この直線の傾きからべき指数－βを求めることができる。しかし実データやシミュレーションでは、大きい k でたびたび$p(k)=0$となる箇所があるため、傾きを求めることが困難となる。そこで次式のように累積次数分布に変換し、べき指数を求めた。

$$p_{cum}(k) \equiv \sum_{k'=k}^{\infty} p(k') \tag{1}$$

ただし、累積次数分布を両対数プロットした場合の直線の傾きは、次数分布の傾きから１ずれるため、求めた傾きから１を引いておく［5］。

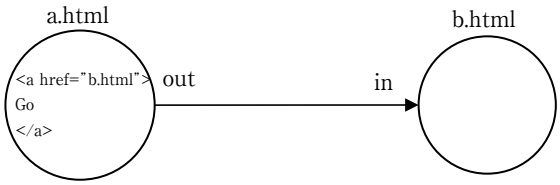

図2．a.htmlのoutリンクとb.htmlのinリンクの関係。

3.2　クラスター係数

クラスター係数とは、ノードに隣接しているノード同士が隣接している割合を示す指標であり、ノード毎に求める。ここで本研究でのノードとは静的、動的ページを指す。Webページのリンクには方向性があるが、方向性を考慮しない場合は以下の式で求められる。

$$C_i = \frac{1}{(k_i)(k_i-1)/2} \sum_{j,k} a_{ij} a_{ik} a_{jk} \tag{2}$$

ここで、C_iはノード i のクラスター係数を、k_i はノード i のリンク数を、a は隣接行列を表す。

しかしWebのリンクにはinとoutの方向性があるので、ここでは方向性を考慮したクラスター係数を下記の式で求めた［6］。

$$C_i^{in} = \frac{1}{(k_i^{in})(k_i^{in}-1)/2} \sum_{j,k} a_{ji} a_{ki} \frac{(a_{jk}+a_{kj})}{2} \tag{3}$$

$$C_i^{out} = \frac{1}{(k_i^{out})(k_i^{out}-1)/2} \sum_{j,k} a_{ij} a_{ik} \frac{(a_{jk}+a_{kj})}{2} \tag{4}$$

また全ノードのクラスター係数を平均したものをネットワークのクラスター係数といい、ネットワークにあるリンクの緊密度を表す。本研究でのネットワークのクラスター係数は、静的、動的ページ間同士のリンクの緊密度のことを指す。

方向性のあるネットワークのinのクラスター係数を例にした計算方法は、図３のネットワーク構造に隣接行列を用いると表１のようになる。たとえばノード１はノード２，３，４からのinリンクがある。このうちから２つを選ぶ組み合わせは、ノード２，３と２，４と３，４の３通りあり、$a_{23}+a_{32}=0$、$a_{24}+a_{42}=2$、$a_{34}+a_{43}=1$となることから、ノード１のinのクラスター係数を求めると1.5/3=0.5となる（表２）。

3.3 東京情報大学内Web

http：//www.tuis.ac.jp/を起点ページとし、24ホップにわたってクローリングを行った。ただしtuis.ac.jpドメイン内を全クロールさせるため、ドメイン外Webへのクロールは行わなかった。発見されたURLは175,028にのぼるが、このうち閲覧権限がない場合や、ページが消滅している等の理由でアクセスできなかったものを除くと、実際にWebページを取得できたのは161,842ページであった。ホップごとの発見URL数をプロットした図４で、16ホップ目以降の発見URL数はほぼ一定となっている。これはいくつかの動的ページではアクセス毎に異なる文字列を生成し、ユニークなURLを次々とリンク先として生成するためである（図５）。したがって、16ホップ程度で通常のページは十分にクローリングを終えていると考えられる。クロールするホップ数が増すにつれ、このような動的ページの数は莫大なものとなるので、全クロールするまでに要する時間やマシン性能を考慮し、既知のものについては可能な限り排除した。しかし全てを排除することはできず、このような動的ページはおよそ67,000ページ残っている。

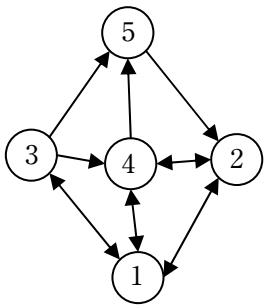

図3．方向性のあるネットワーク例

表1．図3に対応した隣接行列

	1	2	3	4	5
1	0	1	1	1	0
2	1	0	0	1	0
3	1	0	0	1	1
4	1	1	0	0	1
5	0	1	0	0	0

表2．図3のinのクラスター係数

ノード番号	inのクラスター係数
1	1.5/3
2	1.5/3
3	0
4	2.0/3
5	0.5/1
ネットワークのinのクラスター係数：0.4333	

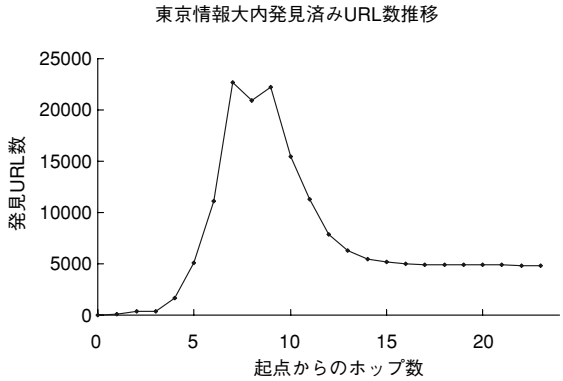

図4．東京情報大学内Webの発見URL数の推移．図5のような動的ページがあるため、常に発見されるURLがある。

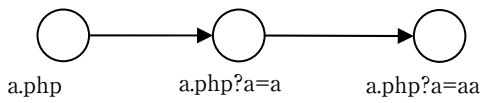

図5．アクセスする度次々にユニークなURLを生成する動的ページ例。

ちなみに、静的ページで最大のinリンク数は48,342（リンク元の1ページから2つ以上のリンクがありえるため、リンク元のページ数は3,427ページ）であった。学内には、学生が授業で参照できるようJavadoc（Javaのリファレンスページ）が存在する。Javadocにはクラス毎に仕様をまとめたページがあり、継承関係等があれば、クラス毎の仕様ページにリンクされている。最大のinリンク数があるページは、Objectクラスの仕様ページである。Objectクラスは全クラスの基底クラスなので、全クラスの仕様ページからリンクを受けている。（図6-a、図6-b）（http：//www.solar-system.tuis.ac.jp/Java/jdk-1_5_0-doc-ja/api/java/lang/Object.html）

また最大のoutリンク数は19,120（リンク先の1ページへ2つ以上のリンクがありえるため、リンク先のページ数は2,181ページ）であった。これもJavadocのページである。このページは、Stringクラスを使用しているクラス、メソッドの一覧ページである。たとえばURLクラスにはtoString（）メソッドがあるが、これはStringクラスを使用している。そのためこの一覧ページにも、URLクラスの仕様ページへリンクされている。Stringクラスは便利なクラスであり、あらゆるクラスで使用されているため、この一覧ページには多くのクラス仕様ページへのリンクがある。（図7）（http://www.solar-system.tuis.ac.jp/Java/jdk-1_5_0-doc-ja/api/java/lang/class-use/String.html）

図8のように静的ページのみを抽出した結果、累積次数分布はスケールフリー性の特徴の一つであるべき乗則に従っていることがわかった［6］。次数分布のべき指数は、この傾きの－1.34と－1.15から1を引けばよいので、outで－2.34、inで－2.15という値になり、out：－2.45、in：－2.1という過去の研究とほぼ一致した［1］［7］。inの傾きがoutの傾きに比べ緩やかになっているが、これは元々inリンクの多いWebページはその他の多くのページからさらにリンクされる可能性があり、そのリンクされる数に上限がないのに対して、他へ極端に多くのリンクを持つページは考えづらいことから、inリンク数がoutリンク数よりも大きくなりやすいことが分かる。この結果から、静的ページのinのリン

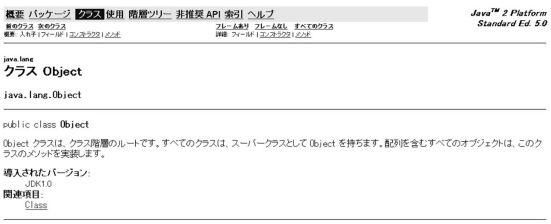

図6-a　Objectクラスの仕様ページ。Objectクラスを継承したクラスの仕様ページからリンクされ48,342のinリンクがある。

図6-b　Objectクラスの仕様ページにリンクしている一例。

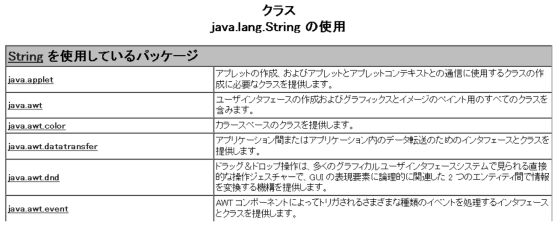

図7　Stringを使用しているクラスの仕様ページを一覧できるページ。

表3．東京情報大内Webクローリング結果。URLからページを分類する際、静的、動的ページのどちらとも判定できないURLが4,265あった。「全て」にはこれらも含まれている。

	ページ数	リンク数	クラスター係数(in/out)
全て	161,842	8,551,676	0.0864/0.4143
静的のみ	49,227	2,268,552	0.1686/0.3904
動的のみ	108,350	4,121,860	0.0462/0.4391

ク次数分布にべき乗則が現れる理由として、現在も優先的選択が機能していると考えられる[2]。outリンクでもべき乗則が成り立っているが、どのようなリンクのつながり方の規則から結果としてリンク次数分布のべき乗則に結びつくのか、いまだに明らかではない。

一方、動的ページでは図9に見られるように直線領域がほとんどなく、べき乗則に従っているとは言えない。すなわち、動的ページを生成するプログラムがoutリンクを自動生成していると考えられるため、優先的選択によってリンク先が選ばれていないことがわかる。

また、表3のクラスター係数をみると、outに比べinのクラスター係数が低い傾向にあるが、静的ページのみ抽出した場合では、動的ページのみ抽出した場合に比べてinのクラスター係数が高い。これは、静的ページは動的ページに比べると図10-aのようなリンク元のページ同士に関連があるということを示している。たとえば、ページのリンク元が同一トピックを扱うページならば、リンク元のページ同士がリンク関係にある確率が高いことが多いと考えられる。動的ページのみ抽出した場合ではoutに比べinのクラスター係数が著しく低く、inリンクの緊密度が低かった。これは、図10-aのようなリンク元のページ間に関連がないか、リンク元のWebページが1ページのみということを示している。この傾向は東京情報大学外の起点からクロールした場合でも見られ、動的ページのみ抽出した場合、表4のようにoutのクラスター係数に対してinのクラスター係数が著しく低い。これは、図5のような動的ページが多数あるためと考えられる。たとえば、図11の動的ページa.php、a.php?s=abc、a.php?s=xyzの3つは、それぞれクエリーストリングで与えられたパラメータに基づいて、新たにリンク先にユニークなURLを生成している。ただし図11ではa.php?s=xyzの生成するリンク先は省略した。図5のようなクエリーストリングを与えられる動的ページのリンク先は、既存のページをリンク先としているものと、クエリーストリングを付加した新たなURLのように、表示する際に

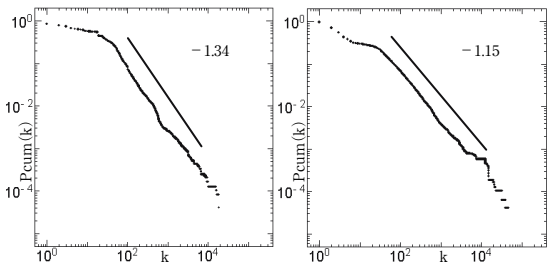

図8．静的ページ間のリンクを対象にした累積次数分布。
左：outリンク，右：inリンク

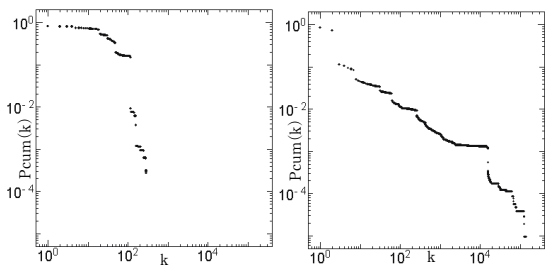

図9．動的ページ間のリンクを対象にした累積次数分布。
左：outリンク，右：inリンク

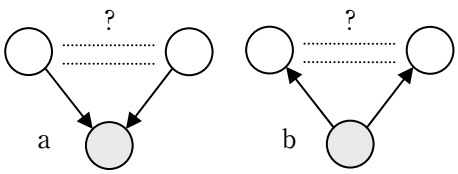

図10．網掛けページからみたクラスター係数の求め方。（a：inのクラスター係数、b：out のクラスター係数。）

表4．http://www.ipa.go.jp/を起点にクロールした際の動的ページ結果。このうち、およそ92,000ページが図5のように生成されたページだった。

	ページ数	リンク数	クラスター係数(in/out)
動的のみ	96,914	857,305	0.0081/0.7426

生成されるものがある。既存ページへのリンク先は動的ページを生成するプログラムを書いた作成者が選んでいるため、outリンクの緊密度は高めになると考えられる。これに対してinリンクでは、クエリーストリングの部分が直前のページで生成されているためリンク数は１～２と非常に少なく、リンク数が１であれば自動的にクラスター係数は０となってしまう。表４のデータではクエリーストリングを付加されたページはおよそ92,000ページあるが、ほぼ全てinのクラスター係数が０であった。一方outのクラスター係数は、図11のようにクエリーストリングを付加したURLのみにリンクしているわけではないため、inのクラスター係数と同じ理由で低くなることはない。

次にクラスター係数のリンク数依存性について解析した。図12、図13は表３で示した東京情報大学内Webのデータ中から、リンク数が８以上、32以上、128以上、512以上、2048以上のWebページを抽出したデータ５つを用意しクラスター係数を求めたものである。静的、動的ページどちらも、抜き出すリンク数 k の値が大きくなるとクラスター係数が高くなることがわかる。特にinのクラスター係数ではその傾向が強い。inのクラスター係数は、リンク元のページ同士に関連があれば高くなるが、図10-aのようなリンク元のページがリンクの少ないページならば、互いに関連し合っている可能性は低い。このようなリンクの少ないページが排除されたことで、inのクラスター係数が高くなる傾向にあると考えられ、図11のような動的ページが排除された場合により顕著になる。またoutのクラスター係数は、inに比べて緩やかに推移している。もし優先選択が機能しているならば、リンク先はリンク数の多いページが選ばれやすい。図10-bのようなリンク先のページがリンクの多いページならば、同じくリンクの多いページ同士に関連がある可能性も比較的高い。このことからリンク数の少ない大多数のページも、outのクラスター係数はinに比べ高いと考えられる。

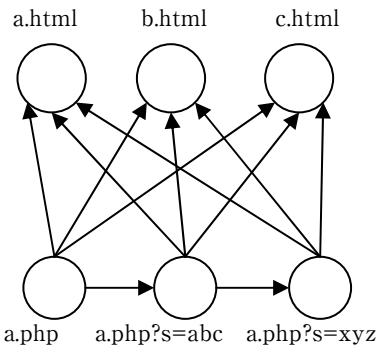

図11．図５のような動的ページ（a.php、a.php?s=abc、a.php?s=xyz）のinのクラスター係数が低くなる例.

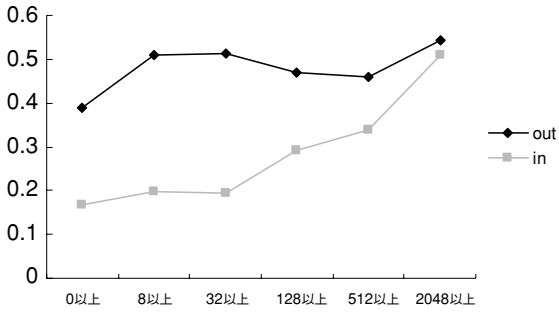

図12．東京情報大学内でリンク数k以上のWebページを抽出後の静的ページ間のクラスター係数。0以上は、表3の「静的ページのみ」と同じ。

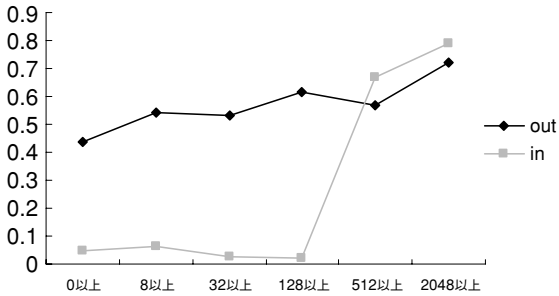

図13．東京情報大学内でリンク数k以上のWebページを抽出後の動的ページ間のクラスター係数。0以上は、表3の「動的ページのみ」と同じ。

4．まとめ

本研究では、東京情報大学内Webページのほぼ全てと、学外のいくつかのWebサイトに対してクローリングを行い、リンク構造を解析した。静的ページは、リンクの次数分布が以前のようにべき乗則に従っていること、動的ページに比べinのクラスター係数が高いことを確認した。一方動的ページでは、次数分布がべき乗則に従っておらず、inのクラスター係数が著しく低かった。また、リンクの多いページを抽出し、inのクラスター係数が高くなる傾向を確認した。これはinのリンク数が多いページ同士のむすびつきが強いことを示している。これにより、静的ページでは以前と同じリンク構造であるため、既存手法［3］［4］のような手法を用いることができると考えられる。しかし、動的ページでは傾向が異なるため、新たなクロール手法が必要であり、静的、動的ページとでクロール手法を分けて行う必要がある。たとえば、リンク数が多いページ同士は緊密なリンク関係にあるという点を利用する手法などが考えられる。

最初のWebリンクのべき乗則発見以来、inリンクに関しては優先的選択がその有力な原因であると考えられている。しかし、今回の研究でも確認された静的ページのoutリンクのべき乗則出現理由は現在でも明らかになっていない。

本研究をさらに進め、Webリンク構造、クラスター係数による緊密度を解析することによって、同一のトピックを扱うWebページ群を効率的に収集するクロール手法の開発にもむすびつくと考えられる。

【参考文献】

［1］R. Albert, H. Jeong, and A. L. Barabási : Diameter of the World-Wide Web, Nature, vol.401, pp.130-131（1999）
［2］A. L. Barabási, and R. Albert : Emergence of Scaling in Random Networks, Science, vol.286, pp.509-512（1999）
［3］Kim et. al.：Path finding strategies in scale-free networks：Phys. Rev. E65, 027103（2002）
［4］Adamic et. al.： Search in power-law networks： Phs. Rev. E64, 046135.（2001）
［5］S. N. Dorogovtsev, Jose F. F. Mendes : Evolution of Networks : From Biological Nets to the Internet and WWW, pp.222-223（Oxford, 2003）
［6］G. Caldarelli : Scale-Free Networks : complex webs in nature and technology（Oxford, 2007）
［7］S. N. Dorogovtsev, Jose F. F. Mendes : Evolution of Networks : From Biological Nets to the Internet and WWW, pp.80-81（Oxford, 2003）