Blog における話題分析のためのランダムサンプリング手法の提案

仲前　晋太郎† 　成　凱‡

†‡九州産業大学大学院 情報科学研究科 〒813-8503 福岡市東区松香台 2-3-1

E-mail: †k09gjk08@ip.kyusan-u.ac.jp, ‡chengk@is.kyusan-u.ac.jp

あらまし　近年、Web 技術の進化に伴い、Blog に代表される、一般利用者より自由に発信された情報が著しく増えており、個人の関心の持ったニュースや出来事について、速報性、リアルタイム性のある新鮮な情報が発信されている。このような情報を分析することによりチャンス発見につながると期待されている。分析するにあたって、調査対象となる Blog 記事を無作為に抽出すること、いわゆる、ランダムサンプリングが重要である。本研究では、ランダムウォークを用いたランダムサンプリング手法として、リンク構造を対象としたサンプリング手法 RRW (Random Walk with a Reservoir)とページを対象としたに対するサンプリング手法 IRW (In-Degree Weighted Random Walk)を提案し、評価実験を行った。実験結果によって、平均クラスター係数、入次数、出次数、強連結成分の数、弱連結成分の数の各評価尺度においては、IRW はページを対象としたサンプリングに優れている。RRW は単純ランダムウォークより優れていることが分かった。

キーワード　ランダムサンプリング、グラフデータ、ランダムウォーク、ブログ、話題分析

Random Sampling Methods for Topic Analysis of Blogs

Shintaro NAKAMAE† 　Kai CHENG‡

†‡Graduate School of Information Science, Kyushu Sangyo University

3-1, Matukadai 2-chome, Higashi-ku, Fukuoka, 813-8503

E-mail: †k09gjk08@ip.kyusan-u.ac.jp, ‡chengk@is.kyusan-u.ac.jp

1. はじめに

Web 技術の進化に伴い、一般利用者より自由に発信された情報が著しく増えており、CGM (Customer Generated Media) として注目が高まっている。とりわけ Blog に代表される CGM では個人の関心の持ったニュースや出来事について、速報性、リアルタイム性のある新鮮な情報が発信されている。これらの情報を元に様々な分析を行うことにより、チャンス発見につながる可能性がある。例えば、異なる言語圏においてある話題に対する関心度は､その話題を言及する ページの割合等を分析することにより評価できる。

より精度の高い調査結果を得るために、理想的には調査対象となる Blog 記事を漏れなく収集することが必要と思われる。しかし、大規模な収集はコストが高いだけではなく、すべてのサイトを一度カバーするまでに必要な時間が長く、最新な情報を分析できない問題点がある。例えば、検索エンジンのクローラは同じサイトを再度訪問するまでに一ヶ月もかかることがある。それゆえ、調査対象から代表的な部分を抽出する方法、いわゆるランダムサンプリングが必要不可欠である。

Web を対象とするランダムサンプリングの主要な手法として、IP アドレスによるランダムサンプリング、検索エンジンを用いたランダムサンプリング、ランダムウォークによるサンプリングが知られている。IP アドレスによるランダムサンプリングでは、IP アドレスとなる 32 桁の二進数を無作為に生成し、その IP アドレスをもつホストはインターネット上で Web サーバとして生きている場合には、そのサーバから Web ページを抽出する。しかし自由に選ばれた IP アドレスはウェブサーバに割り当てた部分が非常に少なく、Web サーバ全体において IP アドレスの分布も均一とはいえなく、偏りが生じる可能性がある。一方、検索エンジンによるサンプリングでは、ランダムに生成された検索条件で検索し返された結果からランダムにページを選ぶことで、必要なサンプルページを収集する。検索エンジンのデータベースにインデックスされたページの範囲や更新頻度等の影響を受ける短所がある。

ランダムウォークによるサンプリングでは、Web をページ（ノード）とリンク（エッジ）からなるグラフと捉え、ランダムウォークをしながら、ページやリン

クを収集している。この方法は、ウェブのハイパーテキスト特徴を利用して、ページだけではなく、代表的なリンク構造も抽出することができるので、注目が高まっている。例えば、Henzinger ら[1]が提案した PageRank に基づくランダムウォーク、Leskovec ら[5]の ForestFire 法等が上げられる。

しかし、これらの手法は以下の問題点が指摘されている。まず、ページ対象の調査と、リンク構造を対象としての調査の違いを考慮に入れていない。そのため、リンクの多いページに偏った結果しか得られない。次に、調査対象である Web の広域分散性を重視せず、サンプルを抽出していくうちに分かってくることをサンプリングのプロセスに反映されていない。

本研究では、ページを対象としたサンプリングのため、ランダムウォークにおける移動先の入次数を考慮した手法 IRW を提案し、グラフの各ノードを偏りなくサンプリングできるようにする。そして、リンク構造を対象としたサンプリングのため、Reservoir を用いたサンプリング手法 RRW を提案する。RRW を用いれば、サンプリングを進めていくとともに、収集されたサンプルの質がよくなっていく。

2. ランダムウォークによるサンプリング

Web は有向グラフとみなして、ページをノード、リンクをエッジとするとして、グラフ上でランダムウォークによって、ページやリンクを無作為に抽出することができる。

ランダムウォークとは、グラフ上で次の移動先が確率的に無作為（ランダム）に決定される運動である。ランダムウォークは、マルコフ性を持っており、現在の状態から次の状態への推移は、それ以前の状態に依存しない。

Web をグラフとみなし、あるページから出ている複数のハイパーリンクを無作為に一つ選んで辿っていく。全てのリンクに対して、偏りなく同じ遷移確率で選択することが一般的である。例えば、ページ v_i において、ページ v_j への遷移確率 $p_i(j)$は ページ v_i の出次数（v_i を始点となるリンクの数）を $\mathrm{outdeg}(v_i)$とすると

$$p_i(j) = \frac{1}{\mathrm{outdeg}(v_i)}$$

人気のあるページは他のページから多くリンクされており、人気のないページは他のページからのリンク数が少ない。これにより、被リンク数の多いページは収集されやすく、被リンク数の少ないページは収集されにくいといった問題がある。これにより、通常のランダムウォークではリンクを無作為に辿ったとしても収集されるデータは人気のあるページを多く収集してしまうことになり、偏りが生じてしまう。

2.1. ページを対象としたサンプリング手法 IRW

入次数の高いページは複数の経路よりランダムウォークで辿りつく可能性があるので、遷移確率を低くすることにより、ページが選ばれたチャンスを均等にすることができる。これに基づいて、アルゴリズム **IRW（In-Degree Weighted Random Walk）**を考案する。IRW は入次数の多いページに対して遷移しにくくすると考えた手法である。

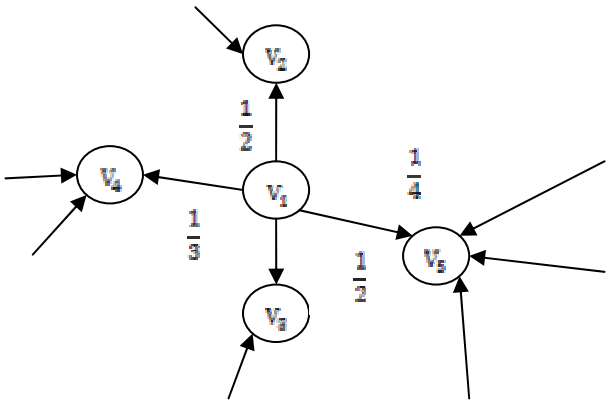

図 1. IRW:入次数が大きいほど遷移確率が小さい

図 1 の有向グラフにおいて、通常のランダムウォークではノード V_1のそれぞれのノードに対する遷移確率は、それぞれ等確率であったのに対して IRW では、隣接するノードの入次数が大きいほど、遷移確率を小さくしている。

$$p_i(j) \propto \frac{1}{\mathrm{indeg}(v_j)} \qquad (1)$$

式(1)で v_i から隣接ノード v_j への遷移確率は v_j の入次数の逆数と比例する。これにより、入次数の多いノードに対して低い確率で遷移させることができる。よって、入次数による偏りを減らすことができると考えられる。

IRW を実現するために、各ノードの入次数と出次数を知る必要がある。出次数は現在のページに埋め込まれたリンク数から分かるが、リンク先のページの入次数を調べる単純な方法が必要である。これについて以下の方法が考えられる。

まずは、Blog システムのトラックバック機能を利用して、記事の被リンク数を推定できる。次に、あるページから、隣接するページへ移動する確率を計算するために、隣接する全てのページの被リンク数を調べるが、コストがかかる。解決策として、隣接するページの数がある閾値 T を超えた場合、そこからランダムに最大 T ページを選び出す。この T 個のページに対して、公式（1）に基づいて、遷移確率を求める。**アルゴリ**

ズム 1 では入次数が全て知っているとする。

アルゴリズム 1：IRW

入力：

(1) G=(V, E)：抽出対象となる有向グラフ。ノードの入次数 *indeg(v)* と出次数 *outdeg(v)*

(2) *samp_size*：サンプルサイズ

(3) *jump_prob*：ジャンプ確率。ランダムウォークを中断し隣接ノード以外のノードを選ぶ確率

(4) *U*：ランダムウォーク始点の集合

出力:

抽出結果グラフ（サンプル）G'=(V', E')

処理手順：

1. 初期化：$u \in U$, *sample=1*, $v=u$, $V'=\{u\}$, $E'=\phi$;
2. *sample*≧*samp_size* ならば、G'=(V', E')を出力して終了
3. $m=outdeg(v)$, $neighbor(v)=\{v_1,v_2,\cdots,v_m\}$, $p_i=1/indeg(v_i)$, $p=p_1+p_2+\cdots+p_m$
4. 隣接点 $neighbor(v)=\{v_1,v_2,\cdots,v_m\}$から次の移動先 v_k をランダムに選ぶ：
   乱数 $r \in [0, p]$を振り、r は以下の範囲内であれば移動先 v_k をノード v_k とする
   $$\sum_{i=0}^{k-1} p_i < r \le \sum_{i=0}^{k} p_i, \quad (p_0 = 0)$$
5. $V'=V' \cup \{ v_k \}$, $E'=E' \cup \{ <v, v_k> \}$
   $v= v_k$, *sample= sample+1*
6. ランダムジャンプ
   (1) 乱数 $r_0 \in [0, 1]$を振る
   (2) r_0<*jump_prob* なら、別始点 $u' \in U$ を選び、$V'=V' \cup \{ u' \}$, $v= u'$, *sample= sample+1*
7. ステップ 2 へ移動し処理が続く

2.2. リンク構造を対象としたサンプリング手法 RRW

リンク構造を対象としたサンプリング手法として、**RRW（Random Walk with a Reservoir）**について述べる。IRW が Web のリンク構造を考慮せず隣接ノードの入次数によって遷移確率を求める手法に対し、RRW は入次数による偏りなどのリンク構造を考慮してサンプリングする手法である。Reservoir とは水槽を意味した言葉であり、水槽から流れ出る水をイメージしており、データが膨大であり、母集団の大きさが分からないことが前提で考えられている。

RRW の考え方としては、通常のランダムウォークにより、サンプルを N 個抽出する。データの収集を通常のランダムウォークを用いることにより Web のリンク構造における他から多くリンクされている人気のあるページは収集されやすい性質があるためページを N 個抽出する際に同じページを何度も抽出してしまう。同じページを抽出した場合は、抽出した回数を数えておく。この処理を N 個の異なるページが抽出されるまで繰り返す。当然、N 個のページのうち、抽出回数が 1 しかないページもあるし、2 以上のページなどがある。

アルゴリズム 2：RRW

入力：

(1) G=(V, E)：抽出対象となる有向グラフ

(2) *samp_size* ：サンプルサイズ

(3) *samp_space*：抽出対象サイズ

(4) *jump_prob*：ジャンプ確率。ランダムウォークを中断し隣接ノード以外のノードを選ぶ確率

(5) *U*：ランダムウォーク始点の集合

出力:

抽出結果（サンプル）グラフ G'=(V', E')

処理手順：

1. 初期化：$u \in U$, *sample=1*, *space=1*, *freq(u)=1*, $V'=\{u\}$, $E'=\phi$; $v=u$,
2. *space*≧*samp_space* なら, G'=(V', E')を出力して終了
3. 隣接ノード $neighbor(v)=\{v_1,v_2,\cdots,v_m\}$から移動先 *t* をランダムに選ぶ
   space= space+1
4. もし $t \notin V'$
   もし *sample*≧*samp_size*, 以下の (1) － (3) を *sample*<*samp_size* になるまで繰り返す
   (1) V'からランダム v'を選ぶ
   (2) freq(v')= freq(v')-1
   (3) freq(v')が 0 であれば、v'を V'から削除、v'に関連するエッジも E'から削除 *sample=sample-1*.
5. そして $V'=V' \cup \{ t \}$, $E'=E' \cup \{ <v, t> \}$, freq(*t*)=1, v= *t*, *sample= sample+1*,
6. もし $t \in V'$
   freq(t)= freq(t)+1, $E'=E' \cup \{<v, t>\}$,
7. ランダムジャンプ
   (1) 乱数 $r_0 \in [0, 1]$を振る
   (2) r_0<*jump_prob* なら、別の始点 $u' \in U$ を選び、$V'=V' \cup \{ u' \}$, $v= u'$, *sample= sample+1, space=space+1*
8. ステップ 2 へ移動し処理が続く

次に、N+1 個目のデータを抽出した時、元の N 個のデータからランダムに 1 つ削除するページを決める。この時、削除されるページの回数が 1 ならばそのページを削除し、回数が 1 より大きければ回数を 1 減らす。次に抽出したページが元の N 個のページと同一のページであるならば、同一ページの回数を増やし、抽出したページを削除する。この処理を繰り返すことにより、Web のリンク構造における他から多くリンクされている人気のあるページの有意性を保ち、さらに、無作為な標本抽出ができると考える。

アルゴリズム 2 はサンプルサイズ samp_size とサンプル空間上限 samp_space を入力パラメタとして RRW を

実現している。samp_size は実際に出力されるサンプルページの数になる。samp_space はより確実な結果を得るためにサンプル空間から最低この数のページを調べる必要がある。ただし、調べたページは乱数によって最終の結果に含まないことが多い。

3. 評価実験

提案手法を評価するために、通常のランダムワーク RW、IRW、RRW の結果を比較する評価実験を行った。評価尺度を用いる[5]。

- ccf: クラスター係数の分布状況。
- inDeg: 入次数の分布状況
- outDeg: 出次数の分布状況
- scc: 強連結成分の分布状況
- wcc: 弱連結部分の分布状況

クラスター係数とは、ノード v に隣接するノードがお互いに連結する度合いを評価する指標であり、$(t/{}_nC_2)$で表す。ここで、n は v の隣接ノード数、t はこれらのノード間に実際に持っているエッジ数である。次数別のノードの平均クラスター係数を用いて、グラフのコミュニティ性を評価できる。

同じデータに対してそれぞれのサンプリング手法で抽出したサンプルグラフの上記の指標を比較する。

3.1. 実験データ

本実験では Google ProgrammingContest2002 年[6]で提供されたデータセットを利用している。このデータセットの内容は**表 1** に示している。

表 1 実験用データセット

ノード数	875713
エッジ数	5105039
最大 WCC のノード数	855802 (0.977)
最大 WCC のエッジ数	5066842 (0.993)
最大 SCC のノード数	434818 (0.497)
最大 SCC のエッジ数	3419124 (0.670)
平均クラスター係数	0.6047

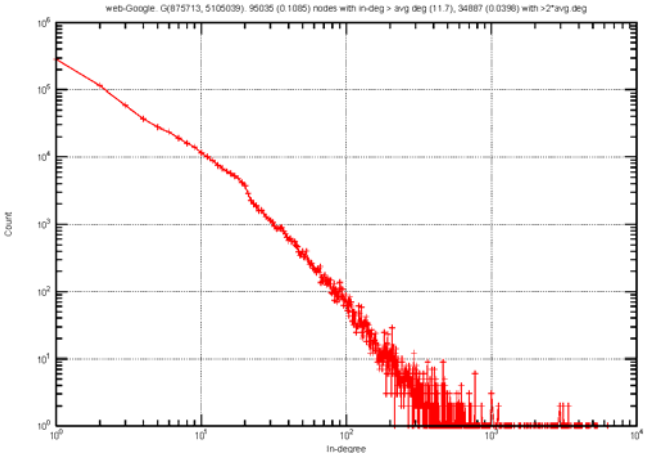

図 2. 元 Google グラフの入次数の分布

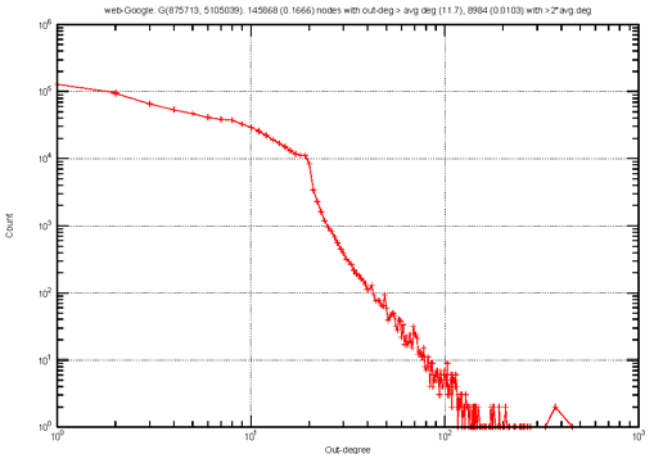

図 3. 元 Google グラフの出次数の分布

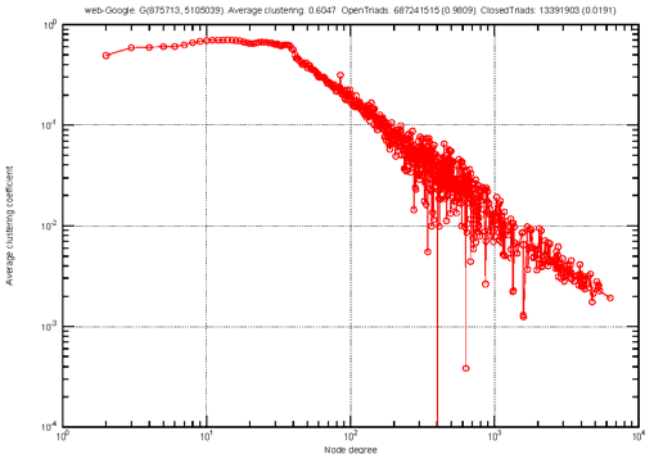

図 4. 元 Google グラフの平均クラスター係数の分布

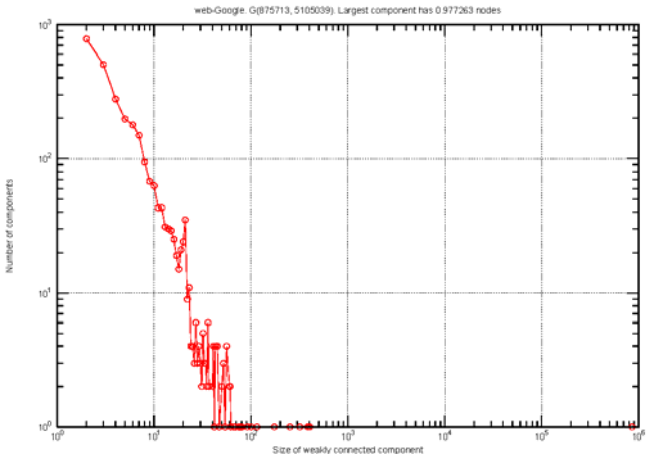

図 5. 元 Google グラフの弱連結成分の分布

図 2 から**図 6** までは、実験用データの特徴を現したものである。横軸と縦軸とも log でスケールダウンしている。

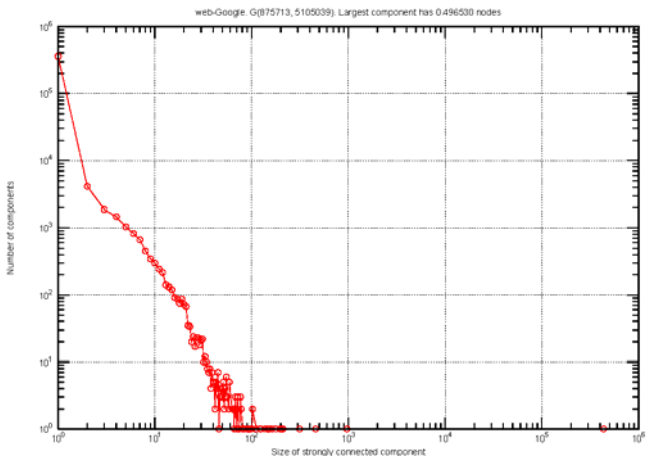

図 6. 元 Google グラフの強連結成分の分布

3.2. 実験結果

実験では各手法でサンプルを抽出し、それぞれのサンプルに対して、各グラフ特徴を評価した。サンプルを抽出する際に、以下のようにパラメーターを設定している。(1) jump_prob=0.15; (2) samp_size=ノード数の20%; (3) samp_space = samp_size*1.382

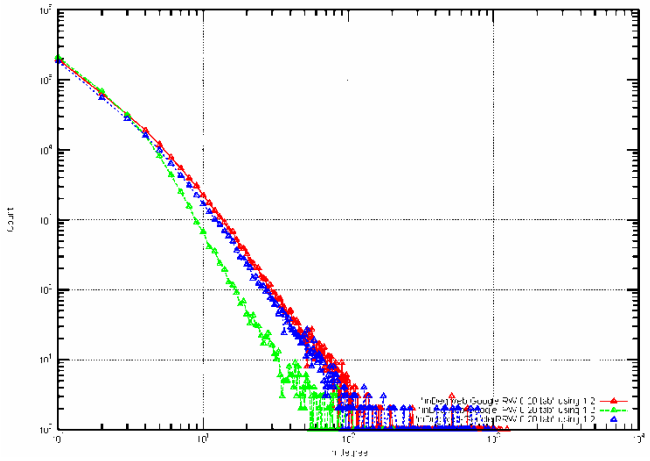

図 7. 抽出サンプルの入次数の分布

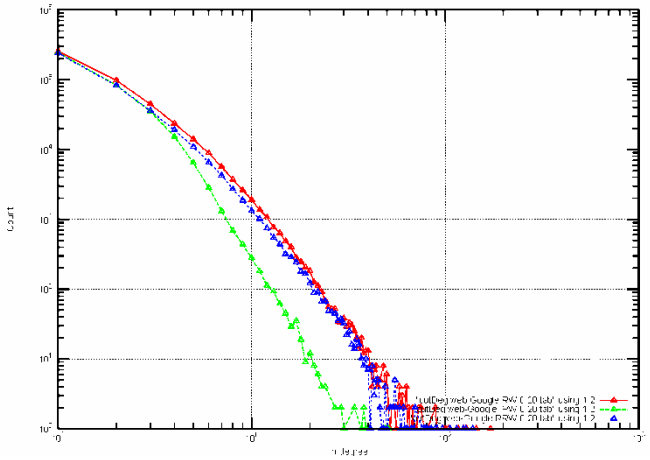

図 8. 抽出サンプルの出次数の分布

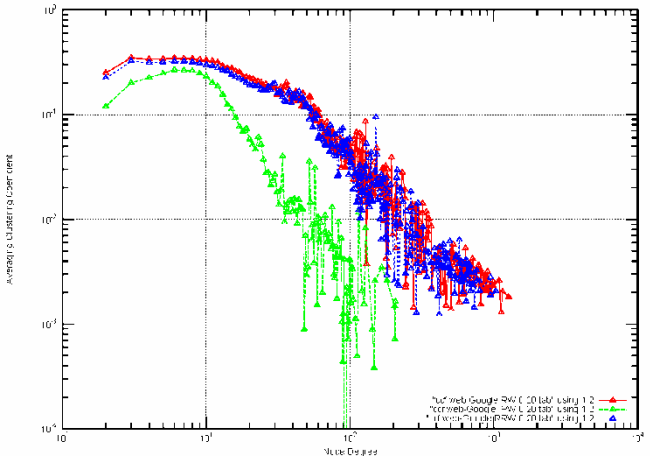

図 9. 抽出サンプルの平均クラスター係数の分布

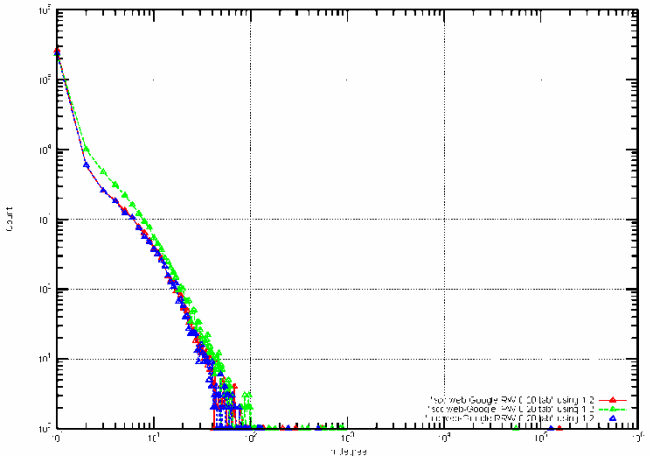

図 10. 抽出サンプルの弱連結成分の分布

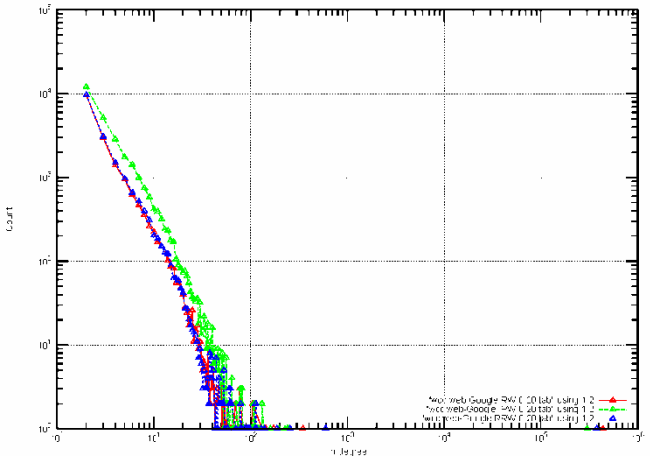

図 11. 抽出サンプルの強連結成分の分布

実験結果は**図 7**～**図 11**の各グラフで示している。グラフから以下の結果が分かった。解析結果として、IRWでは図2の各ノードの入次数の分布からRW、RRWより低いグラフとなっており、リンク構造の入次数による偏りに対して有効であることが分かった。さらに、RRWではリンク関係を表す指標として平均クラスター係数を用いて、比較した結果、RWと近似したグラフ

となりIRWより高い結果が得られたことはリンク構造を考慮したサンプリング手法として有効であることが分かった。

4. 関連研究

Henzigerら[1]はPageRankの逆数を用いてランダムウォークの遷移確率を求めた手法を提案した。PageRankはリンク構造のみでWebページ群から客観的な人気度が高いと思われるページを機械的に算出する尺度である。ランダムワークで人気度の高いページに辿り着く確率がほかのページより高いわけなので、ページに等確率にサンプリングするために、PageRankの大きなページに偏らない方法である。

正確なPageRankを求める計算コストが高いため、Henzigerらは、VR、PRの2つの方法を用いてPageRankを近似的に求める。VRでは、ランダムウォークを行う段階で各ページの訪問率VR (Visit Ratio)をPageRankの近似として求めている。PRでは、ランダムウォークによって得られた部分グラフで標準なPageRankの求め方で近似的PageRankを算出している。

しかし、PageRankは強連結グラフでしか求められないため、[1]の方式は適用できる場合は限られている。本研究で提案したIRWでは入次数のみで遷移確率を計算しているので､計算コストがかからないのみならず、弱連結有向グラフでも適用可能である。入次数はBlogにおいてTrackbackにより近似的に求めることができる。

他に、TrackBackに基づくBlogクローリング手法[10]がある。TrackBackは、関連するBlog記事同士をリンクするためのBlog固有の機能である。この手法では、TrackBackをBlog記事への支持投票とみなすことで、Blog記事へのスコアリングを行い、さらに、特徴語を指定することにより、着目した話題に限定したBlog記事を収集すると同時にTrackBackスパムの排除も可能にしている。そして、収集されたBlog記事集合内でのスコアが高いBlog記事を推薦している。Trackbackを利用する点では本研究と同じであるが、Trackbackを利用する目的が異なる。

5. 終わりに

本研究では、従来のランダムウォークによるサンプリング手法におけるページ対象の調査とリンク構造を対象としての調査の違いを考慮に入れていないという問題点に注目し、ページを対象としたサンプリングのため、ランダムウォークにおける移動先の入次数を考慮した手法 IRW とリンク構造を対象としたサンプリングのため､Reservoir を用いたサンプリング手法 RRW を提案した。IRW では入次数の多いノードに偏らないようにランダムウォークの遷移確率を使うことにより、ページを対象としたランダムサンプリングに相応しい。入次数情報が必要であるが、Blog の場合はトラックバックより得られるので、Blog サンプリングに適していると思われる。一方、RRW では、調査対象の規模が不明な場合でもリンク構造を適切にサンプリングできるので、Web のような膨大で動的であるネットワークのサンプリングに相応しいと思われる。

提案手法の有用性を検証するために、Google ProgrammingContest2002 年[6]で提供されたデータセットを利用し、RW、IRW、RRW の結果を比較する評価実験を行った。

今後の課題として、今回提案した手法の正当性を理論上で証明することや、更なる実験評価を行うことが上げられる。

参考文献

[1] M. Henzinger, A. Heydon, M. Mitzenmacher and M. Najork,On near-uniform URL sampling. In Proceedings of the 9th International World Wide Web Conference、2001, pp.295-308.

[2] K. Bharat and A. Broder, A technique for measuring the relative size and overlap of public Web search engines. In: Proc. of the 7th International World Wide Web Conference, Brisbane Australia, Elsevier, Amsterdam, 1998, pp. 379-388.

[3] S. Brin and L. Page, The anatomy of a large-scale hyper-textual Web search engine, in: Proc. of the 7th International World Wide Web Conference, Brisbane, Australia, Elsevier, Amsterdam, 1998 ,pp.107-117

[4] Vitter,J.S Random Sampling with a reservoir. ACM Trans. Math. Softw. 11(1) .1985 Mar.,pp.37-57

[5] J Leskovec, C Faloutsos Sampling from Large Graphs. Proceedings of the 12th ACM SIGKDD

[6] J. Leskovec, K. Lang, A. Dasgupta, M. Mahoney. Community Structure in Large Networks: Natural Cluster Sizes and the Absence of Large Well-Defined Clusters. arXiv.org:0810.1355, 2008.

[7] M. Henzinger, A. Heydon, M. Mitzenmacher and M. Najork, Measuring index quality using random walks on the Web. In Proceedings of the 8th International World Wide Web Conference(WWW8) 1999.pp 213-225

[8] Bar-Yossef,Z.:Berg,A.;Chin,S.; and Fakcharoeenphol, J. Approximating aggregate queries about web pages via random walks. In Proceedings of the 26th International Conference on Very Large Data Bases. 2000

[9] P.Rusmevichientong, DM.Pennock, S. Lawrence, C. Lee Giles; Methods for Sampling Pages Uniformly from the World Wide Web. In Proceedings of the AAAI Fall Symposium on Using Uncertainty Within Computation, 2001,Cape Cod, MA, pp.121-128

[10] 鎌田基之,福田直樹,石川博, ”TrackBack と特徴語に基づく Blog クローリングと Blog 記事の推薦”,DEWS2007 A8-4,2007.