Reservoir を用いた巨大グラフのランダムサンプリング

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、生命科学、ソーシャルネットワーク分析やマーケティング等の分野では、複雑な構造のもつグラフデータが急増し、グラフデータを扱うニーズが高まっている。膨大なグラフデータを解析するため、一部の代表的データを抽出するランダムサンプリング手法が必要不可欠である。我々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、ノードを対象としたサンプリング手法 IRW (In-Degree Weighted Random Walk)を提案してきた。しかし、IRW では入次数が分ることを前提としており、膨大なグラフデータに対して、入次数を事前に調べることが現実的ではない問題点が残っている。本研究では、Reservoir を用いて入次数を前提条件としないノードのランダムサンプリング手法 IRRW を提案する。評価実験によって、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の結果について報告する。

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

Sampling from Large Graphs with a Reservoir

Shintarou Nakamae† and 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

近年，生命科学，ビジネス分析やマーケティング等の分野では，複雑な構造のもつデータが急増し，グラフデータとして知られている。グラフデータとは、有限個の点からなる集合と 2 点を接続するいくつかの辺からなる集合によって定義される。定義自体は単純であり、点と辺との繋がりの様子を図に描くと視覚に訴える事ができるため、応用範囲はきわめて広い。例えば、地図上の街や点で、それらをつなぐ道路を辺で抽象化して示すと、一つのグラフを得る。鉄道や地下鉄の路線図、空港を結ぶ航空機の経路図も典型的なグラフであり、近年ではインターネット上のサイト間を複雑に結ぶリンク構造が大規模なグラフの例として注目を集めている。一般的には、複数の要素からなるシステムを表現するために、システムの構成要素を点で、それらの関係を辺で表したグラフを基に、さまざまなシステムの特性が研究されている。このようにみると、小規模なものから大規模なものまで、世の中はグラフ構造でみちているといえる。

膨大なグラフデータの特徴を把握するために、調査対象となるデータを分析する必要があるが、しかし、調査対象となるすべてのデータを分析することはコストや時間がかかり難しいといえる。巨大グラフを分析するために、一部の代表的グラフデータを抽出するランダムサンプリング手法が必要である。これまで様々な手法が提案されている。その中でも、グラフから代表的ノードも代表的グラフ構造も抽出することができるランダムウォークを基本とするサンプリング手法には注目されている。例えば、Henzinger ら[1]が提案した PageRank に基づくランダムウォークや Leskovec ら[6]の ForestFire 法等が挙げられる。

我々は、ランダムウォークの入次数の高いノードを訪れやすいという問題点に対して、入次数を考慮したランダムウォーク IRW[11]を提案し、入次数に偏らないサンプリングをできるようにした。しかし、IRW では入次数が分ることを前提としており、膨大なグラフデータに対して、入次数を事前に調べることが現実的ではない問題点が残っている。本研究では、Reservoir を用いて IRW の改善案 IRRW を提案し、入次数を前提条件としないノードのランダムサンプリング手法 IRRW を実現する。IRRW では、入次数を事前に調べる

必要がなく、さらに、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. 関連研究

Henzinger らは Web を対象としたランダムウォークによるサンプリングの手法として、VR (Visit Ratio)と PR (PageRank)を用い他サンプリングを提案した[1]。通常のランダムウォークでは多くのページからリンクを張られているページをより頻繁に訪れるという問題点に対して、PageRank の逆数を用いて均一なサンプルを収集する。PageRank は、Web のリンク構造のみで Web ページの客観的な重要度（人気度）を評価するための尺度である。PageRank を算出する際には、「多くの良質なページからリンクされているページは、やはり良質なページである」という再帰的な関係により、全てのページの重要度を判定する。あるページの現在のスコアを、そのページの出次数で割った値が、それぞれリンク先のページのスコアに加算されるという関係になっている。しかし、PageRank の算出はサンプリングよりコストがかかるので、一般のグラフサンプリングに適用できない問題点がある。

MRW (Metropolized Random Walk) は Metropolis-Hastings 法に基づくサンプリングである[5]。MRW では、サンプリング結果の均一な分布を得るために、Metropolis-Hastings 法を用いて遷移確率を修正している。

$$\varrho(x,y)=\begin{cases} p(x,y)\,min\left(\dfrac{deg(x)}{deg(y)},1\right), x\neq y \\ 1-\sum_{x\neq y}\varrho(x,y), \qquad x=y \end{cases}$$

具体的には、無向グラフにおいて x が隣接ノードから無作為（等確率）に任意の y を候補遷移先として選ぶ。しかし本当に y に遷移するかはさらに確率 Q(x, y)で決める。候補遷移先 y の次数が x の次数より大きいなら、遷移確率 Q(x, y)を小さめに調整する。候補遷移先 y の次数が x の次数より小さいなら、遷移確率 Q(x, y)を調整しない。このように、次数の高いノードに集中する傾向を抑えることができる。

3. 入次数を考慮したランダムウォーク

通常のランダムウォークでは，あるノードから次に進んでいくノードを選ぶために，そのノードの隣接ノードから無作為に一つ選んで移動していく。つまり，全ての隣接ノードに対して，偏りなく同じ遷移確率で選択する。例えば，有向グラフ G のノード v_i の出次数（v_i を始点とする有方向辺の数）を $outdeg(v_i)$deg$_i$とする。v_i からノード v_j への遷移確率$p_i(j)$は次のように求める。

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

この方法では，入次数の低いノードに比べると，入次数の高いノードに移動する確率が大きいため収集されやすい傾向があり，偏りが生じる問題がある。

この問題を対処するために、我々は入次数を考慮したランダムウォーク IRW（In-Degree Weighted Random Walk）を提案してきた[11]。IRW では、入次数の高いノード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，ノードが選ばれたチャンスを均等にすることができる。

アルゴリズム 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. $sample\geq samp\_size$ 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
   $\sum_{i=0}^{k-1} p_i < r \leq \sum_{i=0}^{k} p_i, \quad (p_0=0)$
5. $V'=V'\cup\{v_k\}$, $E'=E'\cup\{\langle v, v_k\rangle\}$
   $v=v_k$, $sample=sample+1$
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

図 1　入次数を考慮したランダムウォーク

IRW ではあるノードから次の移動先を選ぶときに，入次数の高いノードに遷移するチャンスを低めにする手法である。通常のランダムウォークでは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する。つまり

$$p_i(j) \propto \frac{1}{\mathrm{indeg}(v_j)} \quad \cdots\cdots\cdots(2)$$

式(2)で v_i から隣接ノード v_j への遷移する値は 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。

4. 入次数未知の場合の IRW の改善

グラフデータをサンプリングする際、入次数が分かっており、ノードを対象としたサンプリングを行う場合は、提案手法の IRW を用いる事ができる。しかし、グラフデータによっては、入次数が分からない場合がある。その場合は、入次数を調べてからサンプリングを行う方法が考えられるが、すべてのノードに対して入次数を調べることはコストがかかる。よって、本研究では入次数が分からない場合でも、入次数と近似した値を用いることにより入次数を調べなくてもサンプリングを行える手法として、IRRW を提案する。IRRW は、RRW を改良した手法となっており、通常の RRW では、抽出したノードに対して抽出回数を数え、入れ替えを行う際には、サンプル内のノードをランダムに選ぶ。選ばれたノードの抽出回数を減らし、抽出回数が 0 になったらサンプル内から削除するので、グラフ構造上の入次数の多いノードに対しての有意性は保たれている。IRRW では、抽出回数の多いノードに対して、高い確率で削除する。これにより、IRW で求める入次数の逆数を用いた遷移する値と同等の効果を得られると考える。

アルゴリズム：IRRW

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'= ϕ ; *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,⋯,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) freq = freq(v'$_1$)+ freq(v'$_2$),⋯, freq(v'$_n$)
   削除するノード v'$_k$ をランダムに選ぶ
   (2) v'$_k$ をランダムに選ぶ際、乱数 r∈[0, freq] を振り、r が以下の範囲内であればノード v'$_k$ を削除: sample=sample-1

$$\sum_{i=0}^{k-1} freq < r \le \sum_{i=0}^{k} freq, (freq_0 = 0)$$

5. そして V'=V'∪{ *t* }, E'=E'∪{ <*v*, *t* > }, freq(*t*)=1, v= *t*, *sample= sample+1*,
6. もし $t \in V'$
   freq(t)= freq(t)+1, E'=E'∪{<*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'∪{ *u'* }, v= *u'*,
   sample= sample+1, space=space+1
8. ステップ 2 へ移動し処理が続く

図 2 入次数を前提としない RW：IRRW

図 2 のアルゴリズムでは、抽出回数がどれだけ多くても、一度削除されるノードとして選ばれたら削除する。しかし、これだと抽出回数の優位性が保たれておらず、抽出回数の多いノードが不利な状態となってしまう可能性がある。よって、削除するノードに選ばれても抽出回数 freq が 0 になるまで削除されないアルゴリズム **IRRW2** とし、どちらの手法が良いかの評価も行う。

5. 評価実験

提案手法を評価するために、IRRW、IRRW2 の比較実験を行う。そして、通常のランダムウォーク RW、IRW、IRRW の結果を比較する評価実験を行った。評価尺度を用いる[6]。

- 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。

5.1. 実験データ

本実験では Google ProgrammingContest2002 年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

このデータセットの入次数、出次数、強連結成分、クラスター係数の分布状況は**図 3**～**図 7** に示している。**図 3** の横軸は入次数であり、縦軸は入次数に該当するノードの数である。**図 4** の横軸は出次数であり、縦軸は出次数に該当するノードの数である。

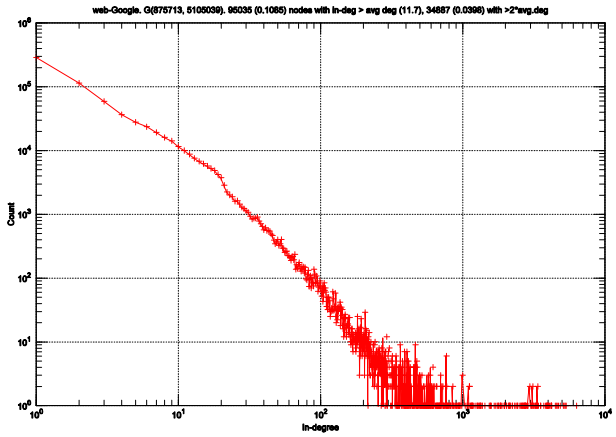

図 3 Google グラフの入次数の分布

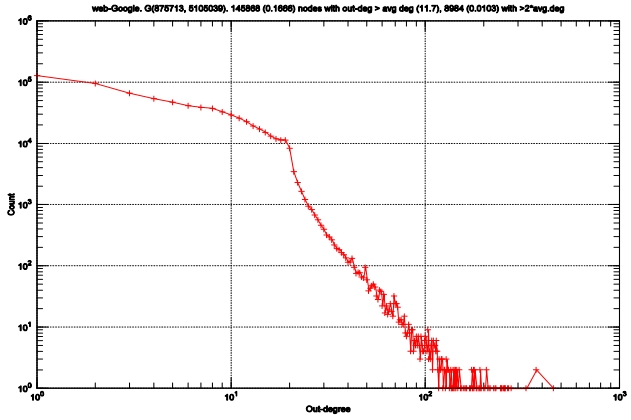

図 4 Google グラフの出次数の分布

図 5 の横軸は次数であり、縦軸は次数に該当するノードの平均クラスタ係数である。**図 6** の横軸は強連結部分グラフのサイズであり、縦軸はサイズに該当する強連結部分グラフの数である。**図 7** の横軸は弱連結部分グラフのサイズであり、縦軸はサイズに該当する弱い連結部分グラフの数である。

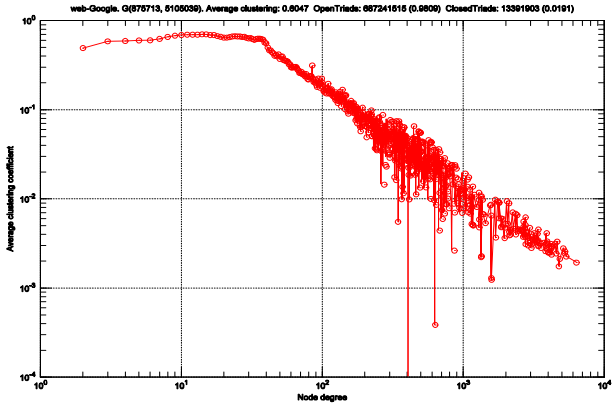

図 5 Google グラフの平均クラスター係数の分布

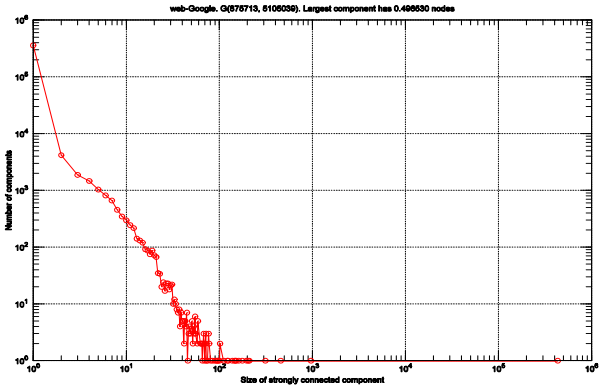

図 6 Google グラフの強連結成分の分布

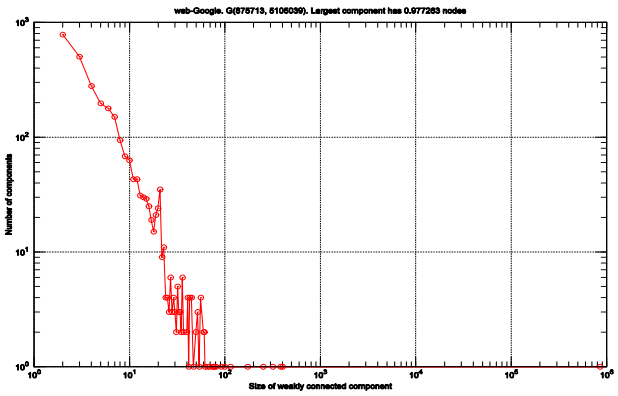

図 7 Google グラフの弱連結成分の分布

5.2. 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を設定している。

(1) jump_prob=0.15

(2) samp_size=51050

(3) samp_space = samp_size*1.5

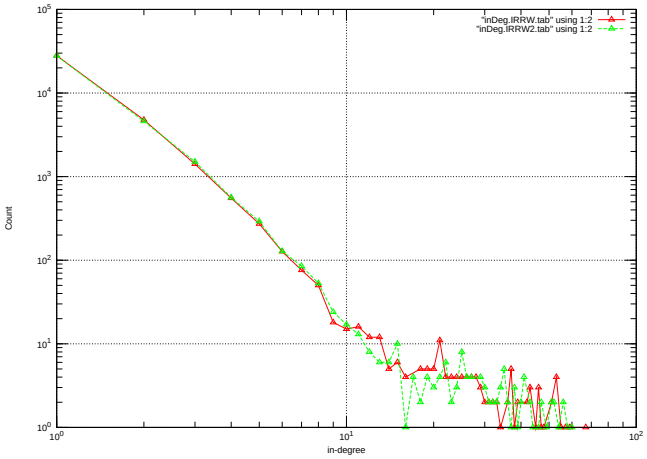

図 8 抽出サンプルの入次数の分布

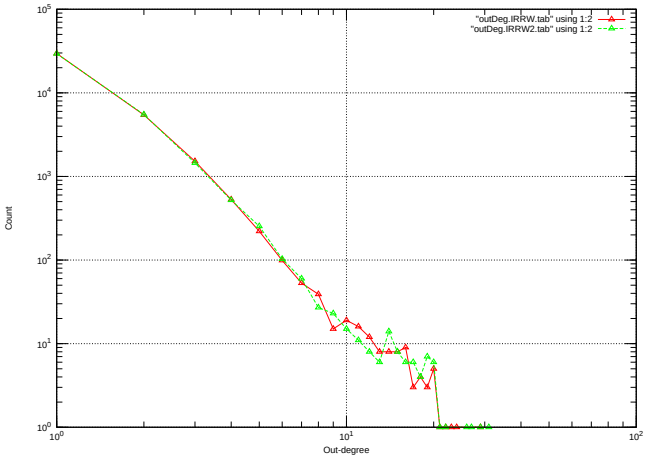

図 9 抽出サンプルの出次数の分布

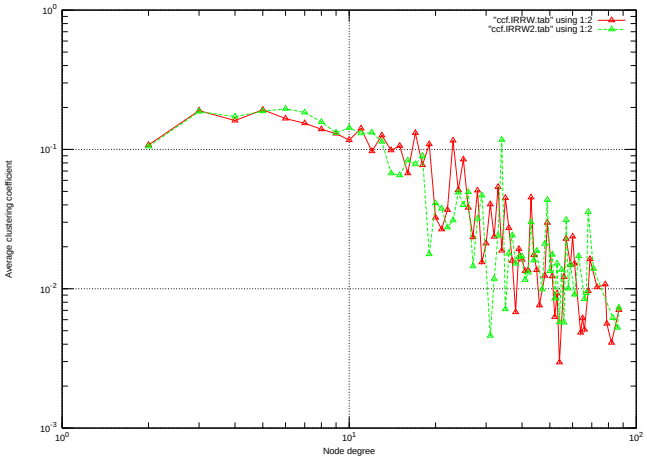

図 10 抽出サンプルの平均クラスター係数の分布

まず、IRRW、IRRW2 の実験結果は、**図 8～図 10** の各グラフに示している。グラフの結果から IRRW、IRRW2 とも似た結果となり、それぞれの手法に差がないことがグラフから分かる。入次数の偏りにおいて必ずしも抑制できておらず、IRW と近似した結果とならなかった。この原因として、IRRW の処理時間の長さから抽出対象サイズを狭めたことが原因に挙げられる。

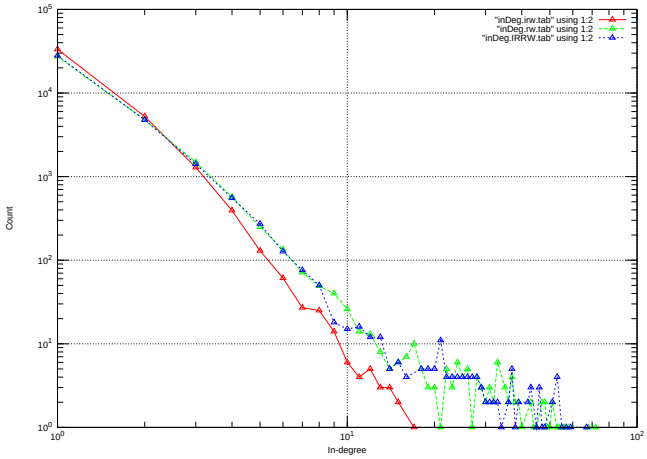

図 11 抽出サンプルの入次数の分布

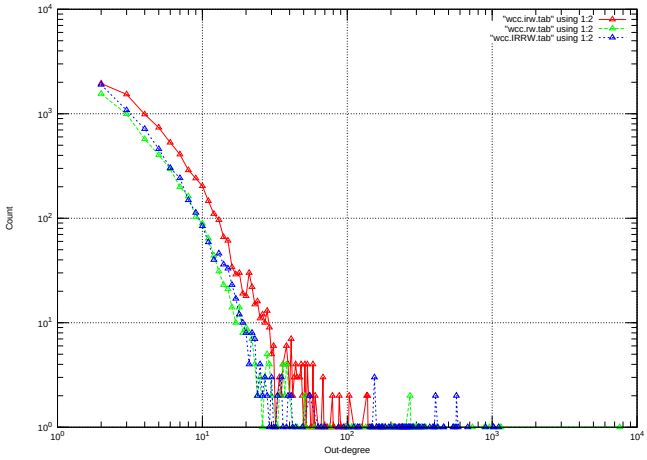

図 12 抽出サンプルの出次数の分布

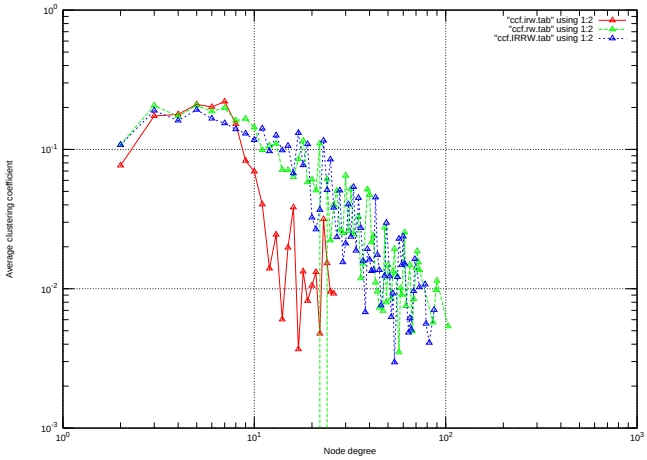

図 13 抽出サンプルの平均クラスター係数の分布

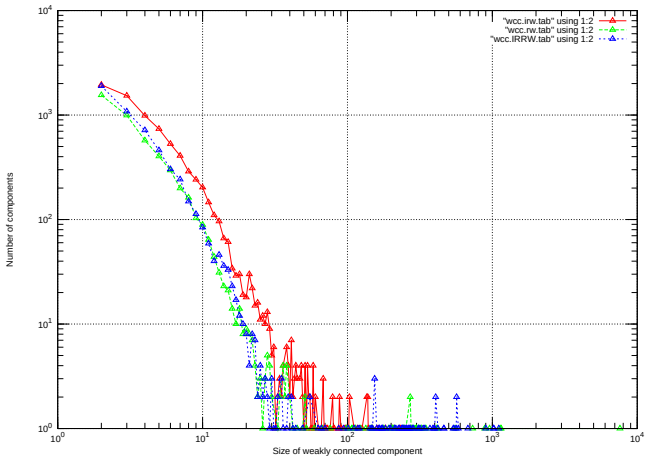

図 14　抽出サンプルの弱連結成分の分布

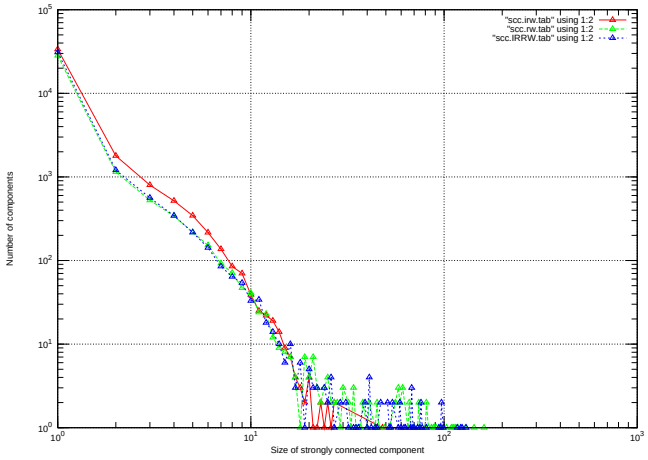

図 15　抽出サンプルの強連結成分の分布

6. 終わりに

本研究では、[11]で提案した IRW の入次数の分からないグラフデータに対して、サンプリングを行うことが難しいといった問題点に対して、入次数が分からなくてもサンプリング手法として IRRW を提案した。IRRW では、[11]で提案した RRW を改良した手法であり、ノードの抽出回数が多いほどサンプルとして削除される確率を高くする。これにより、入次数による偏りを抑えることができると考えられる。

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

今後の課題として、より抽出対象サイズを広くした評価実験が必要であり、また、処理時間の長さの問題についても考察していく。更に既存のほかのサンプリング手法、例えば、MRW との比較も行う予定である。

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] W. Hastings. Monte Carlo Sampling Methods Using Markov Chains and Their Applications. Biometrika, 57,1970.

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

[7] 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

[8] 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

[9] 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

[10] 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

[11] 仲前　晋太郎,成　凱 ;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, DEIM 2010

[12] D. Stutzbach,R. Rejaie, N. Duffield, S. Sen, and W. Willinger. On Unbiased Sampling for Unstructured Peer-to-Peer Networks.TON,16(2),Apr.2008.

[13] Rasti, A.H.; Torkjazi, M.; Rejaie, R.; Duffield, N.; Willinger, W.; Stutzbach, D.: Respondent-Driven Sampling for Characterizing Unstructured Overlays. INFOCOM 2009, IEEE.