D-012

ランダムウォークによるグラフデータのサンプリング手法

Sampling Methods for Graph Data based on Random Walks

仲前　晋太郎†　　中村　三四郎†　　成　凱‡
Shintarou Nakamae　Sanshiro Nakamura　Kai Cheng

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おり注目が集まっている。膨大なグラフデータの特徴を把握するために，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り，これまでは様々な手法が提案されている[1][3]。中でも，グラフから代表的頂点も，代表的リンク構造も抽出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 らが提案した PageRank に基づくランダムウォーク[1]，Leskovec らの ForestFire 法等が挙げられる[3]。

しかし，これらの手法では，頂点を対象とするサンプリングと，グラフ構造を対象とするサンプリングの違いは考慮されていないため，分析目的に応じて適切なサンプリング手法を選ぶことができない。例えば，頂点を対象としたサンプリングでは，一部の頂点に偏らないサンプリング手法が望まれている。しかし，通常のランダムウォークだと，入次数の大きな頂点に偏ってしまい，満足できるサンプルを取得することができない。また，膨大なグラフ構造からサンプルを抽出する際に，必要なサンプルサイズが前もって決めることができない場合がある。そのため，ランダムウォークが止まるまで十分なサンプル数が得られたかどうかは判断しにくい。

そこで本論文では，我々は，頂点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，グラフの各頂点を偏りなくサンプリングできるようにする。また，グラフ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

通常のランダムウォークでは，ある頂点から次に進んでいく頂点を選ぶために，その頂点の隣接頂点から無作為に一つ選んで移動していく。つまり，全ての隣接頂点に対して，偏りなく同じ遷移確率で選択する。例えば，有向グラフ G の頂点 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 から頂点 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，入次数の低い頂点に比べると，入次数の高い頂点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。

† 九州産業大学大学院　情報科学研究科
‡ 九州産業大学　情報科学部

2.1 入次数を考慮したランダムウォーク

入次数の高い頂点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，頂点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 ではある頂点から次の移動先を選ぶときに，入次数の高い頂点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頂点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，隣接する頂点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 から隣接頂点 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，入次数の多い頂点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。

アルゴリズム IRW

1. 初期化：$u \in U$, $z=1$, $v=u$, $V'=\{u\}$, $E'=\phi$;
2. $z \geq samp\_size$ なら，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を頂点 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. $V'=V' \cup \{v_k\}$, $E' = E' \cup \{\langle v, v_k \rangle\}$
   $v=v_k$, $z=z+1$
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，
   $V'=V' \cup \{u'\}$, $v=u'$, $z=z+1$
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 はグラフ構造上の偏りがないため，頂点を対象としたサンプリングが必要なとき，有効である。例えば，ブログ上で様々な話題の分布を調べる際に，IRW のような，リンクの多いブログに偏らない収集方法が必要である。

2.2 Reservoir を用いたランダムウォーク

代表的グラフ構造を抽出することを目的とするサンプリングを行う際に，必要なサンプルサイズを事前に決めることができない場合がある。ここで我々は RRW（Random Walk with a Reservoir）を提案する。Reservoir とは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おり，サンプルの一部を入れ替えながらサン

プリング¥を行う事ができる。

ランダムウォークで抽出された頂点は Reservoir に一時保持しておく。同じ頂点が複数回抽出された場合は，抽出回数を数える。Reservoir の容量を超えた場合は，すでに抽出された頂点をランダムに選んで削除してスペースを空ける。削除となる頂点の抽出回数が 1 より大きい場合は，抽出回数を 1 減らす。この方法ではランダムウォークの範囲が広がるにつれ，サンプルグラフは元のグラフをよく代表できるようになっていく。また，サンプルサイズはランダムウォークの範囲の割合で決めることができ，事前に必要なサンプルサイズを決める必要がなくなる利点がある。

アルゴリズム RRW

1. 初期化：$u \in U$, $z=1, s=1$, $freq(u)=1$,
   $V'=\{u\}$, $E'=\phi$; $v=u$,
2. $s \geq samp\_space$ 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。$s=s+1$
4. もし $t \notin V'$
   もし $z \geq samp\_size$,以下の(1)－(3)を $z < samp\_size$ 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 $m=m-1$.
5. $V'=V' \cup \{t\}$, $E'=E' \cup \{<v, t>\}$,
   $freq(t)=1$, $v=t$, $z=z+1$,
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，$V'=V' \cup \{u'\}$, $v=u'$,
   $z=z+1$, $s=s+1$
8. ステップ 2 へ移動し処理が続く

図 2 Reservoir を用いたランダムウォーク

3. 評価実験

提案手法を評価するために，通常のランダムウォーク RW と提案手法 IRW，RRW の結果を比較する評価実験を行った。以下の評価尺度を用いる：（1）ccf: クラスター係数の分布状況；（2）inDeg: 入次数の分布状況；（3）outDeg: 出次数の分布状況；（4）scc: 強連結成分の分布状況；（5）wcc: 弱連結部分の分布状況。また，本実験で Google ProgrammingContest2002 で提供されたデータセットを利用する。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。

実験結果は紙面の制限で入次数と平均クラスター係数の分布のみ**図 3** と図 4 で示している。グラフから以下の結果が分かった。IRW では**図 3** の各頂点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。

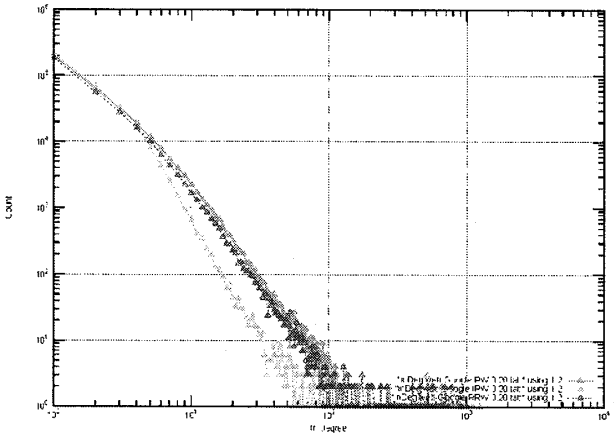

図 3. 抽出サンプルの入次数の分布

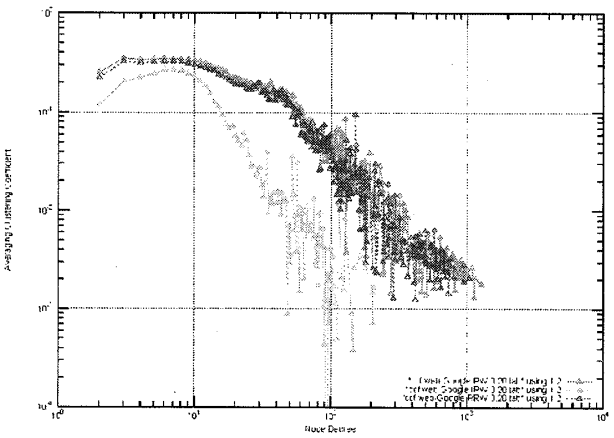

図 4. 抽出サンプルの平均クラスター係数の分布

4. 終わりに

本論文では，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頂点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，頂点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 とグラフ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。提案手法の有用性を検証するために，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] Vitter, J.S Random Sampling with a reservoir. ACM Trans. Math. Softw. 11(1) .1985 Mar., pp.37-57

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