

グラフデータを対象とした重み付き Reservoir Sampling

王 キン^{†1} 成 凱^{†2}

近年、ソーシャルメディアや SNS の爆発的な普及に伴い、データ解析の対象は従来の単純な表構造データから、人・物・場所といった多様な情報のつながりを表現可能なグラフ構造へとシフトしつつある。グラフの様々な特徴量を得るためにサンプリング手法が重要である。これまでのグラフサンプリング手法では、ランダムウォークがよく用いられ、グラフ構造を探索しながらサンプルを抽出する。しかし、ランダムウォークは、一部の次数の高いノードに偏り、一様なサンプルを抽出することができない問題点がある。本研究では、我々は重み付き Reservoir Sampling アルゴリズム RWW を提案し、高次数ノードへの偏りを解消できるだけでなく、グラフ探索範囲の拡大やグラフの進化に対応できるサンプル抽出を実現する。シミュレーション実験では、評価対象のグラフをランダムグラフとして生成し、提案手法によってサンプルを抽出する。抽出した結果の次数分布を中心に、理想なサンプル結果と比べ、類似度を評価した。

Weighted Reservoir Sampling from Large Graphs

XIN WANG^{†1} KAI CHENG^{†2}

Most studies in graph sampling use random walk as a basic technique to collect sample. However, it is well known that random walk will bias towards nodes of high degree, to obtain a uniform sample, transition probabilities of high degree nodes should be corrected. In this paper, we propose a reservoir based approach to uniformly random sampling of large graphs. We developed a sampling algorithm called RWW (Random Walk with a Weighted Reservoir), where random walk is also used to collect nodes from the graph. To deal with the biases, a weighted reservoir is employed to maintain the collected items, where high degree nodes are associated with a low priority. We perform simulation-driven experiments to evaluate the algorithms and show that most of these algorithms perform well compared to the standard sampling results.

1. はじめに

近年、ソーシャルメディアや SNS の爆発的な普及に伴い、データ工学の分野で扱われるデータは、従来の単純な表構造データから、人・物・場所といった多様な情報のつながりを表現可能なグラフ構造へとシフトしつつある。ソーシャルネットワークサービスにおける利用者同士の交友関係、ビジネス分野における企業間の取引関係、オンラインショッピングサイトにおける顧客商品関係、嘘の広がり、意見形成、疫病の拡散など、グラフを対象とする解析が必要不可欠になっている。

大規模のグラフデータを解析するため、一部代表的なデータを手作為に抽出するランダムサンプリング手法がよく用いられる。グラフサンプリングの特徴として、既知のノードから隣接ノードへアクセスし、グラフ構造を探索しながら、サンプルを集めなければならない。グラフの構造や探索方法によって、集められたサンプルは偏りや相関性等の問題が生じる可能性が高い。これまでは様々なサンプリング手法やアルゴリズムが提案された 9)13)15)16)。従来の研究ではランダムウォーク (Random Walk) を用いてグラフ

を探索し、経由したノードをサンプルとして抽出する。ランダムウォークとは、グラフ上であるノードからランダムに隣接ノードを一つ選んでそこに移動し、移動先からまた同じ方法で次のノードを選び移動を続けるグラフ探索の仕組みである。ランダムウォークは現在ノードの情報しか必要とせず、探索のコストが低い利点がある一方、次数の高いノードに偏ったり、連結度の高い部分グラフにとどまったり、グラフ全体をカバーする時間が長くカバー率が低い問題点が指摘されている 15)16)。

そこで、我々は、グラフ探索にランダムウォークを前提に、サンプルを管理するに重み付き Reservoir を導入し新しいグラフサンプリング手法を提案する。高次数ノードへの偏りを解消できるだけでなく、グラフ探索範囲の拡大やグラフの進化に対応できるサンプル抽出を実現する。

2. Random Walk によるグラフ探索

グラフサンプリングでは、グラフ探索にランダムウォークを基本としている。単純ランダムウォーク (random walk) とは、現在訪問中のノードより、次の訪問先を決めるとき、確率的に無作為 (ランダム) に行われるグラフ探索方法である。

グラフ $G=(V, E)$ におけるランダムウォークでは、各ノードの探索される確率はつぎのような定常分布に従う。

^{†1} 九州産業大学
Kyushu Sangyo University
^{†2} 九州産業大学
Kyushu Sangyo University

$$\pi_v = \frac{d(v)}{2|E|}$$

つまり、ノード v の訪問確率は v の次数 $d(v)$ に比例する。ランダムウォークで訪問されたノードはそのままサンプルとして抽出されると、次数の高いノードを高い確率で抽出されることになる。

このような偏りを解消するため、次数の高いノードへ遷移する確率を抑える必要がある。MRW (Metropolized Random Walk) がこのような思想でできた改良の一つとしてよく知られている[8]。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。

$$q(x, y) = \begin{cases} p(x, y) \min\left(\frac{\deg(x)}{\deg(y)}, 1\right), & x \neq y \\ 1 - \sum_{x \neq y} q(x, y), & x = y \end{cases}$$

つまり、ノード x から次の遷移先を決めるために、隣接ノードから無作為（等確率）に任意の y を選ぶ。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。

グラフ探索はサンプリングの効率に影響が大きい。探索方法によって、グラフをカバーする時間(Cover Time)が違う。カバー時間とはグラフ上のあるノードから始め、すべてノードを訪問するまでにかかる移動数 (step) の期待値である。例えば、ノード数 n の連結グラフをすべてカバーするために、単純ランダムウォークのカバー時間の上限が $O(n^3)$ であると知られている[1]。グラフの一部を訪問する場合は、一定の移動数でカバーするグラフ範囲が大きいほどよい。

3. Reservoir を用いたサンプル抽出

サンプル抽出では、グラフ探索で得られたノードをサンプルとして抽出するかどうかを決める。

Reservoir を用いたサンプリング (Reservoir Sampling) は母集団のデータが一度しかアクセスできない場合や、新しいデータが無限に出てくる場合に、データ出現の順序に関係なく等確率でサンプルを抽出する手法である[7]。サンプルサイズを k として事前にきめておく。

母集団から k 番目までのデータはそのままサンプルに入る。そして、 i 番目 ($i > k$) 以降のデータが到来したとき、 $[1..i]$ の範囲で乱数 r を生成し、 r は k より小さい場合のみ、新しいデータをサンプルの r 番目として保存し、元のサンプルが上書きされて消える。 R が k より大きい場合は新しいデータを無視し Reservoir はそのまま変わらない。Reservoir を利用する際に、サンプルサイズ (Reservoir 容量)

を事前にきめる必要がある。

定義 1 (Uniform Reservoir)

一様な Reservoir とは標本空間 V (サイズ未知) からサイズ k のサンプルを一様に抽出するデータ構造であり、標本空間の一部である m 個の要素を処理した時点では各要素の抽出確率は k/m である。

一様な Reservoir は次の Algorithm A にしたがって管理される。

アルゴリズム A:

- (1). サイズ k の配列を初期化する;
- (2). 最初 k 個の要素を配列に格納する;
- (3). $j > k$ の各要素 j に対し i と j の間から乱数 r を振る。
 $r \leq k$ のときのみ、配列の r 番目の要素を削除し、要素 j を r 番目に格納する。

定義 2 (Weighted Reservoir)

重み付き Reservoir とは標本空間 V (サイズ未知) からサイズ k の重み付き要素をサンプルとして抽出するデータ構造であり、標本空間の一部である m 個の要素が処理された時点では各要素の抽出確率はその要素の重みに比例し、つまり $w_i/(w_1 + w_2 + \dots + w_m)$ 。

重み付き Reservoir は次の Algorithm B にしたがって管理される。

アルゴリズム B:

- (1). サイズ k の配列を初期化する;
- (2). 最初 k 個の要素を配列に格納する;
- (3). 重み w_i の要素 i が格納されるとき、パラメータ w_i の指数分布に従う乱数 p_i を発生し要素 i の優先度とする。このとき優先度の最小値を T とする。
- (4). 続いて $j > k$ の要素 j を次のように処理する。まず、パラメータ w_j の指数分布に従う乱数 p_j を発生し要素 j の優先度とする。つぎに $p_j > T$, 配列から優先度最小の要素を削除し、要素 j を配列に格納する。最後に最小優先度 T を更新する。

定理 1:

アルゴリズム B で管理されるサイズ k の Weighted Reservoir において、標本空間の一部である m 個の要素が処理された時点で各要素の抽出確率は要素の重みに比例し、 $w_i/(w_1 + w_2 + \dots + w_m)$ である。

命題 1:

互いに独立する指数分布の確率変数 $X_1, X_2, \dots, X_m$ のパラメータはそれぞれ $\theta_1, \theta_2, \dots, \theta_m$ とする。 $\min(X_1, X_2,$

..., X_m) はパラメータ $\sum_{i=1}^m \theta_i$ の指数分布の確率変数である。さらに

$$\Pr(\min(X_1, X_2, \dots, X_m) = X_i) = \frac{\theta_i}{\sum_{i=1}^m \theta_i}$$

4. Reservoir を用いたグラフサンプリング

これまでの議論からわかるように、適切なグラフサンプリング手法を考案するために、グラフ探索、サンプル抽出、サンプル更新を含む各要素を考慮しなければならない。

本章ではサンプル更新に用いる手法によって2種類のグラフサンプリングを提案する。アルゴリズムを説明するために、表1に示すような記号を使う。関数 Move() と Sample() は探索方法、サンプル抽出方法に依存し、詳細は具体的にアルゴリズムで決める。

Reservoir を用いたグラフサンプリング (RWW : Random Walk with a Weighted Reservoir) を紹介する。グラフ探索で次から次へ移動しながら、サンプルを抽出する。抽出されたノードを Reservoir に格納しサンプルを更新する。

表1 アルゴリズム RWW に使われる記号

記号	説明
$G<V, E>$	探索対象となるグラフ
S	探索開始点の集合
K	サンプルサイズ
Reservoir	サンプル一時保存用 Reservoir
$d(v)$	ノード v の次数

アルゴリズム RWW:

入力:

1. $G<V, E>$ 、S、k
2. Reservoir[1...k]: 重み付き Reservoir

出力:

抽出結果: 部分グラフ $G'=(V', E')$ (サンプル)

- (1). $i=1$;
- (2). 始点集合Sから任意一つをランダムに選ぶ。
- (3). アルゴリズムBに従い、重み付きReservoirにノード v_i を格納する。重みは $1/d(v_i)$ とする。
- (4). v_i から単純ランダムウォークの次の遷移先を訪問する。
- (5). $i = i + 1$;
- (6). サンプリングが終了する?
いいえ: (3) に戻り処理を続ける。
はい: アルゴリズムを終了する。Reservoirよりサンプルを出力する。

定理2:

アルゴリズム RWW は一様な分布のサイズ k のサンプル

が抽出される。また、探索されたグラフの一部であるノードが同じ確率で抽出できる。

定理2の成立は次のように証明できる。まず、単純ランダムウォークでノード v の訪問確率は

$$\pi_v = \frac{d(v)}{2|E|}$$

また、定理1より、ノード v が重み付き Reservoir に残る確率は重み $1/d(v)$ に比例する。事象 A: ノードが訪問されることと事象 B: Reservoir に残ることは互いに独立するので両者を掛け合わせて、 v の抽出確率は定数であることがわかる。

5. 実験評価

提案アルゴリズムの性能を評価するために評価実験を行う。実験では、代表的なグラフをランダムグラフとして生成し、抽出結果の次数分布を中心に評価を行う。

実験用のシミュレーターは、JUNG (Java Universal Network/Graph Framework) を用いて実装した。JUNG は Java でグラフ構造の生成、処理、可視化等ができるオープンソースのライブラリーである。

(1) 評価尺度

グラフサンプリングにおけるグラフ探索アルゴリズムを評価するために、抽出されたサンプルの次数分布が基準となる次数分布との誤差で評価する。

母集団の要素がサンプルとして平等に選ばれることで望ましい。一部の要素が他より選ばれやすい場合に偏りが生じる。偏ったサンプルは一般に誤った推定に与える。ここで以下の三つの尺度を用いてサンプルの次数分布と基準となる次数分布を比較し、その誤差を評価する。

- a. カルバック・ライブラー情報量 (Kullback-Leibler divergence: **KL**): 離散確率分布 P の Q に対するカルバック・ライブラー情報量は以下のように定義される。

$$D_{KL}(P||Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}$$

- b. 対称カルバック・ライブラー情報量 (Symmetric Kullback-Leibler divergence: **SK**)

$$D_{SK}(P, Q) = \frac{1}{2} (D_{KL}(P||Q) + D_{KL}(Q||P))$$

- c. 平均二乗誤差 (Squared Error: **SE**): 2つの確率分布 P と Q の差異を以下の式で評価する。

$$D_{SE}(P, Q) = \frac{1}{n} \sum_i (P(i) - Q(i))^2$$

カルバック・ライブラー情報量は2つの確率分布の差異を計る尺度であり、以下のような性質を持っている。

1. KL 情報量は常に 0 以上である。
2. モデルの分布が真の分布と一致するときに KL 情報量はゼロとなる。
3. カルバック・ライブラー情報量には対称性がない

$$D_{KL}(P||Q) \neq (Q||P)$$

(2) 評価対象

評価対象となるアルゴリズム単純ランダムウォーク RW、MRW (Metropolized Random Walk)、および提案後リズム RWW (Random Walk with a Weighted Reservoir) を評価する。

RW と MRW の場合は、サンプル抽出に Bernoulli サンプルリングを使用し、一定の確率で探索し集めてきたノードをサンプルとして抽出する。

また、実験に使うグラフは以下の三種類のランダムグラフを生成し、それぞれに対して、各サンプリングアルゴリズムを使って、サンプルを抽出し、評価を行う。

Uniform : Erdős-Rényi ランダムグラフ。任意のノード対が一定確率で隣接するグラフである。

PowerLaw : Eppstein ランダムグラフ。次数分布がべき乗則に従う。

SmallWorld : Kleinberg ランダムグラフ。スモールワールド (Small World) の性質をもつグラフ。つまり、辺の数がノード数の数倍程度しかないが、ノード間の平均距離が小さくて高度にクラスター化している。

抽出対象グラフ $G<V, E>$ について以下の共通のパラメータを使っている。

- ・ ノード数 : $n=|V|=22,500$
- ・ 辺数 : $m=|E|=151,8750, (n \times n \times 0.003)$
- ・ 探索開始点集合 $S : s=|S|=22, (n \times 0.001)$
- ・ サンプル抽出率 p :
 $p = \{ 0.05, 0.10, 0.15, 0.2, 0.25, 0.30, 0.35 \}$;
- ・ 最大探索移動数 $q=225,000, (n \times 10)$ 。無限ループに陥らないようにこの数以上探索しない。

誤差を評価するためには基準となる次数の確率分布を用いる。本研究では、元のグラフにおける次数分布を基準とする。また、理想なサンプルとして、グラフ構造を考慮せず、すべてのノードを平等に抽出する。このような理想的状況は現実的には存在しないので Oracle を呼ぶ。

(3) 評価結果

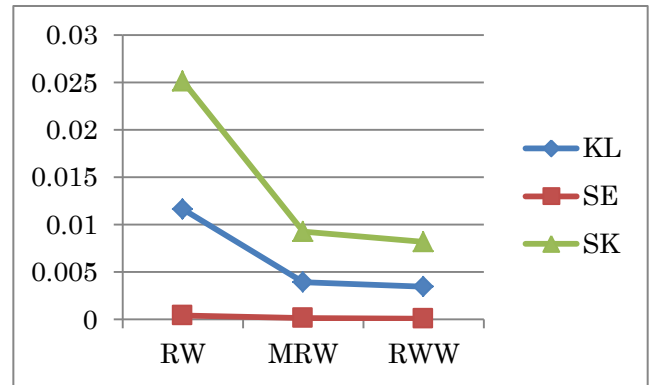


図 1 ランダムグラフ Uniform の評価結果(p=0.25)

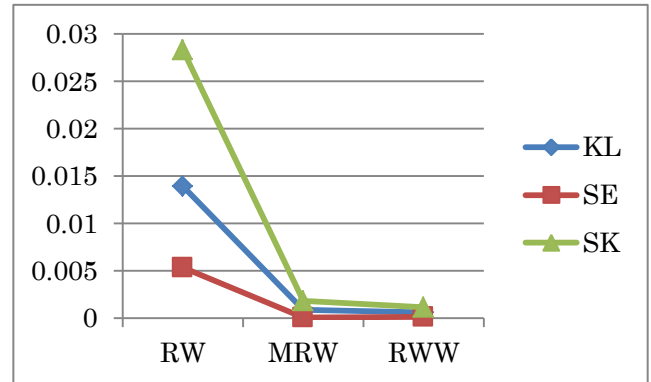


図 2 ランダムグラフ SmallWorld の評価結果(p=0.25)

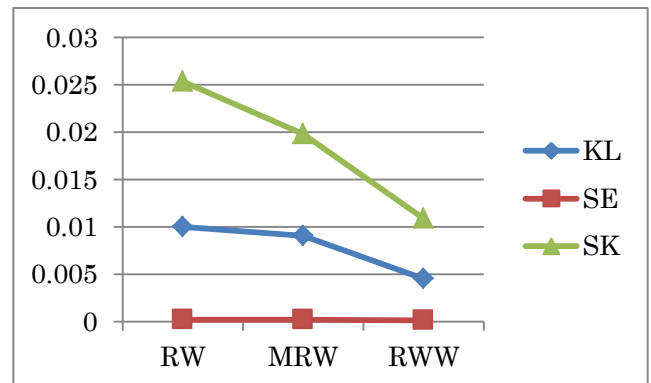


図 3 ランダムグラフ PowerLaw の評価結果(p=0.25)

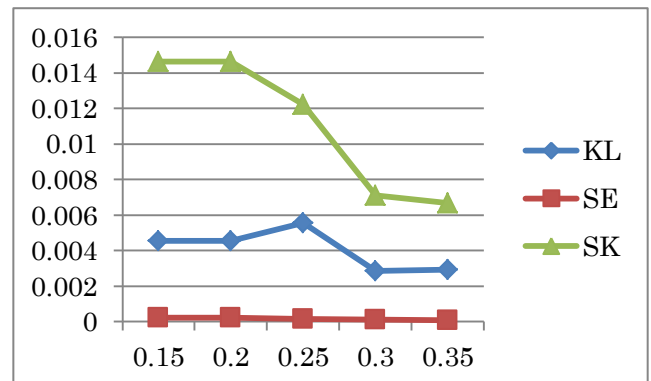


図 4 異なる抽出率の評価結果 (RWW)

実験結果を見ると分かるように、重み付き Reservoir を用

いる場合は、RWW はより偏りの少ない一様なサンプルを抽出ことができる。

また、抽出率が高くなると、サンプルの質がよくなる傾向がある。

6. 終わりに

本研究では大規模グラフより一様適切なサンプルを抽出するために、重み付き Reservoir を用いる手法を提案した。グラフ探索は単純なランダムウォークで良く、抽出されたサンプルは MRW よりよい結果が確認できた。

今回の評価実験は、機器設備の条件の制限で比較的に小規模のランダムグラフを使って行った。今後の予定として、実データを使った大規模な実験は必要と思われる。また、探索アルゴリズムとグラフのカバー時間、カバー率の関係を調べる必要がある。

参考文献

- 1) D. J. Aldous, Lower bounds for covering times for reversible Markov chains and random walks on graphs, J. Theoretical Probability 2(1989), 91-100
- 2) 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.
- 3) 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.
- 4) D. Chakrabarti, C. Faloutsos; Graph mining: laws, tools, and case studies. Morgan & Claypool Publishers, 2012
- 5) Y. Tille; Sampling algorithms, Springer, 2010
- 6) 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
- 7) Vitter, J.S; Random Sampling with a reservoir. ACM Trans. Math. Softw. 11(1) .1985 Mar., pp.37-57
- 8) W. Hastings. Monte Carlo Sampling Methods Using Markov Chains and Their Applications. Biometrika, 57, 1(1970), pp.91-109.
- 9) J Leskovec, C Faloutsos Sampling from Large Graphs. Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2006
- 10) 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
- 11) 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
- 12) Z.A.Bar-Yossef, A. Berg. S. Chin, and J. Fakcharoenphol; Approximating aggregate queries about web pages via random walks. In Proceedings of the 26th International Conference on Very Large Data Bases. 2000
- 13) P. Rusmevichientong, D.M.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
- 14) 仲前 晋太郎, 成 凱; Reservoir を用いた巨大グラフのランダムサンプリング, DEIM 2011

- 15) 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.
- 16) A.H. Rasti, M. Torkjazi, R. Rejaie, N. Duffield, W. Willinger, D. Stutzbach; Respondent-Driven Sampling for Characterizing Unstructured Overlays. INFOCOM 2009, IEEE.
- 17) 鹿島 久嗣, グラフとネットワークの構造データマイニング, 電子情報通信学会誌 93(9):797-802, 2010-09-01
- 18) JUNG : Java Universal Network/Graph Framework, <http://jung.sourceforge.net/>