2H3-3

Yule-Simon過程によるタグ共起ダイナミクスのモデル化と分析

Modeling and Analysis of Tag Co-occurrence Dynamics Using Yule-Simon Process

佐藤晃矢 Koya Sato　岡瑞起 Mizuki Oka　橋本康弘 Yasuhiro Hashimoto　加藤和彦 Kazuhiko Kato

筑波大学大学院 情報工学研究科
Graduate school of SIE, University of Tsukuba

We apply Yule-Simon process describing the cascade phenomena to the generative model of tagging in Social Tagging System, and expand to the model of co-occurrence phenomena. To compare with the empirical data, we discuss the hierarchy of the tag co-occurrence structure.

1. はじめに

Delicious，Flickr，YouTube，Twitter，Facebook などのオンラインコンテンツ共有サービスでは，ユーザが任意の文字列を付与することによって投稿されるコンテンツの管理を行う Social Tagging System(STS) が採用されている．STS においては一般的にコンテンツに対して複数のタグが使われ，タグ間に共起が生まれる．タグには意味が付与され，利用されることから共起しやすいタグと共起しにくいタグが存在している．このような共起を利用することでタグの推薦やタグの階層を抽出することが可能であることから重要な統計量の一つである．そこで，本研究ではタグ生成モデルとして利用される Yule-Simon 過程に共起を考慮する拡張を加える．Yule-Simon 過程はべき分布を表現することが可能なモデルであり，べき分布はタグ付け以外の様々なシステムに現れることから，他の分野でも広く使われるモデルとなっている [Simon 55]．本研究ではタグの共起関係を，Yule-Simon 過程のような意味を含まない古典的なモデルでどの程度再現可能であるかに注目した．

モデルの検証には共起構造を捉えることが可能なタグ共起ネットワークの構築と，モデルと商用サービス RoomClip(http://roomclip.jp/) から得られる実データ，それぞれの統計量の比較を通して行った．

2. 提案モデル

Yule-Simon 過程では実際のタグ付けに見られる，1 つのコンテンツに対して複数のタグが付与されるというタグの共起は考慮していない．一般的に，投稿されたコンテンツに対してユーザは複数のタグを付与する，通常，それらのタグの間には何らかの関係があることが多い．

そこで，タグの共起を扱うためにウィンドウという概念を Yule-Simon 過程に導入した Windowed Yule-Simon 過程を提案する．ここでウィンドウは投稿されるコンテンツのことを指し，J 番目に投稿されたコンテンツに対して付与されるタグの数をウィンドウサイズ (Ω_J) とする．通常，同じコンテンツに対して同じタグが 2 度使われることは無いので，ウィンドウ内で使われるタグの重複は無いような制約を与える．提案モデルの概念図を図 1 に示す．Windowed Yule-Simon 過程ではウィンドウを作成し，ウィンドウ内で試行ごとに新たな種類のタグを生成，またはこれまでに使われたタグの中から選択を行う．新たな種類のタグを選択する確率を，Yule-Simon 過程と同様に (p) とする．既存のタグから選択する確率を $(1-p)$ とし，これまでに選択された全てのウィンドウで利用されたタグの中から，ウィンドウ内ですでに使われたタグを除いて、ランダムに 1 つを選択する．モデルをシミュレートするにあたって，新たな種類のタグの選択確率 p，ウィンドウサイズ Ω の分布関数が必要となる．実データとの比較の観点から $p = 0.05$ を適当な値として与える．また，ウィンドウサイズはポアソン分布（$P(X = q) = \frac{\lambda^q e^{-\lambda}}{q!}$）に従うと仮定し，$\lambda = 2.89$ の分布を与える．また，ウィンドウを導入したことで試行回数 N の小さな場合にはタグの種類 $V(N)$ よりウィンドウサイズ Ω が大きくなってしまう．そこで初めから存在するタグの種類として $V(0) = 10$ 種類のタグをモデルに与える．

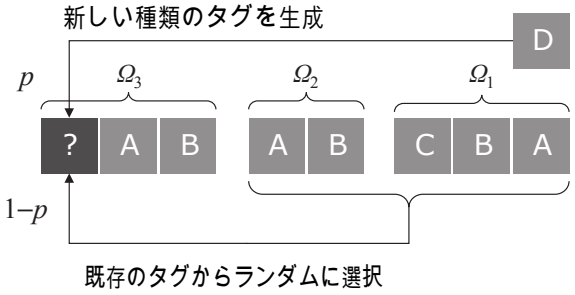


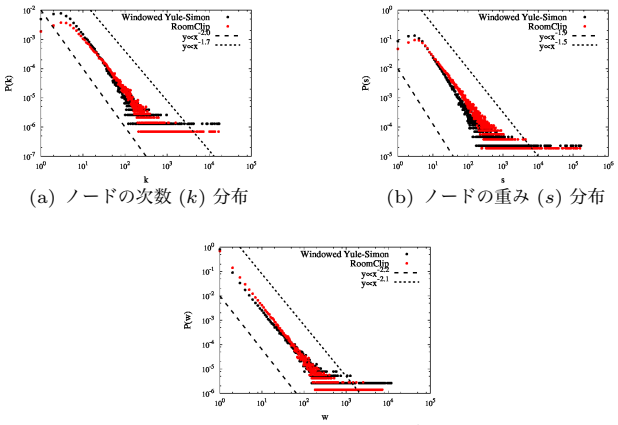
図 1: Windowed Yule-Simon 過程の概念図．ウィンドウはコンテンツを示し，ウィンドウ内においてはタグの重複は許さないという制約を加える．その上で，Yule-Simon 過程と同様の過程でタグを生成，選択する．

Yule-Simon 過程を拡張したモデルにより，タグの共起構造を捉えることが可能となった．次章では実データにより示されるタグの共起構造を，Windowed Yule-Simon 過程により再現できていることを確かめる．

3. 分析

タグの共起構造を分析するために，タグ共起ネットワークを利用する．タグ共起ネットワークは重みありの無向グラフで表現される．ノードはタグの種類に対応し，エッジは同一のウィンドウ内で出現したノード間に貼られ，共起をあらわす．リンクには重みが付与され，ノード間の共起回数で与える．実デー

連絡先: 佐藤晃矢，koya@osss.cs.tsukuba.ac.jp



(a) ノードの次数 (k) 分布　(b) ノードの重み (s) 分布

(c) エッジの重み (w) 分布

図 2: 次数・重みの分布

タに見られるタグの共起構造を Windowed Yule-Simon 過程により再現できているか，ネットワークの階層性に注目し分析を行う，1. ノードの次数 (k_i)・ノードの重み (s_i)・エッジの重み (w_{ij}) の分布，2. 次数-クラスタ係数相関 ($C^{(w)}(k)$)，3. 次数-次数相関 ($k_{nn}^{(w)}(k)$)，以上の 3 つの分析を行うことで確かめる．

1. ノードの次数・ノードの重み・エッジの重みの分布の結果を図 2 に示す．実データとモデルのそれぞれの分布はべき分布を示し，その指数も近い値であることがわかる．

2. 次数-クラスタ係数相関は重み無しの場合 [Ravasz 03] と重みありの場合 [Barrat 04] で計算を行った．その結果を図 3 に示す．重みありの場合も無しの場合も実データとモデルは同様の傾向を示すことがわかる．重み無しの次数-クラスタ係数相関に負の相関が現れるネットワークには階層性が存在することが示唆されており，モデルに負の相関が現れたことからモデルは階層性を生み出す可能性があることがわかる．

3. 次数-次数相関も同様に重み無しの場合 [Pastor-Satorras 01] とありの場合 [Barrat 04] で計算を行った．その結果を図 4 に示す．重みありの場合も無しの場合も実データとモデルは同様の傾向を示すことがわかる．

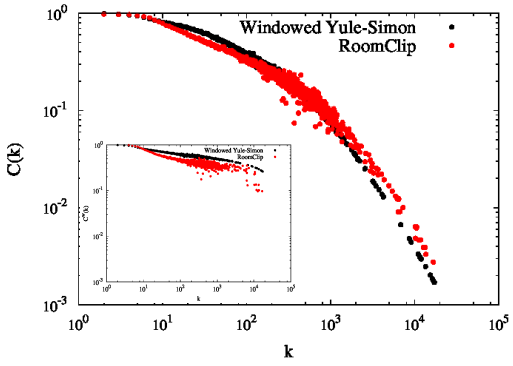


図 3: 次数-クラスタ係数相関．次数 k とクラスタ係数の関係を重み無し ($C(k)$) と重みあり ($C^w(k)$) の場合で示す．

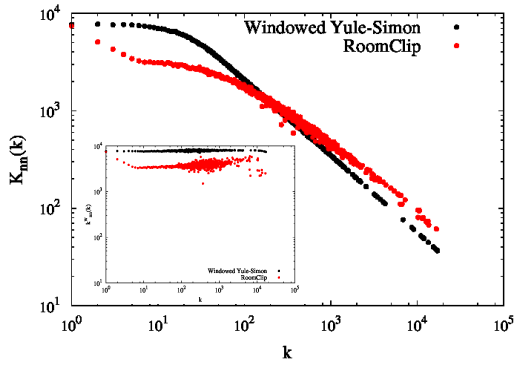


図 4: 次数-次数相関．次数 (k) とその隣接ノードの平均次数の関係を重みなし ($k_{nn}(k)$) の場合と，重みあり ($k_{nn}^w(k)$) の場合で示す．

4. まとめ

本論文ではタグの共起構造に注目し，Yule-Simon 過程を拡張した Windowed Yule-Simon 過程を提案した．タグの共起構造におけるモデルの検証として，実際のサービスに付与されるタグとモデルからタグ共起ネットワークを作成し，分析を行った．Windowed Yule-Simon 過程のタグ共起ネットワークは実データで観測されるものと様々な面で近い傾向を示した．それは次数分布，リンクの重みの分布，ノードの重みの分布がベキ分布を示すという一次の統計量であったり，次数-クラスタ係数相関や次数-次数相関といった二次の統計量である．今回の分析から Yule-Simon 過程を拡張した今回の簡単な確率モデルによって，タグ共起ネットワークは階層性を持つ可能性を示すような統計量も再現可能であることがわかった．以上のことから，個別のタグの利用統計のみならず，階層性の観点で行ったタグの共起構造の統計量の分析結果も同様の傾向を示すことがわかり，Windowed Yule-Simon 過程のタグ共起構造における実データとの類似性が確かめられた．

参考文献

[Barrat 04] Barrat, A., Barthelemy, M., Pastor-Satorras, R., and Vespignani, A.: The architecture of complex weighted networks, *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 101, No. 11, pp. 3747–3752 (2004)

[Pastor-Satorras 01] Pastor-Satorras, R., Vázquez, A., and Vespignani, A.: Dynamical and correlation properties of the Internet, *Physical review letters*, Vol. 87, No. 25, p. 258701 (2001)

[Ravasz 03] Ravasz, E. and Barabási, A.-L.: Hierarchical organization in complex networks, *Physical Review E*, Vol. 67, No. 2, p. 026112 (2003)

[Simon 55] Simon, H. A.: On a Class of Skew Distribution Functions, *Biometrika*, Vol. 42, No. 3/4, pp. pp. 425–440 (1955)