確率ネットワークの時系列分析に基づくソーシャルタグの分類

山岸　祐己・斉藤　和巳

『経営情報イノベーション研究』
静岡県立大学・経営情報イノベーション研究科
第5巻（2016年10月）
（抜刷）

査読論文

確率ネットワークの時系列分析に基づくソーシャルタグの分類

山岸祐己（静岡県立大学大学院 経営情報イノベーション研究科 博士後期課程 3 年）
斉藤和巳（静岡県立大学大学院 経営情報イノベーション研究科 教授）

ソーシャルタグの機能やダイナミクスを分析し，タグを分類する手法を提案する．提案手法は，確率ネットワークにおけるノードの人気度，入次数確率，PageRank を用いた分析に基づいており，タグ間の類似度を確率として正規化することによってネットワークを構築することを特徴としている．本手法は完全ネットワークを扱うが，多様な状況下でも高速に PageRank を求めることが可能であることを示す．タグの分類においては，貪欲法と局所改善に基づいたクラスタリング手法を用いる．

キーワード 確率ネットワーク，ソーシャルタグ，PageRank，k-medoids．

1 はじめに

ソーシャルメディアの発展により，近年，様々な Web オブジェクトに対してソーシャルタグが付与されるようになった．Web 上のソーシャルタグというものは，ユーザの個々の意思決定に基いて常に生成され続けているため，各ソーシャルタグの役割や機能が，時間とともに変化することは自然と推察される．よって我々は，巨視的分析の見地として，ソーシャルタグ間の確率ネットワークを構築し，それを分析することでソーシャルタグに関する規則性や重要性を見出すことを試みる．一般に，様々な側面から重要ノード群を発見することは，ネットワーク分析において基礎的な問題とされている [1]．ソーシャルネットワーク分析の分野では，次数中心性，近接中心性，媒介中心性といった，中心性の指標が幅広く研究されており [1]，一方，Web 情報検索の分野では，PageRank [2] と HITS [3] によるノードランキングが広く認識されている．これらの手法の中でも，次数に関する指標と PageRank は，今回の確率ネットワークに適用することができるため，これら 2 指標を実験で用いる．より詳細には，まず，ソーシャルタグ間の類似度を確率として正規化し，それらを確率有向リンクとしてソーシャルタグ間の確率ネットワークを構築する方法を提案する．次に，それら条件付き確率に基づいた各ノードの PageRank 値を高速に求める計算法を提案し，それを用いたソーシャルタグの分類手法を提案する．

提案分析手法の評価には，ニコニコ動画 [*1] のデータセットから生成した，タグ共起データを用いる．かねてより，大規模なデータを利用した複雑ネットワークの構造や機能に関する研究は，社会学，生物学，物理学，コンピュータ科学等の様々な分野で注目されている [4]．特に，これらのネットワークにおけるスケールフリー性は幅広く研究されており [5][6]，次数相関 [7] 等のより複雑な特徴が提案されてきた．本論文においても，ネットワークが持つとされるこれらの特徴に着目し，タグのスケールフリー性と次数相関を調べる．

本論文の構成は以下の通りである．まず，提案ネットワーク生成法，提案ノード指標，提案

[*1] www.nicovideo.jp

分類手法について説明する．そして，ニコニコ動画から取得したデータセットの調査結果を示し，提案手法による実験結果を述べる．最後に，今回得られた主要な結果と今後の展開についてまとめる．

2　提案分析手法

2.1　確率ネットワーク生成法と PageRank 値計算法

与えられたソーシャルタグ集合と，ニコニコ動画などタグが付与されるオブジェクト集合のそれぞれを自然数と同一視し，$\mathcal{M} = \{1, \cdots, m, \cdots, M\}$ と $\mathcal{H} = \{1, \cdots, h, \cdots, H\}$ で表す．ここで，$M = |\mathcal{M}|$ と $H = |\mathcal{H}|$ は総タグ数と総オブジェクト数である．また，時刻 t の時点でタグ m が付与されていたオブジェクト集合を $\mathcal{H}_m^{(t)} \subset \mathcal{H}$ とする．ただし，時刻 t も自然数と同一視して $t \in \{1, \cdots, T\}$ とし，T は最終観測時刻を表すとする．さらに，オブジェクト集合 $\mathcal{H}_m^{(t)}$ より，$h \in \mathcal{H}_m^{(t)}$ ならば $x_{m,h}^{(t)} = 1$ とし，さもなければ $x_{m,h}^{(t)} = 0$ として，各タグ m に対して H-次元縦ベクトル $\mathbf{x}_m^{(t)}$ を定義する．ただし，時刻の指定が不要な場合には，$\mathbf{x}_m^{(t)}$ を $\mathbf{x}_m$ と略記する．以降では，ベクトルとして定義するものは全て縦ベクトルとする．

いま，任意の時刻 t におけるソーシャルタグ間の $M \times M$ 類似度行列 $\mathbf{S}^{(t)}$ （時刻の指定が不要な場合には $\mathbf{S}$ ）をコサイン類似度に基づき定義する．すなわち，任意のタグのペア $n, m \in \mathcal{M}$ に対し，$\mathbf{x}_m \neq \mathbf{0}_H$ かつ $\mathbf{x}_n \neq \mathbf{0}_H$ ならば，$S(m,n) = (\mathbf{x}_m^T \mathbf{x}_n)/(||\mathbf{x}_m|| ||\mathbf{x}_n||)$ とし，さもなければ $S(m,n) = 0$ とする．ただし，$\mathbf{0}_H$ は H-次元の 0 ベクトルであり，$\mathbf{x}_m^T$ はベクトル $\mathbf{x}_m$ の転置を表し，$||\mathbf{x}_m||$ は $\sqrt{\mathbf{x}_m^T \mathbf{x}_m}$ で定義されるベクトル $\mathbf{x}_m$ のノルムである．次に，この類似度行列 $\mathbf{S}$ を正規化して，任意の時刻 t における $M \times M$ 推移確率行列 $\mathbf{P}^{(t)}$ （時刻の指定が不要な場合には $\mathbf{P}$ ）を構成する．ここで，類似度行列 $\mathbf{S}$ の要素 $S(m,n)$ を用いて，第 m 行の和を $S(m) = S(m,1) + \cdots + S(m,M)$ で定義する．推移確率行列 $\mathbf{P}$ の要素 $P(m,n)$ は，$S(m) \neq 0$ かつ $m \neq n$ ならば，$P(m,n) = S(m,n)/S(m)$ とし，$S(m) \neq 0$ かつ $m = n$ ならば，$P(m,n) = 0$ とし，そして $S(m) = 0$ ならば $P(m,n) = 1/M$ とする．すなわち，推移確率行列 $\mathbf{P}$ は，自己リンクなしで，タグ m から類似が高いタグに高い確率で推移し，$\mathbf{x}_m = \mathbf{0}_H$ のときなど任意のタグとの類似が 0 の場合は任意のタグへのランダムな推移となる．

推移確率行列 $\mathbf{P}$ に対し，一様ジャンプ確率 $\alpha \in (0,1)$ を用いて，任意の時刻 t における Google 行列 $\mathbf{G}^{(t)}$ （時刻の指定が不要な場合には $\mathbf{G}$）を $\mathbf{G} = (1-\alpha)\mathbf{P} + \alpha \mathbf{1}_M \mathbf{1}_M^T$ で定義すれば，ソーシャルタグ間の確率ネットワークにおいて，各タグの PageRank 値を求めることができる．ここで，$\mathbf{1}_M$ は任意の要素値が 1 の M-次元ベクトルを表す．しかしながら，任意の時刻 t で Google 行列 $\mathbf{G}^{(t)}$ を求めて，各タグの PageRank 値の時系列を求めるとすれば，ある時刻 t の類似度行列 $\mathbf{S}$ を求めるのに $O(M^2 H)$ の計算量が必要となり，それを時刻 $t = 1$ から最終観測時刻 T まで求めるため，全体で $O(TM^2 H)$ の計算量が必要となる．よって，この計算量では大規模データへの適用は困難な場合も起こる．

以下では，確率ネットワークでの PageRank 値を高速に求める計算法を提案する．いま，$S(m) \neq 0$ となるタグ数が $N(\leq M)$ のとき，$m \leq N$ ならば $S(m) \neq 0$ となり，$m > N$ ならば $S(m) = 0$ となるように，$\mathcal{M}$ の要素を並び替えても一般性を失わない．また，$S(m) \neq 0$ となるタグに対し，これらベクトル $\mathbf{x}_m$ を並べて構成する $H \times N$ 行列を $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_N)$ とし，これらの正規化値 $S(m)$ を要素とする $N \times N$ 対角行列を $\mathbf{\Delta} = diag(S(1), \cdots, S(N))$ とし，そして，これらベクトルのノルム $||\mathbf{x}_m||$ を要素とする $N \times N$ 対角行列を $\mathbf{\Gamma} = diag(||\mathbf{x}_1||, \cdots, ||\mathbf{x}_N||)$ とする．このとき，推移確率行列 $\mathbf{P}'$ は以下のように

表せる．

$$\mathbf{P}'=\begin{pmatrix} \mathbf{\Delta}^{-1}\left(\mathbf{\Gamma}^{-1}\mathbf{X}^T\mathbf{X}\mathbf{\Gamma}^{-1}-\mathbf{I}_{N,N}\right), & \mathbf{0}_{N,M-N} \\ \frac{1}{M}\mathbf{1}_{M-N}\mathbf{1}_M^T & \end{pmatrix}.$$

ここで，$\mathbf{I}_{N,N}$ は $N\times N$ 単位行列であり，$\mathbf{0}_{N,M-N}$ は全ての要素が 0 の $N\times(M-N)$ 行列を表す．

各タグの PageRank 値ベクトルを $\mathbf{y}^T=(\mathbf{u}^T,\mathbf{v}^T)$ とする．ここで，$\mathbf{u}$ は $S(m)\neq 0$ となるタグに対する N-次元ベクトルであり，$\mathbf{v}$ は $S(m)=0$ となるタグに対する $(M-N)$-次元ベクトルである．よって，Google 行列 $\mathbf{G}$ の定義に従えば，ベクトル $\mathbf{y}$ から PageRank 更新式より求まる次のステップのベクトル $\bar{\mathbf{y}}$ は以下となる．

$$\bar{\mathbf{y}}^T=\left((1-\alpha)\mathbf{u}^T\mathbf{\Delta}^{-1}\left(\mathbf{\Gamma}^{-1}\mathbf{X}^T\mathbf{X}\mathbf{\Gamma}^{-1}-\mathbf{I}_{N,N}\right),\mathbf{0}_{M-N}\right) + \frac{\mathbf{v}^T\mathbf{1}_{M-N}+\alpha}{M}\mathbf{1}_M^T. \quad (1)$$

明かに，行列 $\mathbf{\Delta}^{-1}$ や $\mathbf{\Gamma}^{-1}$ との積は，それぞれ対角行列なので，$O(N)$ 回の乗算で求まる．一方，行列 $\mathbf{X}$ の任意の要素は 0 または 1 であり，1 の要素数は各オブジェクトに付与されたタグ数の合計で，$L=\mathbf{1}_H^T\mathbf{X}\mathbf{1}_M$ となることより，行列 $\mathbf{X}^T$ や $\mathbf{X}$ との積は高々 L 回の加算で求まる．したがって，式 1 の更新は，$O(N)$ 回の乗算と $2L$ 回の加算で実現できる．

以下に提案法のアルゴリズムを示す．

1. PageRank 値ベクトルを $\mathbf{y}=(1/\sqrt{M},\cdots 1/\sqrt{M})^T$ と初期化する；
2. 式 1 で PageRank 値ベクトルを $\bar{\mathbf{y}}$ を求める；
3. $\sum_{m\in\mathcal{M}}|y_m-\bar{y}_m|<\epsilon$ ならば $\bar{\mathbf{y}}$ を出力し終了する；
4. $\mathbf{y}\leftarrow\bar{\mathbf{y}}$ としステップ 2. へ戻る．

実験では，Google 行列を構成するための一様ジャンプ確率を $\alpha=0.15$ とし，終了条件を制御するパラメータを $\epsilon=10^{-8}$ に設定した．なお，推移確率行列 $\mathbf{P}$ を陽に求め，上記ステップ 2. を $\bar{\mathbf{y}}^T\leftarrow\mathbf{y}^T\mathbf{P}$ で求める方法をベースライン法と呼び，実験では，ベースライン法との比較により，提案法の性能を評価する．

2.2 k-medoids クラスタリング

k-medoids 法は，非階層クラスタリングで有名な k-means 法と同様に，H-次元の M 個のオブジェクト集合 $\mathcal{M}$ が与えられたとき，オブジェクト集合を K 個のクラスターに分割する問題を解くための手法である．任意のオブジェクトペア $u,v\in\mathcal{M}$ 間に類似度 $\rho(u,v)$ が定義されていれば，オブジェクト集合の中から他のオブジェクトとの類似度の和が高い代表オブジェクトを選定することが可能であるため，最適な代表オブジェクトが選定されれば，類似度の高いオブジェクトペアは同じクラスターに，類似度の低いオブジェクトペアは異なるクラスターに属するように分割されるはずである．このような問題では，一般的に平均（mean）より中央値（median）の方が頑健であることが知られている．ただし，大域最適解を求めるためには $O(M^KH)$ の計算量が必要であるため，オブジェクト集合の規模や次元，分割数 K がある程度大きくなると，実用的な時間で解を求めることが難しくなる．よって，k-medoids にも局所最適解を求めるための反復法や貪欲法が存在するが，今回は解の一意性が保証される貪欲法に基づく解法を採用する．この解法は，目的関数のサブモジュラ性により，厳密解ではないものの，ある程度妥当な精度で最悪ケースの解品質が理論的に保証されている [8]．貪欲法とは，既に選定した代表オブジェクトを固定し，目的関数値を最大にするオブジェクトを求め，目的関数が増加するならば代表オブジェクト集合に追加することで，結果の代表オブジェクト集合を求める方法である．各オブジェクトは，最も類似度の高い代表オブジェクトと同じクラスターに割り当てられる．既に選定した代表オブジェクト集合を $\mathcal{P}$ とし，新たに追加を試みるオブジェクトを w とす

るとき，ここでは，以下の目的関数を考える．

$$f(\mathcal{P}\cup\{w\})=\sum_{v\in\mathcal{M}}\max\{\mu(v\,;\mathcal{P}),\rho(v,w)\}. \tag{2}$$

ここで，$\mu(v\,;\mathcal{P})$ は既に選定された代表オブジェクトとの類似度の最大値を表し，$\mu(v\,;\mathcal{P})=\max_{w\in\mathcal{P}}\{\rho(v,w)\}$ で定義される．以下に k-medoids における貪欲アルゴリズムを説明する．ここで，$\cdot\backslash\cdot$ は集合差を表す．

A1-1. $k\leftarrow 1,\mathcal{P}_0\leftarrow\emptyset$，各オブジェクト $v\in\mathcal{M}$ に対し，$\mu(v\,;\emptyset)\leftarrow 0$ と初期化する；

A1-2. 式 2 で $\hat{p}_k=\underset{w\in\mathcal{M}\backslash\mathcal{P}_{k-1}}{\arg\max}\{f(\mathcal{P}_{k-1}\cup\{w\})\}$ を求め，$\mathcal{P}_k\leftarrow\mathcal{P}_{k-1}\cup\{\hat{p}_k\}$ とする；

A1-3. $k=K$ ならば $\hat{\mathcal{P}}_K=\{\hat{p}_1,\ldots,\hat{p}_K\}$ を出力し終了する；

A1-4. 各オブジェクト $v\in\mathcal{M}$ に対し，$\mu(v\,;\mathcal{P}_k)$ を求める；

A1-5. $k\leftarrow k+1$ とし，ステップ A1-2. へ戻る．

各オブジェクトを，最も類似度の高い代表オブジェクト $p_k\in\mathcal{P}$ のクラスター $\mathcal{C}_k$ に割り当てる．

明らかに，上記のアルゴリズムの計算量は $O(M^2KH)$ となるため，大域最適解を得るために必要な計算量 $O(M^KH)$ に比べて非常に高速である．しかし，貪欲法に基づく単純な手法であるため，比較的アーな局所解にトラップされる危険性が伴う．よって，ここからは貪欲アルゴリズムで得た $\hat{\mathcal{P}}_K$ の解品質を向上させるための局所改善アルゴリズムについて述べる．

A2-1. $k\leftarrow 1,h\leftarrow 0$ と初期化する；

A2-2. 式 2 で $p'_k=\underset{w\in\mathcal{M}\backslash\hat{\mathcal{P}}_K\backslash\{\hat{p}_k\}}{\arg\max}\{f(\hat{\mathcal{P}}_K\backslash\{\hat{p}_k\}\cup\{w\})\}$ を求める；

A2-3. $p'_k=\hat{p}_k$ ならば $h\leftarrow h+1$ とし，さもなければ $h\leftarrow 0,\hat{\mathcal{P}}_K=\hat{\mathcal{P}}_K\backslash\{\hat{p}_k\}\cup\{p'_k\}$ とする；

A2-4. $h=K$ ならば $\hat{\mathcal{P}}_K$ を出力して終了する；

A2-5. 各オブジェクト $v\in\mathcal{M}$ に対し，$\mu(v\,;\hat{\mathcal{P}}_K)$ を求め，$k=K$ ならば $k\leftarrow 1$，さもなければ $k\leftarrow k+1$ とし，ステップ A2-2 へ戻る．

貪欲アルゴリズムの後にこのアルゴリズムを使用すると，明らかに貪欲アルゴリズムだけのときよりも計算量を必要とするが，高々 $O(M^2KH)$ の数倍程度で計算が終わることを我々は既に実験によって示している [9].

先にも述べたように，このクラスタリング手法には解の一意性が保証されているため，基本的にはこの計算処理は一度しか行われない．よって，多少の計算量の増加を負ってでも，一度で解精度が高い結果を出すために，ここでは貪欲アルゴリズムと局所改善アルゴリズムを反復して使用する手法を述べる．

I1. A1-1 から処理を開始する；

I2. A1-4 の処理後に $k>1$ ならば $\mathcal{P}_k$ を $\hat{\mathcal{P}}_K$ として出力する；

I3. $\hat{\mathcal{P}}_K$ を A2 で改善し，改善した $\hat{\mathcal{P}}_K$ を $\mathcal{P}_k$ として出力する；

I4. A1-5 から処理を再開させ，ステップ I2 へ戻る；

この手法は，逐次的に貪欲アルゴリズムと局所改善アルゴリズムを使用するときよりも計算量が増加するが，解の精度は安定して良い結果が得られることを我々は既に実験により示している [9].

*2 1 つの動画につき 11 個まで登録できる関連文字列

3 データセット

今回の実験で用いるデータセットは，ニコニコ動画における VOCALOID オリジナル楽曲動画のタグ [*2]の時系列データである．このデータセットは，VOCALOID オリジナル楽曲動画が一般に有するタグによる検索結果から，二次創作系や加工系のタグを有する動画を除外して取得したものであり，取得期間は 2013/04/03 から 2016/03/16（24 時間毎 1078 日間 $T = 1078$），最終日の動画数 H は 136192，タグ数 M は 125196 である．今回，取得期間中に一度でも 10 以上の動画に登録されていたタグを分析対象としたため，最終的に使用したタグの種類 M は 7787 である．

ここでは，データセットに対して行った統計的処理の結果について述べる．図 1，2 は，それぞれ取得期間中の動画数と分析対象タグ数の推移であり，どちらも一定のペースで増加し続けていることが分かる．これにより，時間の経過によるタグネットワークの構造変化が期待できる．図 3 は，最終観測時刻における全てのタグを用いたときのタグの度数分布である．多くの Web オブジェクトと同様のように，ニコニコ動画のタグの使用頻度にもスケールフリー性が見られる．

4 評価実験

今回の実験結果について述べる．まず，提案手法とベースライン法による PageRank 値の計算時間の比較を図 4 に示す．図より，ベースライン法は使用されたタグ数の増加と共に計算時間が増加し続けているが，提案手法はどのようなタグ数に対しても計算時間が 1 秒未満となっているため，提案手法が圧倒的に高速であることは明らかである．次に，タグ確率ネットワークの期待される性質を調べるため，分析対象タグの共起関係を次数相関に見立ててプロットしたものを図 5 に示す．図の横軸はタグの出現数を，縦軸は共起しているタグの出現数の平均を表している．この分布は，代表的なネットワークモデルの一つである BA モデル [10] の次数相関によく似ている．すなわち，生成されるネットワークの特徴としては，次数が大きい（出現数と共起数が共に多い）ノードがリンク（他ノードからの確率）を得やすいことが期待される．

ここからは，提案 k-medoids 法によるタグのクラスタリング結果について述べる．今回のデータセットにおける PageRank 値の時系列データを *prk*，すなわち

$$prk = \left(\ \bar{\mathbf{y}}^{(1)}, \ \cdots, \ \bar{\mathbf{y}}^{(T)} \ \right),$$

（$\bar{\mathbf{y}}^{(t)}$ は任意の時刻 t において求められた PageRank 値ベクトル）とし，推移確率行列 $\mathbf{P}$ における各ノードの入次数確率合計の時系列データを *ind*，すなわち

$$ind = \begin{pmatrix} \sum_{m=1}^{M} P^{(1)}(m,1) & \cdots & \sum_{m=1}^{M} P^{(T)}(m,1) \\ \vdots & \ddots & \vdots \\ \sum_{m=1}^{M} P^{(1)}(m,M) & \cdots & \sum_{m=1}^{M} P^{(T)}(m,M) \end{pmatrix},$$

タグの出現数の時系列データを *pop*，すなわち

$$pop = \begin{pmatrix} \sum_{h=1}^{H} \mathbf{x}_1^{(1)} & \cdots & \sum_{h=1}^{H} \mathbf{x}_1^{(T)} \\ \vdots & \ddots & \vdots \\ \sum_{h=1}^{H} \mathbf{x}_M^{(1)} & \cdots & \sum_{h=1}^{H} \mathbf{x}_M^{(T)} \end{pmatrix},$$

として比較に用いる．なお，時系列データの類似度 $\rho(u,v)$ は相関係数とした．図 6 にクラスター数 K と解品質の関係を示す．図の縦軸は，各クラスターの代表オブジェクトとの類似度の最大値の総和 $\sum_{v \in \mathcal{M}} \mu(v \, ; \, \hat{\mathcal{P}}_K)$ を表している．図より，出現数 *pop* に基づくクラスタリングにおけるクラスターの解品質が高く，入次数確率 *ind* に基づくクラスタリングにおけるクラスターの解品質が低いことがわかる．また，K が 20 を超えたあたりでどの指標においても解品質の改善が鈍くなるなるため，以下，分析用のクラスター数は $K = 20$ とした．図 7 は，$K = 20$

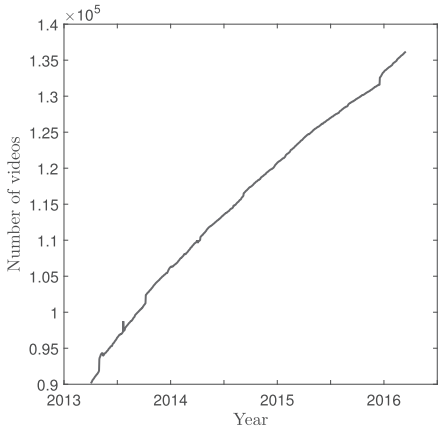

図 1　動画数の推移

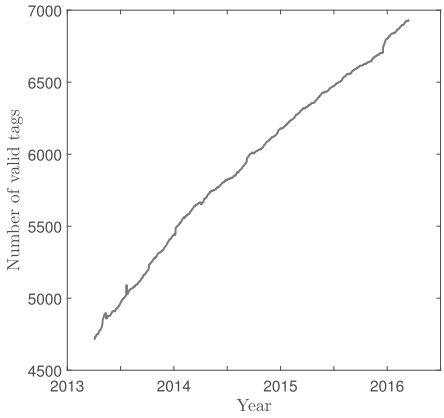

図 2　対象タグ数の推移

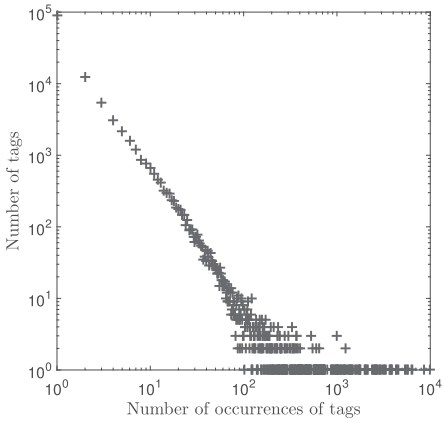

図 3　タグの度数分布

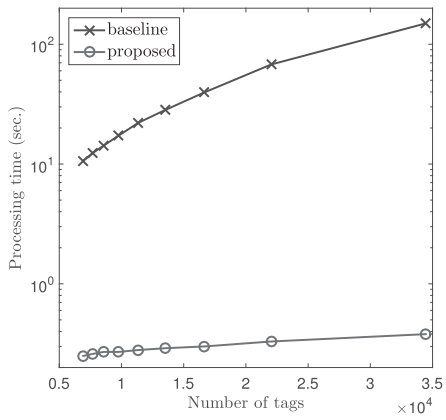

図 4　PageRank 値の計算時間比較

のときの各クラスターの代表オブジェクトとの類似度の平均を示したものである．図の横軸は類似度の平均が高い順で降順ソートしたときのランクである．クラスター内の類似度平均が最も高いクラスターは *prk* 内に存在しているが，全体を見るとやはり *pop* のクラスター品質が高い傾向にある．詳細な検証として，*pop* と *prk* のクラスター品質上位 2 クラスターを比較する．まず，クラスター品質 1 位の比較を表 1,2 に示す．どちらも似たようなタグが並んでいるため，大きな差異が無いように思えるが，タグの種別頻度で見てみると，*pop* は ‘category’ 3, ‘software’ 13, ‘genre’ 3, ‘information’ 1, となっており，*prk* は ‘category’ 2, ‘software’ 15, ‘genre’ 3, となっているため，多少なり *prk* の方がタグの種別のばらつきが小さいことが分かる．続いて，クラスター品質 2 位の比較を表 3,4 に示す．ここでも似たようなタグが並んでおり，タグの種

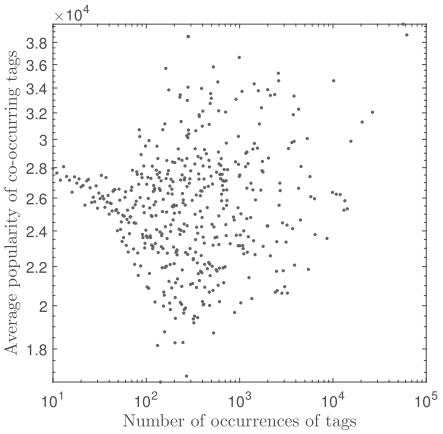

図 5　タグの共起関係に基づく次数相関

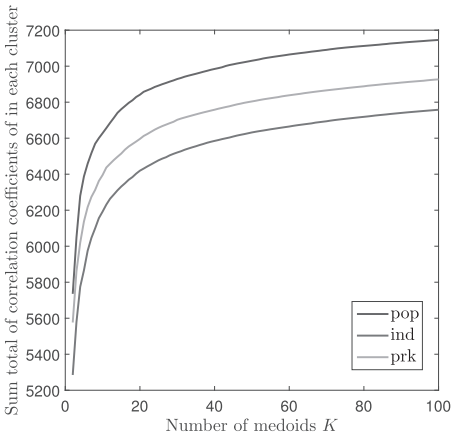

図 6　クラスター数 K と解品質の関係

表 1　*pop* の品質 1 位のクラスター内における出現数上位 20 タグ

	pop	
rank	type	tag
1	category	vocaloid
2	software	初音ミク
3	software	ミクオリジナル曲
4	genre	vocarock
5	software	鏡音リン
6	software	gumi
7	software	gumi オリジナル曲
8	software	巡音ルカ
9	software	リンオリジナル曲
10	information	vocaloid 処女作
11	software	ルカオリジナル曲
12	software	鏡音レン
13	software	レンオリジナル曲
14	software	ia -aria on the planetes-
15	software	kaito
16	software	ia オリジナル曲
17	genre	ボカロクラシカ
18	category	vocaloid オリジナル曲
19	category	vocaloid-pv
20	genre	ボカロバラード

表 2　*prk* の品質 1 位のクラスター内における出現数上位 20 タグ

	prk	
rank	type	tag
1	category	vocaloid
2	software	初音ミク
3	software	ミクオリジナル曲
4	category	音楽
5	genre	vocarock
6	software	鏡音リン
7	software	gumi
8	software	巡音ルカ
9	software	リンオリジナル曲
10	software	ルカオリジナル曲
11	software	鏡音レン
12	software	レンオリジナル曲
13	software	kaito
14	software	ニコニコムービーメーカー
15	genre	ボカロクラシカ
16	software	メグッポイド
17	software	vocaloid3
18	software	meiko
19	software	mikumikudance
20	genre	爽やかなミクうた

別も ‘genre’, ‘artist’, ‘dancer’, が主であるため，目立った差異が無いように見える．恐らく，どちらも Dubstep という音楽ジャンル周辺で見られるタグと思われるが，もし Dubstep が基軸となったクラスターであるなら，*pop* のラインナップはそこまで妥当とは言えない．現に，*pop* の表においては，上位 2 つを除いて Dubstep に関連しているタグは 5 位の ‘skrillex’(Dubstep artist)

表 3　*pop* の品質 2 位のクラスター内における出現数上位 20 タグ

	pop	
rank	type	tag
1	genre	dubstep
2	genre	ダブステップ
3	genre	electro
4	genre	洋楽
5	artist	skrillex
6	genre	アニメ色のない作業用 bgm
7	dancer	remotekontrol
8	genre	street dance 統一タグ
9	genre	amv
10	artist	uk
11	genre	dj
12	genre	nonstop
13	dancer	marquese scott
14	dancer	左のおっさん
15	genre	フリームーブ
16	genre	東方 dubstep
17	genre	dnb
18	genre	アングラアニソン remix リンク
19	genre	アニメーションダンス
20	dancer	chibi

表 4　*prk* の品質 2 位のクラスター内における出現数上位 20 タグ

	prk	
rank	type	tag
1	genre	dubstep
2	genre	ダブステップ
3	genre	洋楽
4	artist	skrillex
5	artist	remotekontrol
6	artist	uk
7	genre	nonstop
8	dancer	marquese scott
9	dancer	左のおっさん
10	genre	東方 dubstep
11	genre	アングラアニソン remix リンク
12	dancer	chibi
13	brand	ukf
14	artist	klaypex
15	artist	nero
16	genre	ブロステップ
17	dancer	bryan gaynor
18	artist	skream
19	artist	knife party
20	genre	脱糞ステップ

と 16 位の ‘東方 dubstep’(Dubstep genre) の 2 つだけである．それに対し，*prk* の表においては，4 位の ‘skrillex’(Dubstep artist), 10 位の ‘東方 dubstep’(Dubstep genre), 13 位の ‘ukf’(Dubstep artists bland), 14 位の ‘klaypex’(Dubstep artist), 15 位の ‘nero’(Dubstep artist), 16 位の ’ ブロステップ ‘(Dubstep genre), 18 位の ‘skream’(Dubstep artist), 19 位の ‘knife party’(Dubstep artist), 20 位の ‘脱糞ステップ’(Dubstep genre) の 9 個のタグが Dubstep と深く関わっている．

5　おわりに

巨視的分析の見地として，ソーシャルタグ間の確率ネットワークを構築し，それを分析することでソーシャルタグに関する規則性や重要性を見出すことを試みた．今回提案した確率ネットワーク生成法と PageRank 値計算法は，ベースライン手法と比較して圧倒的に高速に PageRank 値を算出することができるため，多数の観測時刻におけるデータを PageRank 時系列データとして容易に扱えることを示した．また，提案した k-medoids アルゴリズムによる PageRank 時系列データのクラスタリングでは，比較的解釈しやすい出力が見られたため，ソーシャルタグの分析手法としての有用性が期待できると言える．しかし，図 8 に示すように，提案 k-medoids アルゴリズムはクラスター数 K の増加による計算時間の増加が著しいため，今後はクラスタリングアルゴリズムの高速化を考案するつもりである．

謝辞

本研究は，JSPS 特別研究員奨励費 15K00311 の支援を受けて行ったものである．

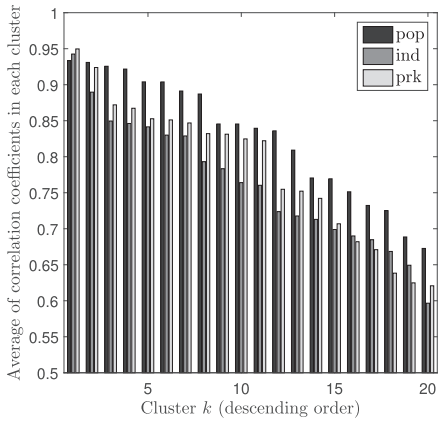

図 7　クラスター品質の検証

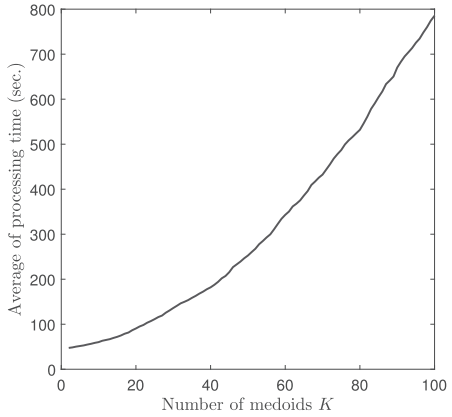

図 8　貪欲アルゴリズムと局所改善アルゴリズムの反復による計算量の増加（3 指標による平均）

参考文献

[1] Wasserman, S. and Faust, K.: *Social network analysis*, Cambridge University Press, Cambridge, UK (1994).

[2] Brin, S. and Page, L.: The anatomy of a large-scale hypertextual web search engine, *Computer Networks and ISDN Systems*, Vol. 30, pp. 107–117 (1998).

[3] Kleinberg, J.: Authoritative sources in a hyperlinked environment, *Journal of the ACM*, Vol. 46, No. 5, pp. 604–632 (1999).

[4] Newman, M.: The structure and function of complex networks, *SIAM Review*, Vol. 45, pp. 167–256 (2003).

[5] Song, C., Koren, T., Wang, P. and Barabási, A.-L.: Modelling the scaling properties of human mobility, *Nature Physics*, Vol. 6, pp. 818–823 (2010).

[6] Easley, D. and Kleinberg, J.: *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*, Cambridge University Press, New York, NY, USA (2010).

[7] Vázquez, A.: Growing network with local rules: Preferential attachment, clustering hierarchy, and degree correlations, *Physical Review*, Vol. 67, No. 5, p. 056104 (2003).

[8] Nemhauser, G. L., Wolsey, L. A. and Fisher, M. L.: An analysis of approximations for maximizing submodular set functions, *Mathematical Programming*, Vol. 14, pp. 265–294 (1978).

[9] Yamagishi, Y., Okubo, S., Saito, K., Ohara, K., Kimura, M. and Motoda, H.: A Method to Divide Stream Data of Scores over Review Sites, *Proceedings of the 13th Pacific Rim International Conference on Artificial Intelligence (PRICAI 2014)*, pp. 913–919 (2014).

[10] Barabási, A.-L. and Albert, R.: Emergence of scaling in random networks, *Science*, Vol. 286, pp. 509–512 (1999).

THE CONTENTS OF THIS ISSUE IS SUMMARIZED IN ENGLISH BELOW

A Classification of Social Tags Based on Time Series Analysis of Stochastic Network

Yuki YAMAGISHI
Graduate School of Management and Information of Innovation, University of Shizuoka
Kazumi SAITO
Graduate School of Management and Information of Innovation, University of Shizuoka

We focus on the functions or dynamics of social tagging and propose a method for classification of social tags. Our method is based on an analytics about node popularity, in-degree probability, and PageRank in stochastic networks. In generating the stochastic networks, we define a similarity between social tags based on their co-occurrence vectors and normalize them to probabilities in order to appropriately treat under the stochastic networks. Although we handling complete graphs in experiments, our computational method can obtain the PageRank values in a short time because the transition probability matrix is converted to a convenience shape which needs minimum necessary calculations. We use k-medoids clustering in the social tag classification and employed an algorithm constructed by a combination of a greedy search and a local search. In experimental results, we show an availability about tag classification using time series PageRank values and our k-medoids algorithm.