人工知能学会研究会資料
SIG-KBS-B501-03

トピックを考慮した情報拡散現象のモデル化について

On a Topic-based Information Diffusion Model

大原剛三 [1*]　小田切亮祐 [1]　白川真一 [2]
Kouzou Ohara[1], Ryosuke Odagiri[1], Shinichi Shirakawa[2]

[1] 青山学院大学理工学部
[1] College of Science and Engineering, Aoyama Gakuin University
[2] 筑波大学システム情報系
[2] Faculty of Engineering, Information and Systems, University of Tsukuba

Abstract: Information diffusion over a social network can be modeled as stochastic processes of state changes. In this paper, we propose an information diffusion model that takes into account topics of information. More specifically, the proposed model determines the diffusion probability for a directed link by using the content attribute of a target document that will propagate over the link, which represents the topic distribution in the document, and the link attribute that expresses the topic distribution in documents that have propagated over the link. The number of model parameters to be learned is only twice of the number of topics considered, which is much less than the one for traditional models, and they can be efficiently and accurately learned from observed diffusion sequences based on the framework of the maximum likelihood estimation. Through an experiment using real world retweet sequences, we confirmed that the proposed model allows us to estimate the length of an information diffusion sequence more accurately compared to an existing model that do not consider topics at all.

1　はじめに

近年，スマートフォンなどの携帯情報端末の高機能化，高性能化と，社会におけるインターネット接続環境の急速な普及により，Facebook や Twitter などの様々なソーシャルメディアを手軽に利用することが可能となっている．それらのソーシャルメディア上に展開される大規模な社会ネットワークを通して，多様な情報が短時間，かつ広範囲に拡散されており，我々の日常における意思決定にも多大な影響を与えるに至っている．そのような背景の下，従来の社会学のみならず，数多くの分野において社会ネットワークの分析が進められている [1, 2].

これらの研究では，情報拡散現象をモデル化するために独立カスケード（IC: Independent Cascade）モデルや線形閾値（LT: Linear Threshold）モデル [3] などの基本的な確率モデルが用いられることが多い．一方，IC モデルや LT モデルは，実際の情報拡散現象を再現するには必ずしも十分とは言えないため，これまでに幾つかの拡張が提案されている [4, 5, 6, 7].

Saito らは，離散時間間隔でノードの状態変化が同期して起こることを前提としている IC モデルを，実時間間隔でノードの状態が非同期に変化し得る AsIC モデル [5] を提案し，さらにノードに事前に付与された属性に基づき 2 ノード間の情報伝播確率が決まるノード属性を考慮した情報拡散モデルに拡張している [4]．このモデルでは，ノード属性ベクトルと同じ次元数の重みベクトルのみを過去の情報拡散系列を用いて学習するため，リンク個々に対する情報拡散確率を個別に学習する従来の枠組みと比べると，比較的少ない情報拡散系列からでも精度よくその値を学習することができるという特徴がある．

一方，類似したアプローチとして，ノードの属性ではなく，拡散する情報のトピックに基づいてリンクに対する拡散確率を決定するモデルも近年，提案されている [6, 7]．Barbieri らの AIR モデル [6] は，ユーザ u がトピック z に対してもつ影響度（Authoritativeness）と興味度（Interest），および情報 i に対するトピック分布（Relevance）に基づき拡散確率を決定する．また，Chen らのモデルは，IC モデルを基礎とし，各リンクにおけるトピック分布と拡散する情報に対するトピック分布からリンクごとの拡散確率を求めるものである [7]．ただし，Barbieri らのモデルはリンク数に比例するだけの数のパラメータを学習する必要があるため，過学習を回避するには一定量の観測データが必要となる．また，Chen らのモデルでは過去の拡散情報からトピックモデ

*連絡先：青山学院大学理工学部情報テクノロジー学科
〒 252-5258 相模原市中央区淵野辺 5-10-1
E-mail: ohara@it.aoyama.ac.jp

ルを学習し，そのトピックモデルに基づいて各リンク，および新規に流れる情報に対するトピック分布を求めているが，拡散確率の計算には過去の情報拡散系列の情報は反映されない．そのため，得られる拡散確率が実際の拡散現象を精度よく近似できる保証はない．

以上のような背景の下，本研究では Saito らのモデルを基礎とし，ネットワーク上を流れる情報のトピックを考慮した情報拡散モデルを提案する．基本的な考え方は，Chen らのモデルと同様に，新規の情報に対するトピック分布（コンテンツ属性）とリンクごとのトピック分布（リンク属性）から情報拡散確率を計算するというものである．ただし，Chen らのモデルとは異なり，2 つの属性を結合する際に，Saito らのモデルと同様に重みベクトルを導入し，その値はモデルパラメータとして観測した拡散系列から学習する．Twitter 上で観測された実際の情報拡散系列を用いた実験では，トピックを考慮した場合と，そうでない場合とで情報拡散シミュレーションの結果を比較し，提案モデルの有用性を示す．

2 ノード属性を考慮した情報拡散モデル

以下では，まず基本となる AsIC モデル [5] について概説し，その後，文献 [4] に従いノード属性を考慮した情報拡散モデルについて説明する．

2.1 AsIC モデル

AsIC モデルは，時間遅れを考慮していない従来の IC モデルに対して，非同期実時間遅れの概念を導入したものである．なお，本稿では議論の簡単化のために遅延時間は指数分布に従うものとするが，べき乗分布など他の分布も同様に用いることが可能である．

以下では，社会ネットワークを有向グラフ $G = (V, E)$ で表す．ここで，V と E（$\subset V \times V$）はそれぞれ全ノードの集合と全リンクの集合を表す．また，任意のノード $v \in V$ について，v を始点とするリンクの終点ノードの集合を $F(v) = \{u \in V;\ (v, u) \in E\}$，$v$ を終点とするリンクの始点ノードの集合を $B(v) = \{u \in V;\ (u, v) \in E\}$ とする．このとき，ネットワーク G 上の情報拡散において，あるノードが対象情報の影響を受けた場合，そのノードはアクティブであると呼び，各ノードはアクティブと非アクティブという二つの状態のいずれか一方の状態を取るものとする．また，ノードは非アクティブからアクティブに変化するが，その逆には変化しないと仮定する．

AsIC モデルは，各リンク $(u, v) \in E$ に対する拡散確率 $p_{u,v}$（$0 < p_{u,v} < 1$）と，遅延パラメータ $r_{u,v}$（$r_{u,v} > 0$）という 2 つのパラメータをもつ．このとき，AsIC モデルにおける情報拡散過程は与えられた初期アクティブノード（情報源ノード）を起点とし，連続時間 t の下で次のように展開される．あるノード u が時刻 t でアクティブとなったとき，u はその時点で非アクティブであるノード $v \in F(u)$ をアクティブにする機会を 1 度だけ与えられる．このとき，遅延時間 δ が $r_{u,v}$ をパラメータとする指数分布に従い決められる．そして，u は時刻 $t + \delta$ までに v がアクティブになっていなければ v をアクティブにすることを試み，その試行は確率 $p_{u,v}$ で成功する．u の試行が成功した場合，v は時刻 $t + \delta$ にアクティブとなる．ノードをアクティブにする試行がそれ以上実行できなくなった時点で，この情報拡散過程は終了する．

2.2 ノード属性を考慮した AsIC モデル

ノード属性を考慮した AsIC モデルでは，各ノードは名義属性，もしくは数値属性のいずれかであるような属性を 1 つ以上もち，各ノードのもつ属性値に応じて拡散確率，ならびに遅延パラメータの値が決定される．いま，すべてのノードが J 個の属性をもち，ノード v が j 番目の属性に対して取る値を v_j とする．ここで，各リンク $(u, v) \in E$ に対して，各要素が u_j と v_j に対する関数値 $x_{u,v,j} = f_j(u_j, v_j)$ となる J 次元ベクトル $\boldsymbol{x}_{u,v}$ を考える．実際には，パラメータ学習時の計算を簡単化するために，$x_{u,v,0} = 1$ を加えた $J+1$ 次元ベクトル $\boldsymbol{x}_{u,v}$ をリンク属性として考える．このとき，リンク $(u, v) \in E$ に対する拡散確率 $p_{u,v}$ と遅延パラメータ $r_{u,v}$ は次式により求められる．

$$p_{u,v} = p(\boldsymbol{x}_{u,v}, \boldsymbol{\varphi}) = \frac{1}{1 + \exp(-\boldsymbol{\varphi}^T \boldsymbol{x}_{u,v})} \tag{1}$$

$$r_{u,v} = r(\boldsymbol{x}_{u,v}, \boldsymbol{\phi}) = \exp(\boldsymbol{\phi}^T \boldsymbol{x}_{u,v}) \tag{2}$$

ここで，$\boldsymbol{\varphi}^T = (\varphi_0, \cdots, \varphi_J)$ と $\boldsymbol{\phi}^T = (\phi_0, \cdots, \phi_J)$ は，それぞれ拡散確率と遅延パラメータに対する $J+1$ 次元のパラメータベクトルであり，φ_0 と ϕ_0 はそれぞれ定数項であり，$\boldsymbol{\varphi}^T$ はベクトル $\boldsymbol{\varphi}$ の転置を表すものとする．すなわち，拡散確率と遅延パラメータはリンクごとに定まる $J+1$ 次元ベクトル $\boldsymbol{x}_{u,v}$ とそれぞれに対応するパラメータベクトルから導出されることになる．このパラメータベクトルの値は，観測した情報拡散系列から最尤推定を用いて学習する．

文献 [4] では，関数 f_j として，j 番目の属性が数値属性の場合は $x_{u,v,j} = \exp(-|u_j - v_j|)$ を，名義属性の場合は $x_{u,v,j} = \delta(u_j, v_j)$ を用いている．ここで，$\delta(u_j, v_j)$ は，$u_j = v_j$ のとき $\delta(u_j, v_j) = 1$ となり，それ以外の場合は $\delta(u_j, v_j) = 0$ となる関数である．直観的には，2 つのノード u，v が似ているほど，それらの j 番目の属性値 u_j は

v_j 近い値を取り，そのとき，対応するパラメータ値 θ_j が正の値であれば拡散確率 $p_{u,v}$ はより大きな値をとり，負であればより小さな値となる．

3 トピックを考慮した情報拡散のモデル化

前節で概説したモデルは，各ノードが J 個の属性をもつことを仮定していた．これに対し，ここでは，拡散対象となる各情報（文章）に対するトピック分布をコンテンツ属性，各リンクを流れる情報に対するトピック分布をリンク属性とし，両者によって拡散確率と遅延パラメータの値が決まる AsIC モデルの拡張を考える．

3.1 コンテンツ属性とリンク属性を考慮した情報拡散モデル

いま，K 個のトピック $z_1, \cdots, z_K$ を仮定したとき，ネットワークを伝播する情報（文書）d に対する k 番目のトピックの生起確率を z_k^d とする．このとき，z_k^d を k 番目の要素にもつ K 次元のベクトル $\boldsymbol{\theta}_d = (z_1^d, \cdots, z_K^d)$ を d に対するコンテンツ属性と呼ぶ．一方，リンク $e = (u, v) \in E$ に対して，ノード u からノード v に伝播したすべての情報に対するトピックの分布 $\boldsymbol{\theta}_e = (z_1^e, \cdots, z_K^e)$ を e のリンク属性と呼ぶ．ここで，$\boldsymbol{x}_{d,e}$ を各要素が z_k^d と z_k^e に対する関数値 $x_{d,e,k} = g_k(z_k^d, z_k^e)$ となる J 次元ベクトルとし，以下では，2.2 節と同様に，$x_{d,e,0} = 1$ を加えた $J+1$ 次元ベクトル $\boldsymbol{x}_{d,e}$ を考える．このように定義することで，リンク $(u, v) \in E$ に対する拡散確率 $p_{u,v}$ と遅延パラメータ $r_{u,v}$ を，それぞれ式 (1) と式 (2) を用いて定義することが可能となる．ただし，関数 g_k は $g_k(z_k^d, z_k^e) = \exp(-|z_k^d - z_k^e|)$ とする．直観的には，情報 d のトピック分布（コンテンツ属性）と，リンク $e = (u, v)$ をそれまでに流れた情報のトピック分布（リンク属性）が近いほど，拡散確率 $p_{u,v}$ は高い値を取ることになる．また，モデルパラメータ $\boldsymbol{\varphi}$，$\boldsymbol{\phi}$ は共に文献 [4] と同じ枠組みの下で観測した拡散系列データから学習可能である．

3.2 属性ベクトルの生成とパラメータの学習

前節で導入したコンテンツ属性とリンク属性は，それぞれ拡散対象となる情報，および各リンクをそれまでに流れた情報のトピック分布を表す．本稿では，これらの属性ベクトルの生成に LDA（Latent Dirichlet Allocation）[8] によるトピックモデルを用いる．具体的には，対象ネットワーク上で拡散が観測された N 個の情報（文書）の集合を $D = \{d_1, \cdots, d_N\}$ としたとき，各属性の生成の流れは以下のようになる．

1. D から LDA によりトピックモデル $\mathcal{M}$ を学習する．
2. リンク $e = (u, v)$ 上を伝播した文書集合 $\{d_1^e, \cdots, d_h^e\}$ を 1 つの文書と見なし，それに対するトピック分布（リンク属性）$\boldsymbol{\theta}_e$ を $\mathcal{M}$ を用いて求める．
3. 拡散対象となる新規文書 d に対するトピック分布（コンテンツ属性）$\boldsymbol{\theta}_d$ を $\mathcal{M}$ を用いて求める．

ここで，文献 [4] におけるパラメータ学習の枠組みでは，ノード属性は拡散系列に依存せずに一定であることに注意する．本稿で提案するトピックに基づく情報拡散モデルでは，拡散対象となる文書ごとにそのコンテンツ属性は異なるため，文献 [4] におけるパラメータ学習の枠組みをそのまま用いることはできない．そのため，ここでは特定の対象に関する文書は同一のトピック分布をもつという仮定をおき，それらの文書の属性ベクトルの平均ベクトルを用いてモデルのパラメータを学習する．具体的には，学習に用いる観測拡散系列の集合を $\mathcal{D} = \{\mathcal{D}_1, \cdots, \mathcal{D}_M\}$ とし，各拡散系列は $\mathcal{D}_m = \{(u, t_{m,u}), (v, t_{m,v}), \cdots\}$ という形式で表現されるものとする．ここで，$(u, t_{m,u})$ は m 番目の系列において，ノード u が時刻 $t_{m,u}$ にアクティブになったことを意味する．このとき，$\mathcal{D}' \subset \mathcal{D}$ を特定の対象に関する文書 $\{d'_1, \cdots, d'_n\}$ に対する拡散系列の集合としたとき，各文章に対応する n 個のコンテンツ属性ベクトル $\boldsymbol{\theta}_{d'_1}, \cdots, \boldsymbol{\theta}_{d'_n}$ の平均ベクトル $\bar{\boldsymbol{\theta}}_{d'}$ を $\mathcal{D}'$ 中の各拡散系列に対する属性ベクトルとして利用することで，文献 [4] におけるパラメータ学習の枠組みがそのまま適用可能となる．

4 評価実験

4.1 実験設定

本実験では，2011 年 3 月 5 日から 3 月 10 日までの間に Twitter[1] 上で観測されたリツイート系列を用い，提案モデルの有用性を評価した．具体的には，ニュースサイト GIGAZINE[2] のアカウントが発信したツイートに対する 79 本のリツイート系列を抽出して利用した．それらのリツイート系列に対して，そこに現れる 3,118 のユーザをノードとし，ノード u のツイートをノード v がリツイートした場合にそれらの間に有向リンク (u, v) を張ることで，リンク数 3,784 の有向ネットワークを構築した．このネットワークを構築する際に用いたツイートからトピックモデルを学習し，それに基づき，各リンクに対するリンク属性（トピック分布），流れる情報に対するコンテンツ属性（トピック分布）を生成した．なお，LDA に関しては GibbsLDA++[3] を用い，トピック数は 10 とした．

[1] https://twitter.com
[2] http://gigazine.net/
[3] http://gibbslda.sourceforge.net

表 1: 提案モデルによる拡散系列長推定結果

元文書	実際の拡散系列長	推定拡散系列長
T1	42	41.4
T2	29	43.2
T3	73	37.9
T4	78	36.7
T5	17	45.1
T6	24	43.6
T7	17	44.7
T8	18	43.6
T9	17	37.1
平均値	41.6	41.5

実験では，iPhone に関するものと解釈できる 9 つのツイートに対するリツイート系列 9 本を対象に，提案モデルの下での情報拡散の予測結果と，従来の AsIC モデルによる情報拡散の予測結果を得られた情報拡散系列の長さという点で比較した．また，各リツイート系列の実際の系列長とも比較した．提案モデルに関しては，予測対象となるリツイート系列を除く 8 本の系列を用いてモデルのパラメータベクトルを学習した．AsIC モデルに関しては，従来研究 [3, 5] と同様にすべてのリンクにおける拡散確率が一様であるという仮定の下，拡散確率を 0.1 から 0.9 まで 0.1 刻みで変化させた．AsIC モデルにおける情報拡散はトピックに依存しないため，同一のネットワーク上で，同一の初期情報源ノード（GIGAZINE の公式アカウントノード）の下での情報拡散シミュレーションの結果を用いた．情報拡散系列長の予測値は，提案モデル，AsIC モデルいずれにおいても，100 回の拡散シミュレーション結果の平均値とした．

4.2 実験結果

実験で用いた 9 本のリツイート系列に対する元文書 T1〜T9 に対する実際の拡散系列長と提案モデルによって推定した拡散系列長を表 1 にまとめる．また，AsIC モデルの下で推定した拡散系列長を図 1 に示す．表 1 から，平均値を見ると，提案モデルにより実際の拡散系列長を精度よく予測できていることがわかる．ただし，個々の系列を見てみると，拡散系列長の平均値と実際の系列値の差が大きな系列に関しては，予測誤差も大きくなっている．これは，提案モデルのパラメータが観測系列に対する最尤推定の枠組みで学習されるためと考えられる．言い換えると，学習結果となる最尤推定量は，観測系列の平均系列長が得られるように最適化されるからである．

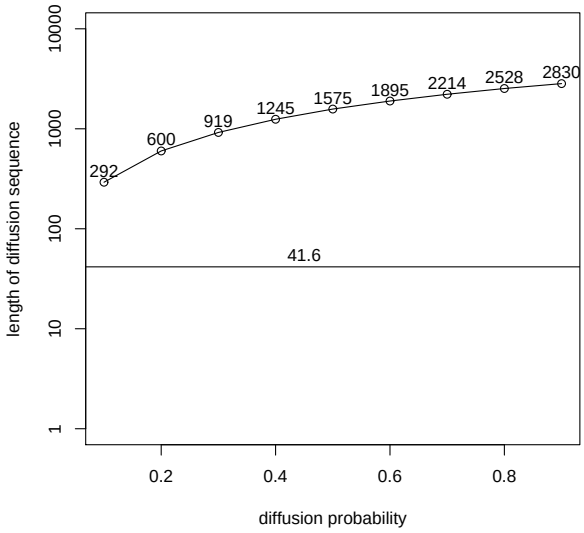

図 1: AsIC モデルによる拡散系列長の推定結果

一方，図 1 からは，トピックを考慮しない AsIC モデルによる拡散系列長の予測値は実際の拡散系列長と大きく乖離していることがわかる．実際には，拡散確率の値を 0.1 大きくするごとに予測系列長は約 300 ほどの割合で増加している．このような結果となった理由は，情報拡散ネットワークの形状に大きく依存しているものと考えられる．本実験で用いたネットワークは，実際の情報源ノードから拡散したリツイート系列を統合することで構築したものであるが，実際には，ほとんどのノードが情報源ノードに 1 本のリンクで直接つながるような形状になっていた．これは，実際の情報拡散が情報源ノードの周辺で多く生じ，直線的な連鎖がそれほど多く生じないことを意味している．その結果，今回のネットワークでは情報源ノードの出次数が約 3,000 となり，拡散確率 p が実際にはこの約 3,000 本のリンクの選択確率として機能するため，確率値が 0.1 変化するごとにその期待値が約 300 ずつ変化したものと考えられる．

以上の結果から，実際の情報拡散系列を精度よく推定するためには，流れるコンテンツのトピックを考慮することはきわめて重要であり，リンクに対する拡散確率が一様であるという従来用いられていた仮定は，情報拡散系列を過大予測するものと言える．

5 まとめ

本論文では，社会ネットワーク上の情報拡散を対象に，拡散する情報のトピックに着目した情報拡散モデルを提案した．従来のトピックに着目した情報拡散モデルに対して，提案したモデルは，観測した拡散系列からモデルのパラメータを学習することでより現実的

な拡散予測が可能であり，かつ学習すべきパラメータが仮定するトピック数の 2 倍程度と少ないという特徴を有する．また，実際のリツイート系列を用いた評価実験を通して，従来の社会ネットワーク分析で多用されていたすべてのリンクに対して拡散確率が一様であるという仮定は，情報拡散系列長の予測において非現実的な結果をもたらすこと，それに対して提案モデルは精度よく情報拡散系列長を予測できることを確認した．

ただし，本稿で用いた実験データは小規模であるため，今後は，より大規模，かつ多様なネットワークと情報拡散系列を用いて提案モデルを評価する必要がある．加えて，これまでに提案されているトピックを考慮した情報拡散モデルとの直接比較も必要である．また，従来のノード属性を考慮した情報拡散モデルにおけるパラメータ学習の枠組みをそのまま利用したため，現状ではコンテンツ属性ごとにモデルパラメータを学習する必要がある．今後は，複数のコンテンツ属性が混在する場合のパラメータ学習についても検討する必要がある．

なお，本稿で用いたツイートデータは，東京大学の鳥海不二夫准教授，和歌山大学の風間一洋教授から提供を受けた．また，モデルのパラメータ学習，情報拡散シミュレーションなどの実験環境は，静岡県立大学の斉藤和巳教授から提供いただいたプログラムを基に構築した．

参考文献

[1] Kleinberg, J.: The convergence of social and technological networks, *Communications of ACM*, Vol. 51, No. 11, pp. 66–72 (2008)

[2] Chen, W., Lakshmanan, L., and Castillo, C.: Information and influence propagation in social networks, *Synthesis Lectures on Data Management*, Vol. 5(4), pp. 1–177 (2013)

[3] Kempe, D., Kleinberg, J., and Tardos, E.: Maximizing the spread of influence through a social network, in *Proceedings of KDD'03*, pp. 137–146 (2003)

[4] Saito, K., Ohara, K., Yamagishi, Y., Kimura, M., and Motoda, H.: Learning Diffusion Probability based on Node Attributes in Social Networks, in *Proceedings of ISMIS'11*, pp. 153–162 (2011)

[5] Saito, K., Kimura, M., ohara, K., and Motoda, H.: Learning Asynchronous-Time Information Diffusion Models and Its Application to Behavioral Data Analysis over Social Networks, *Journal of Computer Engineering and Informatics*, Vol. 1, No. 2, pp. 30–57 (2013)

[6] Barbieri, N., Bonchi, F., and Manco, G.: Topic-aware social influence propagation models, *Knowledge and Information Systems*, Vol. 37, pp. 555–584 (2013)

[7] Chen, S., Fan, J., Li, G., Feng, J., Tan, l. K., and Tang, J.: Online topic-aware influence maximization, *Proceedings of the VLDB Endowment*, Vol. 8(6), pp. 666–677 (2015)

[8] Blei, D. M., Ng, A. Y., and Jordan, M. I.: Latent Dirichlet Allocation, *Journal of Machine Learning Research*, Vol.3, pp.933–1022 (2003).