F-014

社会ネットワーク上での情報拡散データの分析

Data Analysis of Information Diffusion in Social Networks

吉川 友也† 伏見 卓恭† 斉藤 和巳† 元田 浩‡ 木村 昌弘§
Yuya Yoshikawa Takayasu Fushimi Kazumi Saito Hiroshi Motoda Masahiro Kimura

1. はじめに

近年情報ネットワーク技術の進展により，Web 上で誰とも簡単にコミュニケーション出来るブログが普及してきている．このブログ空間では様々な情報が行き交い，まさに情報拡散現象の原理研究のための重要なメディアになっている．

このようなブログネットワーク上の情報拡散分析を行うための基本的なモデルに独立カスケードモデル（Independent Cascade model，以下 IC モデル）[1]がある．しかし，IC モデルでは情報拡散の時間遅れを考慮できず，現実の情報拡散の仕組みとは異なる面もある．また，Gruhl et al.[2]は IC モデルに時間遅れを取り入れた初めての研究を行ったが，彼らのモデルにおいて，時間は離散時間として扱われている．現実の世界が連続時間であることを考えると，彼らのモデルにも不十分な点がある．

以上を踏まえて，本論文では，従来の IC モデルに情報拡散の時間遅れを考慮した連続時間遅れ付き独立カスケードモデル（Continuous Time delay Independent Cascade model，以下 CTIC モデル）を採用し，CTIC モデルにおけるパラメータを推定する方法について述べる．さらに，実際のブログデータに CTIC モデルを適用し，情報拡散を分析する．

2. CTIC モデルにおけるパラメータの推定法

2.1 連続時間遅れ付き独立カスケード CTIC モデル

有向ネットワークを $G=(V, E)$ で定義する．$V=\{u, v, w, \cdots\}$ はノード集合を，$E=\{(u, v), (v, w), \cdots\}$ はリンク集合を表す．ここで，ノード v がリンクする子ノード集合を $F(v)=\{w; (v, w) \in E\}$ とし，ノード v へリンクする親ノード集合を $B(v)=\{u; (u, v) \in E\}$ で表す．今，ノードが情報を保持している状態をアクティブと呼び，そうでない状態を非アクティブと呼ぶ．CTIC モデルでは，非アクティブからアクティブへ状態は変わるが，逆は起こらない．アクティブなノード v は，各出リンクを通し独立に子ノード集合 $F(v)$ の各ノードを確率 κ $(0 \leq \kappa \leq 1)$ でアクティブにすることができる．この情報拡散試行が行われるのは一度限りで，時刻 t_v で子ノード w をアクティブにする試行に成功したとき，w がアクティブになる時刻は指数分布 $p(t)=r \exp(-r(t-t_v))$ で与えられるとする．r は指数分布のパラメータを表す．なお，リンク毎に，拡散確率や指数分布パラメータが異なるように一般化したモデルも同様に定義できる．

2.2 CTIC モデルでの学習問題とパラメータ推定

情報拡散データとしてアクティブとなったノード集合 $D \subseteq V$ が与えられたとする．いま，$u, v, \cdots \in D$ とし，各ノードがアクティブとなった時刻リストを $\langle t_u, t_v, \cdots \rangle$ とする．また，時刻 t 以前にアクティブとなったノード集合を $C(t)=\{u \in D; t_u < t\}$ で表記する．このとき，情報拡散試行に成功した可能性のあるリンク集合を

$$E^{+}=\{(u, v) \in E; v \in D, u \in B(v) \cap C(t_v)\}$$

情報拡散試行に失敗した可能性のあるリンク集合を

$$E^{-}=\{(v, w) \in E; v \in D, w \in F(v) \setminus D\}$$

で定義することができる．

時刻 t_v でノード $v \in D$ がアクティブになったとする．あるノード $u \in B(v) \cap C(t_v)$ が情報拡散に成功する確率は

$$\Phi_{u,v}=\kappa \cdot r \cdot \exp(-r(t_v-t_u))$$

で与えられる．情報拡散に失敗するか，もしくは成功しても時刻 t_v までにノード v をアクティブにできない確率は

$$\Psi_{u,v}=1-\int_{t_u}^{t_v} \kappa \cdot r \cdot \exp\left(-r(t-t_u)\right) dt$$

$$=\kappa \cdot \exp\left(-r(t_v-t_u)\right)+(1-\kappa)$$

となる．よって，親ノード集合の部分集合 $B(v) \cap C(t_v)$ のどれか一つのノードがノード v を時刻 t_v でアクティブとする確率は

$$h_v=\sum_{u \in B(v) \cap C(t_v)} \Phi_{u,v} \cdot \prod_{u' \in B(v) \cap C(t_v), u' \neq u} \Psi_{u',v}$$

$$=\prod_{u \in B(v) \cap C(t_v)} \Psi_{u,v} \cdot \sum_{u \in B(v) \cap C(t_v)} \Phi_{u,v} \Psi_{u,v}^{-1}$$

で計算できる．一方，観測されたデータは，最後にアクティブとなったノードの時刻から十分時間が経過するまで情報拡散データを観測した結果とすれば，親ノードの一つがアクティブとなったにもかかわらず，最終的にアクティブとならなかったノード集合では，情報拡散試行に失敗したと見なすことが可能である．よって，情報拡散データから拡散確率や指数分布パラメータを推定する問題は，次式の尤度関数 (likelihood function) を最大化する κ と r を求める問題として定式化できる．

$$L(\kappa, r; D)=\prod_{v \in D, t_v>0} h_v \prod_{(v,w) \in E^{-}} (1-\kappa)$$

尤度関数 $L(\kappa, r; D)$ を最大にする拡散確率 κ と指数分布パラメータ r は，EM (Expectation-Maximization) アルゴリズムと同様な反復法により求めることができる．導出の詳細はページ数の制限のため割愛するが，最終的な更新式のみを以下に示す．まず，現時点の推定値を用いて，事後確率に相当する以下の値を求める．

$$\alpha_{u,v}=\frac{\Phi_{u,v} \Psi_{u,v}^{-1}}{\sum_{u \in B(v) \cap C(t_v)} \Phi_{u,v} \Psi_{u,v}^{-1}}$$

$$\beta_{u,v}=\frac{\kappa \cdot \exp(-r \cdot (t_v-t_u))}{\Psi_{u,v}}$$

次いで，以下の式で拡散確率 κ と指数分布パラメータ r のそれぞれを更新する．

† 静岡県立大学経営情報学部 University of Shizuoka
‡ 大阪大学 Osaka University § 龍谷大学 Ryukoku University

$$\kappa = \frac{1}{|E^+| + |E^-|} \sum_{(u,v)\in E^+} (\alpha_{u,v} + (1-\alpha_{u,v})\beta_{u,v})$$

$$r = \frac{\sum_{(u,v)\in E^+} \alpha_{u,v}}{\sum_{(u,v)\in E^+} (\alpha_{u,v} + (1-\alpha_{u,v})\beta_{u,v}) \cdot (t_v - t_u)}$$

上記反復を繰り返せば局所最適解への収束を保証することができる．また，リンク毎に，拡散確率や指数分布パラメータが異なるように一般化したモデルでも同様な推定法を導くことができる．

3. 評価実験

3.1 実験データ

Adar & Adamic [3]は，トピックがブログ間で伝わることを URL がブログ間で伝わることと等価であると見なした方法を提案した．この方法でブログに現れる URL を追跡すれば，トピック拡散の分析をすることができる．分析には，Doblog※1 のデータ※2 から作成されたデータベースを利用した．ブログ数（ユーザー数）は 52,525，ブログロールのリンク数は 115,552 であった．

3.2 実験結果

実験では，ブロガ 10 人以上が関与した 172 件の URL を対象とした．

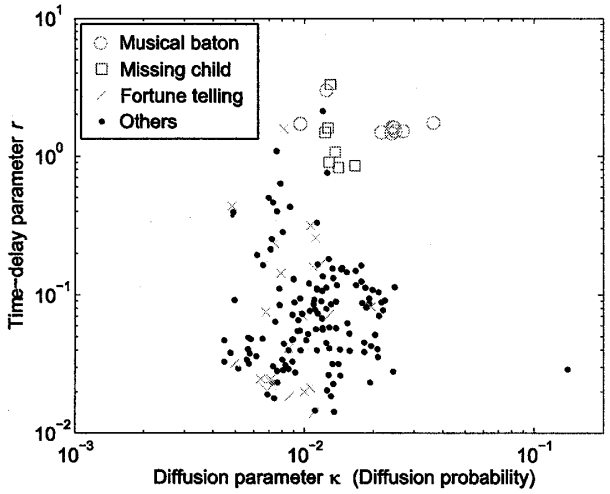

図 1　Doblog データによる分析結果

図 1 は，ブログロール上での各 URL の拡散データからパラメータを計算し，縦軸に時間遅れパラメータ r，横軸に拡散確率パラメータ κ をプロットした結果である．r は，r=1 のとき 1 日の遅れを表し，r=0.1 のとき 10 日の遅れを表すように設定している．ここでは，グループ分けすることができたトピックについての結果について説明する．「○印」は「ミュージカルバトン」についてのトピックを表している．これは以下のルールに基づいたインターネット上の伝言ゲームである．まず，ブロガはあるブロガによって，音楽についての 5 つの質問に返答することを要求される（バトンを受け取る）．そして，要求されたブロガは質問に答え，次に質問を答える 5 人を指名し，自分自身が受けた質問と同じ質問をする（バトンを渡す）．このトピックは素早く拡散した（平均一日以内）．これはこの種の伝言ゲームのようなものに対して，人々は興味をひかれやすいからだと推測される．「□印」は子供の失踪についての記事の URL を表し，これも素早く拡散した URL のひとつである．これは，このトピックの持つメッセージの切迫感によるものだと推測できる．「×印」は占いについての記事の URL を表す．これに対する人々の反応は様々で，素早く反応（1 日以内）する人もいれば，遅れて反応（一か月以上の後）する人もあり，ほぼ均一に分布している．

※1 © （株）NTT データ．http://www.doblog.com.

※2 （株）ホットリンクと（株）NTT データの共同事業契約に基づき，（株）ホットリンクより提供．2003 年 10 月から 2005 年 6 月までのデータを利用．

4. 考察

時系列データの分析には多くの手法の研究がある．我々の方法もこの種の研究の一つになる．しかし，我々の研究が従来の研究と異なるのは，複雑ネットワークの構造を前提としている点であり，時系列データの分析において新しい視点によるものと言える．トピック拡散分析に関しても多くの研究があるが，それらはほとんど平均拡散スピード（拡散スピード分布）と平均生存時間の分析に焦点を当てている．我々は時間遅れとネットワーク構造を合わせた拡散現象に取り組んでおり，その点で他の研究と異なる．

CTIC モデルを現実のブログネットワークに適用した評価実験では，トピックの特性によって拡散確率と拡散スピードが違うことが明確に示された．中でも，ミュージカルバトンと子供の失踪についてのトピックは素早い拡散スピードと高い拡散確率を持ち，我々がこれらトピックに対して持つ想像と一致する結果になった．

5. おわりに

本論文では，従来の基本的な情報拡散モデルである IC モデルに連続時間遅れを取り入れた CTIC モデルにおいて，時間遅れと拡散確率の 2 つのパラメータを推定する方法について述べた．さらに，現実のブログデータに CTIC モデルを適用し，データの分析を行った．その結果拡散する情報の種類によって，拡散する確率や時間遅れに違いがあることが明らかとなった．言い換えれば，拡散確率 κ と時間遅れ r によってトピックをグループ分けすることができた．今後は，CTIC モデルを様々な現実のネットワーク上での拡散データに適用し，この手法の有用性を検証していく．

謝辞

Doblog データは（株）NTT データおよび（株）ホットリンクより提供を受けた．記して感謝致します．

参考文献

[1] Kempe, D., Kleinberg, J., Tardos, E., “Maximizing the spread of influence through a social network.”, Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2003)., 137–146, (2003).

[2] Gruhl, D., Guha, R., Liben-Nowell, D., Tomkins, A., “Information di_usion through blogspace.”, SIGKDD Explorations 6 (2004), 43–52,(2004).

[3] Adar, E., Adamic, L.A., “Tracking information epidemics in blogspace.”, Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence (WI’05)., 207–214,(2005).