ノード属性を考慮した情報拡散影響度の推定

山岸　祐己† 　風間　一洋†† 　斉藤　和巳†

† 静岡県立大学 経営情報学部　〒 422–8526 静岡県静岡市駿河区谷田 52–1
†† 日本電信電話株式会社 未来ねっと研究所　〒 180–8585 東京都武蔵野市緑町 3–9–11
E-mail: †{b08107,k-saito}@u-shizuoka-ken.ac.jp, ††kazama@ingrid.org

あらまし　社会ネットワーク上の代表的な情報拡散モデルの一つである IC モデルを拡張し，各リンクの情報拡散確率がノード属性の関数として規定されるモデルを提案する．また，その応用として情報拡散影響度の推定問題に対する解法を提案する．大規模社会ネットワークを用いた実験では，情報拡散確率がノード属性で規定されるケースにおいて，リンクに一様な拡散確率を付与する従来アプローチと比較し，提案法による情報拡散影響度の推定精度が大幅に優ることを示す．

キーワード　社会ネットワーク，情報拡散モデル，情報拡散影響度，ノード属性

Estimation of Information Diffusion Degree by Considering Nodes' Attribute Values

Yuki YAMAGISHI†, Kazuhiro KAZAMA††, and Kazumi SAITO†

† School of Administration and Informatics, University of Shizuoka
52–1 Yada, Suruga–ku, Shizuoka, 422–8526 Japan
†† Nippon Telegraph and Telephone Corporation Network Innovation Laboratories
3–9–11 Midori-cho, Musashino-shi, Tokyo 180–8585, Japan
E-mail: †{b08107,k-saito}@u-shizuoka-ken.ac.jp, ††kazama@ingrid.org

Abstract　By extending the independent cascade model, one of representative information diffusion models over social networks, we propose a new model in which information diffusion probabilities over links are formulated as a function of nodes' attribute values. As its application, we focus on a problem of estimating the expected information diffusion degree based on the extended model, and propose its solution method. In our experiments using large-scale social networks, we evaluate the performance of the expected information diffusion degree estimation for the case that the information diffusion probabilities are determined by nodes' attribute values. As our experimental results, we show that our proposed method is much better than the conventional approach assuming that all of the information probabilities over links are the same.

Key words　social networks, information diffusion model, expected information diffusion degree, nodes' attribute values

1. はじめに

World Wide Web の発展とソーシャルメディアの登場は，大規模な社会ネットワークの発生を促進させている．よって，情報を普及させるための重要メディアとして，現在，社会ネットワークが注目されている [1]．ここで社会ネットワークとは，例えばノードを人とし，その友人関係をリンクとして繋いだネットワークである．また，特に画期的な新製品の普及には、マスメディアによる不特定多数を対象とした宣伝よりも，現実の人間関係やソーシャルメディアを介した口コミ（word-of-mouth）の方が重要な役割を果たしている．このような情報拡散に対する基本的な確率モデルとしては，独立カスケード（IC: Independent Cascade）モデルや線形閾値（LT: Linear Threshold）モデルなどが広く研究されている [2], [3]．このようなモデルを適用すれば，情報拡散影響度の高いノード（人物）を求めることができる．本論文では，ノード属性を考慮した情報拡散モデルについて検討する．ここでノード属性とは，例えばノードを人としたとき，その人の性別や年齢，または趣味などを意図する．明らかに，属性値の類似度の高いノード間においては，そうでないものと比較し，一般に情報拡散確率が大きくなると自然に想

定できる．ところが，これまでの情報拡散モデルでは，ノード属性など考慮することなく，情報拡散確率が規定される枠組みとなっていた．よって，ノード属性を考慮した情報拡散モデルの構築とともに，このようなモデルに基づき，各ノードの情報拡散影響度を推定する手法の考案は重要な研究課題と言える．

本論文の構成は以下となる．まず，本研究の土台となる情報拡散の基本モデルについて説明する．次に，ノード属性を考慮した情報拡散モデルについて述べるとともに，実験による評価結果を報告する．最後に，本研究のまとめについて述べる．

2. 情報拡散の基本モデル

本稿では，有向グラフ $G = (V, E)$ により表現される社会ネットワーク上において，ある種の情報が伝播し広がっていく現象を論じる．その情報が伝わり，かつその情報を受け入れたノードを，**アクティブノード**と呼び，そうでないノードを**非アクティブノード**と呼ぶ．グラフ G 内のノードの総数を $N\ (= |V|)$ とし，リンクの総数を $L\ (= |E|)$ とする．ノード $v \in V$ の親ノード全体の集合を $\Gamma(v)$ とする．ここで親ノードとは他のノードをアクティブにし得るノードのことを指す．情報拡散モデルは，情報拡散過程は連続時間 $t \geqq 0$ で展開していく．また，ノードは非アクティブからアクティブに変化することはできるが，アクティブから非アクティブへは変化できないと仮定する．アクティブノードの初期集合 S が与えられたとき，S に属するノードは時刻 0 で初めてアクティブになったと見なし，その他のノードは時刻 0 では非アクティブであると見なす．

2.1 独立カスケードモデル

非同期時間（asynchronus time）IC モデルでは，各有向リンク (u, v) に対して，実数値 $p_{u,v} \in [0, 1]$ を前もって指定しなければならない．ここに，$p_{u,v}$ はリンク (u, v) を通しての**拡散確率**と呼ばれる．本モデルの情報伝播過程は，アクティブノードの初期集合 S が与えられたとき，次のように進んでいく．ノード u は，時刻 t で初めてアクティブになったとしよう．このとき，u はその未だ非アクティブである子ノード v をアクティブにする唯一のチャンスを与えられ，その試行は確率 $p_{u,v}$ で成功する．そしてもし u が成功したならば，v は時刻 $t + \delta t$ でアクティブとなる．ここで，δt は，例えば指数分布のような確率分布に従い指定される．u が時刻 t で v をアクティブにするのに成功したか否かにかかわらず，時刻 t 以降では，u はもはや v をアクティブにする試行を行うことはできない．非アクティブノードをアクティブにする新たな試行が不可能になったとき，本情報伝播過程は終了する．

2.2 情報拡散期待影響度

ネットワーク $G = (V, E)$ において，IC モデルの拡散確率ベクトルを $\Theta = (p_{v,w})_{(v,w)\in E}$ とする．初期アクティブノード集合 S に対して，本 IC モデルの拡散過程終了後のアクティブノード数を $\varphi(S; \Theta)$ とする．IC モデルの拡散過程は確率過程であるので，$\varphi(S; \Theta)$ は確率変数となる．よって，$\varphi(S; \Theta)$ の期待値を $\sigma(S; \Theta)$ とし，$\sigma(S; \Theta)$ を，拡散確率ベクトル Θ の IC モデルにおけるノード集合 S の**影響度**と定義する．特に，$S = \{v\}$ のとき，$\sigma(S; \Theta)$ を $\sigma(v; \Theta)$ と表記し，$\sigma(v; \Theta)$ を拡散確率ベクトル Θ の IC モデルにおけるノード v の**影響度**と呼ぶ．ここに，影響度の高いノードは，IC モデルに基づく情報拡散において有力なノードである．

3. ノード属性を考慮した情報拡散モデル

まず，ノード属性を考慮した情報拡散モデルについて述べる．ノード v の属性値からなる M-次元ベクトルを $\mathbf{x}_v = (x_{v,1}, \cdots, x_{v,M})^T$ で表す．ここで，$\mathbf{x}_v^T$ によりベクトル $\mathbf{x}_v$ の転置を表すものとする．ノード u から v へリンクが存在するとき，情報拡散確率を以下で定義するモデルを考える．

$$p_{u,v}(\mathbf{w}) = f\left(w_0 + \sum_{m=1}^{M} w_m \delta(x_{u,m}, x_{v,m})\right). \tag{1}$$

ここで，$\mathbf{w}$ は，各属性ごとの情報拡散確率への寄与度を制御する $(M+1)$-次元のパラメータベクトルを表す．一方，$f(\cdot)$ は，情報拡散確率のレンジを $(0, 1)$ とするためのシグモイド関数 $f(y) = 1/(1 + \exp(-y))$ である．また，$\delta(\cdot, \cdot)$ は，属性値ペアに関して，性別のような名義値の属性に対しては，

$$\delta(x_{u,m}, x_{v,m}) = \begin{cases} 1 & \text{if } x_{u,m} = x_{v,m}, \\ 0 & \text{otherwise;} \end{cases} \tag{2}$$

年齢のような数値の属性に対しては，

$$\delta(x_{u,m}, x_{v,m}) = \exp(-|x_{u,m} - x_{v,m}|). \tag{3}$$

により定義される関数である．

次に，各ノードの情報拡散影響度を推定する手法について述べる．上述したモデルにおいて，パラメータ $\mathbf{w}$ は，観測した拡散データ D から推定するものとする．そのためには文献 [4] と同様な方法に，拡散データ D に対するモデルの尤度式を構築して推定することができる．最尤推定パラメータ $\hat{\mathbf{w}}$ が得られれば，式 1 に従い，拡散確率 $p_{u,v}(\hat{\mathbf{w}})$ が求まる．一方，拡散確率 $p_{u,v}(\hat{\mathbf{w}})$ が得られれば，文献 [5] で提案されたボンドパーコレーション法を用い，各ノードの影響度を効率良く計算することができる．詳細には，1 回のボンドパーコレーションで，各ノード v のアクティブノード数 $\varphi(v; \Theta)$ を同時に求め，これを L 回繰り返して影響度 $\sigma(v; \Theta)$ を求める方法である．このアルゴリズムをまとめれば以下となる．

1. 拡散データ D を用いて最尤推定パラメータ $\hat{\mathbf{w}}$ を求める．
2. L 回のボンドパーコレーションにより各ノードの影響度 $\sigma(v; \Theta)$ を求める．

なお，以下で述べる実験では，ボンドパーコレーション回数を $L = 1,000$ に設定した．

4. 実験による評価

評価に用いるネットワークデータ，実験設定，そして実験結果について述べる．

4.1 ネットワークデータ

本論文で取り扱う実験は，3 つのネットワークデータを用い

て行ったものである．これら３つのネットワークは，多くの大規模なネットワークと同じなように，自分に向かうリンク次数の分布（入次数分布）も自分から向かうリンク次数の分布（出次数分布）もべき則分布に従うという特徴を有する．

１つ目のネットワークデータは，ブログのトラックバックネットワークのデータである．ブログすなわち Weblog のトラックバックとは，他人のブログの記事の内容を引用・参照した時，あるいは他人のブログの記事が自身のブログの記事と関連性のある話題を書いている場合などに，自身のブログの記事が引用・参照した事や関連性がある事を通知する目的で行われるものである．ブログネットワークのデータは「goo ブログ」（http://blog.goo.ne.jp/usertheme/）の「JR 福知山線脱線事故」というテーマからトラックバックを 10 段辿ることにより，2005 年 5 月に収集したものである．データ中のノードを記事，トラックバック関係をリンクとしてネットワークを構築した．このネットワークは 12,047 ノードと 79,920 リンクをもつ有向ネットワークで構成される．以下このネットワークをブログネットワークと呼ぶ．

２つ目のネットワークデータは，日本の Wikipedia 内の「人名一覧」の人名共起ネットワークのデータである．実際に「人名一覧」に登場する人物において，Wikipedia 内の同一記事中の共起回数が６回以上の２人の人物をリンクで結ぶことによりネットワークを構築した．以下このネットワークを Wikipedia ネットワークと呼ぶ．ノード数は 9,481，有向リンク数は 245,044 である．

３つ目のネットワークは，エンロンの Email 送受信ネットワークのデータである．まず最初に，データセット中に現れる Email アドレスを送信者と受信者として抽出した．そして各 Email アドレスをノードとみなし，メールが送信された向きに従ってノード間に有向リンクを張り，ネットワークを構築した．このネットワークは 4,254 ノードと 44,314 有向リンクで構成される．以下このネットワークをエンロンネットワークと呼ぶ．

4.2 実験設定

提案法の基本能力を評価するために，各ノードに対してランダムな属性値を付与して実験を行った．詳細には，属性数は $M = 10$ とし，その内の５つを名義属性値と見なし $\{1, 2, 3\}$ の値を一様ランダムに付与し，残りの５つについては数値属性と見なし $1 \leqq n \leqq 20$ の範囲の整数値を同様に一様ランダムに付与した．パラメータ $\mathbf{w}$ については，真のパラメータベクトルを以下のように設定した．

$$\mathbf{w}^* = (-2, 2, -1, 0, 0, 0, 1, -2, 0, 0, 0,)^T \qquad (4)$$

ここで，最初の数値により $w_0 = -2$ に設定し，$w_m = 0$ と設定した属性は拡散確率に影響を与えない不要属性としている．また，パラメータには負の値のものも設定し，類似していれば情報が伝わりにくくなるような属性も許容する枠組みとなっている．

次に，拡散データ D 生成法について述べる．すなわち，真のパラメータ $\mathbf{w}^*$ より求めた拡散確率を用い，全ノード集合からランダムに一つ情報源ノードを選定し，１つの拡散系列を生成するとした．実験では，このようにして生成した５個の拡散系列により拡散データ D を構成し，最尤推定パラメータ $\hat{\mathbf{w}}$ を求めた．

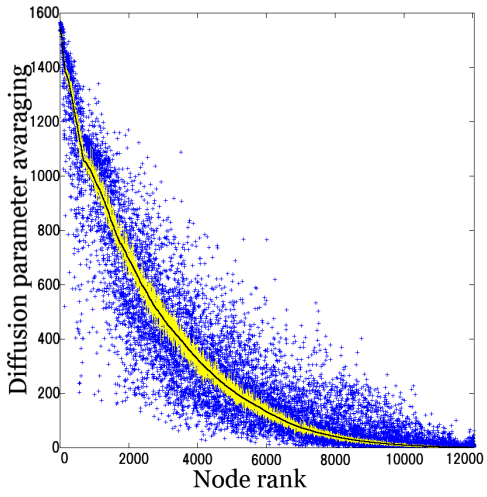

図 1　ブログネットワークの影響度比較

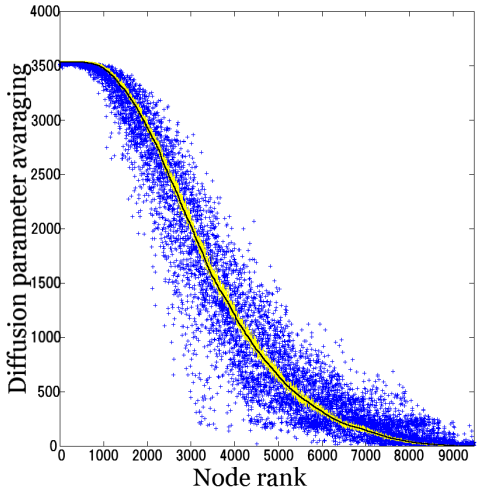

図 2　Wikipedia ネットワークの影響度比較

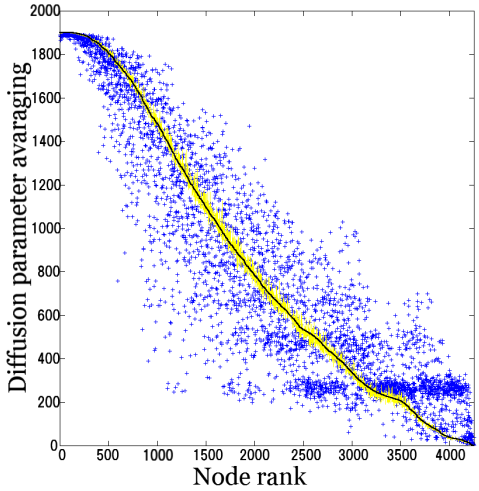

図 3　エンロンネットワークの影響度比較

4.3 実験結果と考察

実験では，真のパラメータ $\mathbf{w}^*$ を用いて求めた影響度（真の

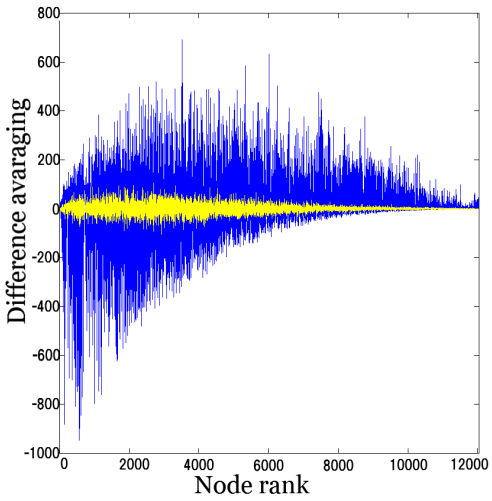

図 4　真の影響度からの推定誤差（ブログ）

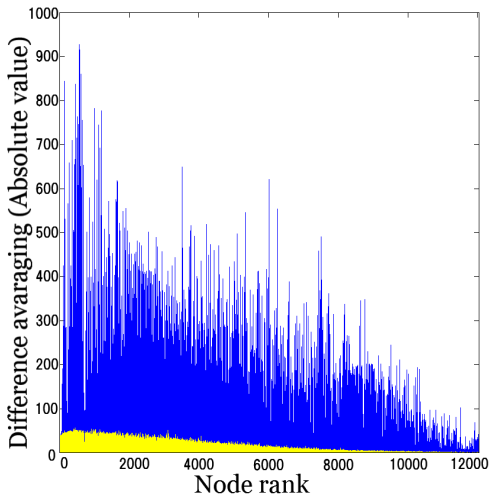

図 5　試行 100 回における誤差絶対値の平均（ブログ）

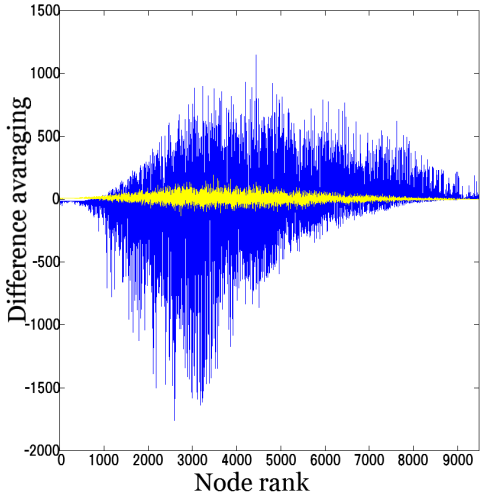

図 6　真の影響度からの推定誤差（Wikipedia）

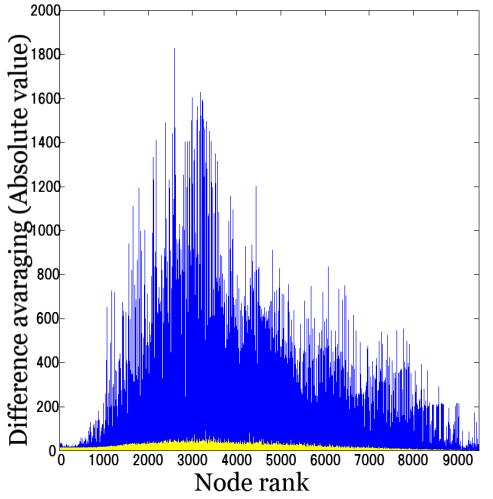

図 7　試行 100 回における誤差絶対値の平均（Wikipedia）

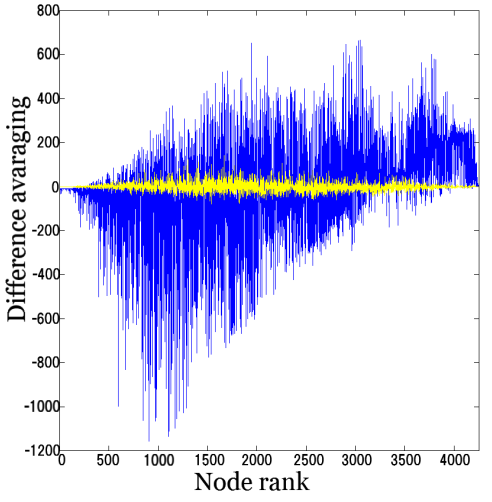

図 8　真の影響度からの推定誤差（エンロン）

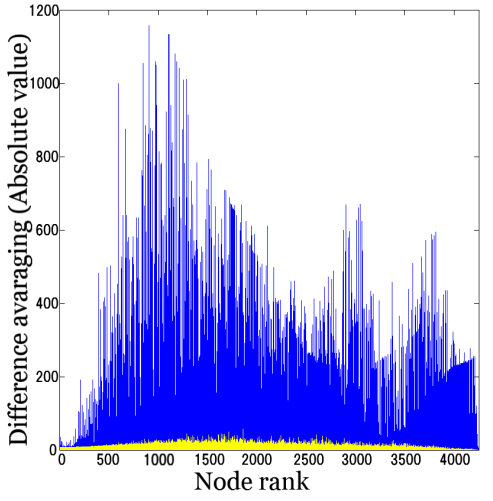

図 9　試行 100 回における誤差絶対値の平均（エンロン）

影響度と呼ぶ）に対して，観測データからパラメータ推定して影響度を求める提案法を比較した．また，比較対象として，多くの従来研究 [2]〜[5] で採用されているように，全てのリンクの拡散確率を一様に設定し求めた影響度（従来法と呼ぶ）とも比較評価する．ここで，一様に設定する従来法の拡散確率は，真のパラメータ $\mathbf{w}^*$ で求めた拡散確率 $p_{u,v}(\mathbf{w}^*)$ の平均値とした．

図 1 には，ブログネットワークにおける影響度の比較結果を示す．ここで横軸は，真の影響度に基づき降順にソートしたときのノード番号に対応し，縦軸は影響度を表す．図より，黒線

で示す真の影響度は単調に減少し，これと近い数値として，黄色線で示す提案法による影響度が求まっていることが分かる．これに対して，青点で示す従来法の影響度は，提案法と比較して大幅に違っていることが分かる．図 4 には，真の影響度からの差分を求めることにより，提案法と従来法での推定誤差を比較する．図より，提案法の推定精度が従来法より優っていることが分かる．図 5 には，生成する観測データ D を変えて行った 100 回の試行における誤差絶対値の平均を示す．この図より，観測データに大きく影響されることなく，提案法は平均して優れた性能を示していることが分かる．

図 2, 6 と 7 には，上述したブログネットワークと同様にして行った，Wikipedia ネットワークにおける影響度の比較結果を示す．また，図 3, 8 と 9 には，エンロンネットワークにおける影響度の比較結果を示す．これらの図から分かるように，ある程度異なったネットワーク構造に影響されることなく，従来法と比較して，提案法は優れた性能を示していることが分かる．

5. おわりに

社会ネットワーク上の代表的な情報拡散モデルの一つ IC モデルを拡張し，各リンクの情報拡散確率がノード属性の関数として規定されるモデルを提案した．また，その応用として情報拡散影響度の推定問題に対する解法を提案した．大規模社会ネットワークを用いた実験では，情報拡散確率がノード属性で規定されるケースにおいて，リンクに一様な拡散確率を付与する従来アプローチと比較し，提案法による情報拡散影響度の推定精度が大幅に優ることを示した．今後は，さらに多様なデータに提案法を適用し，その有効性を評価する予定である．

謝　辞

本研究は，科学研究費補助金基盤研究 (C) (No. 22500133) の補助を受けた．

文　献

[1] Leskovec, J., Adamic, L., Huberman,B. A.,”The dynamics of viral marketing”, EC’06, 228-237 (2006)

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

[3] Kimura, M., Saito, K., and Motoda, H., ”Blocking links to minimize contamination spread in asocial network”, ACM Transactions on Knowledge Discovery from Data, Vol.3, No.2, Article 9 (2009)

[4] Saito, K., Kimura, M., Ohara, K., Motoda, H., ”Learning continuous-time information diffusion model for social behavioral data analysis”, Proceedings of the 1st Asian Conference on Machine Learning (ACML-2009), 322-337 (2009)

[5] Kimura, M., Saito, K., Nakano, R., ”Extracting influential nodes for information diffusion on a social network”, Proceedings of the 22nd AAAI Conference on Artificial Intelligence (AAAI-2007), 1371-1376 (2007).