情報拡散データに基づいた社会ネットワークのノードランキング

Finding Influential Nodes in a Social Network from Information Diffusion Data

木村昌弘[*1] 斉藤和巳[*2] 中野良平[*3] 元田浩[*4]
Masahiro Kimura Kazumi Saito Ryohei Nakano Hiroshi Motoda

[*1]龍谷大学 Ryukoku University
[*2]静岡県立大学 University of Shizuoka
[*3]中部大学 Chubu University
[*4]大阪大学 Osaka University

We address the problem of ranking influential nodes in complex social networks by estimating diffusion probabilities from observed information diffusion data using the popular independent cascade (IC) model. For this purpose we formulate the likelihood for information diffusion data which is a set of time sequence data of active nodes and propose an iterative method to search for the probabilities that maximizes this likelihood. We apply this to two real world social networks in the simplest setting where the probability is uniform for all the links, and show that the accuracy of the probability is outstandingly good, and further show that the proposed method can predict the high ranked influential nodes much more accurately than the well studied conventional four heuristic methods.

1. はじめに

イノベーション，ホットトピックス，さらには悪意のある噂も，人々の間の社会ネットワークを通じて，所謂 “クチコミ” という形で伝搬しうる．インターネットや World Wide Web の興隆は，大規模な社会ネットワークの発生を加速している．したがって，情報を普及させるための重要メディアとして，最近，社会ネットワークが注目されている [Backstrom 06, Leskovec 06].

社会ネットワーク上での情報拡散の基本確率モデルとしては，“independent cascade (IC) モデル” が広く用いられている [Kempe 03, Gruhl 04]．IC モデルを用いて，情報の普及に対して有効である指定された数のノード群を抽出するという，組み合わせ最適化問題が研究されている [Kempe 03, Kimura 07]．本問題は，影響最大化問題と呼ばれている．一方，指定された数のリンク群を封鎖することにより好ましくない情報の普及を最小化するという，影響最大化問題と反対の問題も最近研究されている [Kimura 09]．本論文でも，与えられた社会ネットワーク上での IC モデルに基づく情報拡散現象を考える．

一般に，与えられた社会ネットワークから影響力が強いノード群を抽出することは，社会ネットワーク分析分野における最も中心的な課題の一つであり，ネットワーク構造に基づいた幾つかのノードランキング法が提案されている [Wasserman 94]．本論文では，異なった角度からこの課題に取り組む．すなわち，ネットワーク上での情報拡散の観測データに基づいて，IC モデルにおける “影響度” に関しノードをランキングすることにより，影響力が強いノード群を抽出するという手法を提案する．ところで，IC モデルはパラメータをもっている．より具体的には，ネットワーク上の各リンクに対して．“拡散確率” を前もって指定しなければならない．我々は，情報拡散データの観測集合を得る尤度を繰り返しアルゴリズム（EM アルゴリズム）により最大化することで，拡散確率を推定する．二つの実ネットワークを用いた実験により，提案法の有効性を検証する．まず，拡散確率の推定精度を評価し，次に，推定したモデルを影響力が強いノード群を抽出するために用い，その結果を真の結果と比較するとともに，また，社会ネットワーク分析からの四つのヒューリスティクスによる結果とも比較する．

連絡先: 木村昌弘，龍谷大学理工学部電子情報学科，〒 520-2194
大津市瀬田大江町横谷 1-5，kimura@rins.ryukoku.ac.jp

以下の構成は次のとおりである．2 節では，提案法を機械学習問題として定式化する．3 節では，実験設定および実験結果を述べる．4 節では，拡散確率がノードの影響度に対してどのような影響を与えるかについて考察する．5 節はまとめである．

2. 提案法

2.1 問題の定式化とランキング法

与えられた有向ネットワーク（グラフ）$G = (V, E)$ に対して，V をノード（頂点）全体の集合，E をリンク（辺）全体の集合とする．ノード v からノード w への有向リンク e を，$e = (v, w)$ と記述する．ノード v の子ノード全体の集合を，$F(v) = \{w; (v, w) \in E\}$ とし，ノード v の親ノード全体の集合を，$B(v) = \{u; (u, v) \in E\}$ とする．

IC モデルでは，各有向リンク $e = (v, w)$ に対して，$0 < p_{v,w} < 1$ なる実数 $p_{v,w}$ を前もって指定する必要がある．ここに，$p_{v,w}$ はリンク (v, w) を通じての “拡散確率” と呼ばれる．IC モデルの拡散過程は離散時間 $t \geq 0$ で展開していく．情報が伝わったノードを “アクティブ” と呼ぶ．ノードはその状態が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される．初期アクティブノード集合 $D(0)$ が与えられたとき，拡散過程は次のように進んでいく．ノード v が時刻 t で初めてアクティブになったとき，v は，非アクティブであるその各子ノード w をアクティブにする試行を時刻 t で行い，その試行は確率 $p_{v,w}$ で成功する．もし，w の複数の親ノードが時刻 t で初めてアクティブになった場合は，それら親ノードが w をアクティブにする試行は任意の順序で独立に順々に行われることになるが，これらの試行はすべて時刻 t で行われる．そして，w をアクティブにする試行のうち，少なくとも一つの試行が成功したとき，w は時刻 $t+1$ においてアクティブとなる．ところで，v が時刻 t で w をアクティブにするのに成功したか失敗したかにかかわらず，時刻 $t+1$ 以降では，v はもはや w をアクティブにする試行を行うことはできない．新たにアクティブとなるノードが存在しなくなったとき，拡散過程は終了する．

拡散確率

$$\Theta = \{p_{v,w}; (v, w) \in E\}$$

とノード $v \in V$ が与えられたとき，Θ における v の “影響

度”$\sigma(v;\Theta)$ を，拡散確率が Θ の IC モデルにおける初期アクティブノード v からの拡散過程終了後のアクティブノード数の期待値として定義する．IC モデルに関して影響力が強いノード群を抽出するという我々の問題は，影響度 $\sigma(v;\Theta_0)$ に基づいたノードランキング問題として定式化される．ここに，Θ_0 は真の拡散確率である．しかしながら，実際問題においては真の拡散確率は入手不可能である．そこで，我々は，過去の情報拡散履歴から推定した拡散確率 $\widehat{\Theta}$ を用いることを考える．ここに，情報拡散履歴はアクティブノード集合の時系列として観測される．このとき，$\sigma(v;\Theta_0)$ に従ったランキングと $\sigma(v;\widehat{\Theta})$ に従ったランキング間の類似度を評価する必要があることに注意しておく．

2.2 拡散確率推定法

$D = \langle D(0), D(1), \cdots, D(T)\rangle$ を一つの情報拡散結果とする．ここに，$D(t)$ は時刻 t でアクティブになったノードの集合である．いま，$v \in D(t)$, $e = (v,w) \in E$, $w \in D(t+1) \cap F(v)$ であるとしよう．このとき，ノード v はリンク e を通じてノード w をアクティブにすることに成功した可能性がある．しかしながら，$(D(t) \cap B(w)) \setminus \{v\} \neq \emptyset$ ならば，他のノード $v' \in D(t) \cap B(w)$ が w をアクティブにしたという可能性もまたある．よって，w が時刻 $t+1$ でアクティブになる確率は，

$$P(w;t+1) = 1 - \prod_{v \in B(w) \cap D(t)} (1 - p_{v,w})$$

と計算される．ここに，$w \in D(t+1)$ ならば $D(t) \cap B(w) \neq \emptyset$ であることに注意する．

さて，

$$C(t) = D(0) \cup \cdots \cup D(t)$$

と定義する．$C(t)$ は時刻 t でのアクティブノード全体の集合であることに注意する．$v \in D(t)$ で $w \in F(v) \setminus C(t+1)$ ならば，v はリンク $e = (v,w)$ を通じて w をアクティブにすることに失敗した，ということがわかる．明らかに，$v \in D(t)$ で $w \in F(v) \cap C(t)$ であるとき，または $v \notin D$ であるとき，リンク $e = (v,w)$ を通じての試行についての情報は得られない．したがって，拡散確率 $\Theta = \{p_{v,w}\}$ に関する観測集合 D の尤度関数は，

$$\mathcal{L}(\Theta; D) = \prod_{t=0}^{T-1} \prod_{w \in D(t+1)} \left(1 - \prod_{v \in B(w) \cap D(t)} (1 - p_{v,w})\right) \prod_{t=0}^{T} \prod_{v \in D(t)} \prod_{w \in F(v) \setminus C(t+1)} (1 - p_{v,w})$$

と定義できる．

$\{D_m; 1 \leq m \leq M\}$ を M 個の独立な情報拡散結果とする．このとき，Θ に関する目的関数は，

$$\mathcal{J}(\Theta) = \sum_{m=1}^{M} \log \mathcal{L}(\Theta; D_m). \quad (1)$$

と定義できる．よって，我々の問題は，式 (1) を最大にする拡散確率 Θ を求めることである．本推定問題に対しては，解を安定に求めるために，我々は EM アルゴリズムに基づいた推定法をすでに提案している [Saito 08].

我々の提案法の基本性能を評価するために，本論文では，すべてのリンクが同じ拡散確率 p をもつという最も単純な場合を考える．このような問題設定は，多くの従来研究でも採用されている [Kempe 03, Kimura 07, Kimura 09]．ここで行われる定式化は，そのような制限がないより一般の場合に対しても有効である．

3. 実験

3.1 実験設定

実験では，[Kimura 09] で使用された，社会ネットワークの顕著な特徴を多く持つ二つの大規模な実ネットワーク，ブログネットワークとウィキペディアネットワークを用いた．これらは双方向ネットワークである．ブログネットワークは 12,047 ノードと 79,920 有向リンクをもち，ウィキペディアネットワークは 9,481 ノードと 245,044 有向リンクをもっていた．本予備実験では，前にも述べたように，拡散確率 p がネットワーク上で一様であるという最も単純な場合を考えた．そして，p の値を，ブログネットワークでは $p = 0.1$，ウィキペディアネットワークでは $p = 0.01$ と設定した．影響度 $\{\sigma(v;p);\ v \in V\}$ の値は，パラメータ値 10,000 のボンドパーコレーション法 [Kimura 07] を用いて評価した．ここに，そのパラメータ値はボンドパーコレーション過程の試行回数を表している．影響度の平均値と標準偏差値は，ブログネットワークでは 87.5 と 131 であり，ウィキペディアネットワークでは 8.14 と 18.4 であった．

学習段階においては，訓練サンプルは，情報拡散結果 $D = \langle D(0), D(1), \cdots, D(T)\rangle$ であり，ランダムに選ばれた一つの初期アクティブノードからスタートするアクティブノード集合の時系列である．我々は拡散確率 p の推定に M 個の訓練サンプルを用いた．ここに，M はパラメータである．

3.2 比較法

ランキング上位のノード群の予測性能に関して，提案法を社会ネットワーク分析からの四つのヒューリスティクスと比較した．

“次数中心性”，“closeness 中心性” および “betweenness 中心性” は，ノードの影響力を測定する尺度として社会ネットワーク分析の分野で一般に用いられている [Wasserman 94]．ここに，ノード v の次数は v のリンク数として定義され，ノード v の closeness は v とネットワーク上の他ノードとの平均距離の逆数として定義され，そして，ノード v の betweenness は v を通るノードペア間の最短パスの数として定義される．

また，ウェブページのハイパーリンクネットワーク上で権威ある（影響力がある）ページを同定するための手法として，“PageRank 法”[Brin 98] がよく知られている．PageRank 法によって得られる “authoritativeness” により，ノードの影響力を測定することが自然に考られる．この手法はパラメータ λ をもっている．ここに，λ は，PageRank 法をランダムウェブサーファーモデルと見たとき，サーファーが一様ランダムに選んだウェブページにジャンプする確率を表している．実験では，典型的な設定 $\lambda = 0.15$[Ng 01] を用いた．

3.3 実験結果

まず，提案法による拡散確率の学習性能を評価した．p_0 を IC モデルの真の拡散確率とし，$\hat{p}$ を提案法により推定された拡散確率の値とする．学習性能は，誤差率，

$$\mathcal{E} = \frac{|p_0 - \hat{p}|}{p_0}$$

で評価した．表 1 は，訓練サンプル数 M に対する，$\mathcal{E}$ の平均値と括弧内にその標準偏差を示している．ここに，同じ実験を独立に 5 回実行した．十分な量の訓練データがあるとき，我々のアルゴリズムによる推定確率は真の確率に効率よく収束していくことが見て取れる．本結果は提案法の有効性を実証している．

表 1: 拡散確率の学習性能.

Results for the blog network

M	$\mathcal{E}$
20	0.036 (0.024)
40	0.018 (0.014)
60	0.016 (0.007)
80	0.009 (0.006)
100	0.006 (0.004)

Results for the Wikipedia network

M	$\mathcal{E}$
20	0.138 (0.081)
40	0.109 (0.066)
60	0.080 (0.041)
80	0.047 (0.018)
100	0.021 (0.013)

次に，ネットワーク $G = (V, E)$ から影響力が強いノード群を抽出するためのランキング法として，提案法を，次数法，betweenness 法，closeness 法および PageRank 法と比較した．任意の正整数 r ($\leq |V|$) に対して，$L_0(r)$ を真の上位 r 個のノード集合，$L(r)$ を与えられたランキング法による上位 r 個のノード集合とする．ランキング法の性能を，ランク r でのランキング類似度 $F(r)$,

$$F(r) = \frac{|L_0(r) \cap L(r)|}{r}.$$

により評価した．我々は，影響力が強いノード群を抽出することに興味があるので，上位ランクでのランキング類似度に焦点をあてた．図 1 と図 2 は，ブログネットワークに対する結果とウィキペディアネットワークに対する結果を，それぞれ示している．ここに，円印，三角印，ダイアモンド印，四角印およびアスタリスク印は，それぞれ，提案法，次数法，betweenness 法，closeness 法および PageRank 法に対して，ランク r の関数としてランキング類似度 $F(r)$ を表している．提案法に関しては，$M = 100$ の場合での 5 回の実験結果に対する平均値をプロットしている．提案法は，両方のネットワークに対して，他のヒューリスティクスよりもはるかに良い結果を示したということが見て取れる．これらの結果は，提案法の有効性を実証している．

4. 議論

まず，提案ランキング法は，情報拡散モデルである IC モデルに基づいた新たな中心性の概念を与えていると考えられる．実際，図 1, 2 から，提案法が上位にランキングしたノード群は，各従来法のそれらと大きく異なっていたことがわかる．すなわち，提案法は，もし過去の情報拡散データが入手できるならば，新たなタイプの社会ネットワーク分析を可能としている．もちろん，各従来法にはそれぞれのメリットと利用法があることに議論の余地はなく，我々の手法は，情報拡散という観点で異なるメリットをもつものとして，それら従来法に追加するものである．

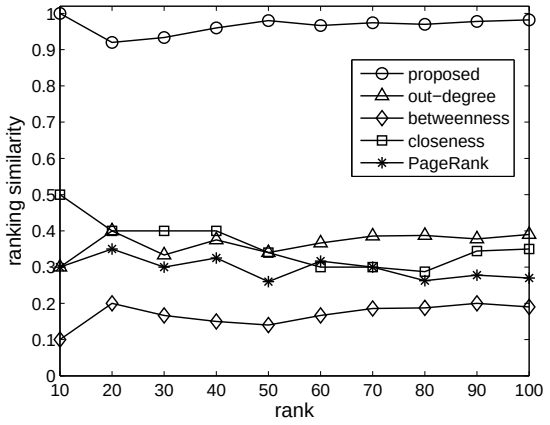

図 1: ブログネットワークでの影響力が強いノード群の抽出に関する性能比較.

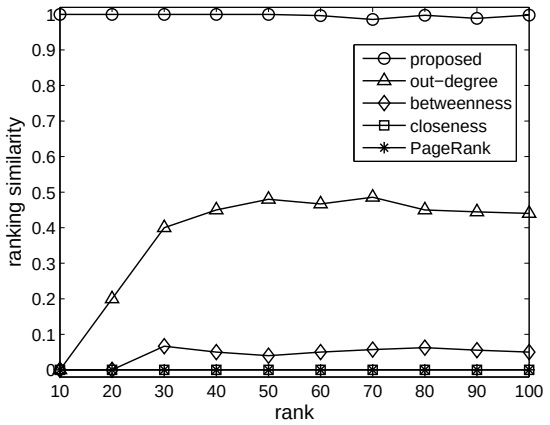

図 2: ウィキペディアネットワークでの影響力が強いノード群に抽出に関する性能比較.

次に，影響力が強いノード群を抽出することにおいて，なぜ拡散確率を知ることが重要であるのかを，単純な解析により説明する．もし拡散確率がノードランキングに全く影響しないならば，我々はその値に注意を払う必要がない．しかしながら，以下のような単純な解析により，拡散確率 p の値はノードランキングに影響するということが示される。まず，$\sigma(v; p)$ は，v の出次数がゼロでないなら，p に関して単調増大する非負関数であることに注意しよう．いま，図 3 に示すように，ネットワーク $G = (V, E)$ 上には，

$$(v, v_1), (v, v_2), (v, v_3) \in E,$$
$$(w, w_1), (w, w_2), (w_1, w_3), (w_2, w_3) \in E$$

なる 2 つのノード $v, w \in V$ があると仮定する．このとき，v の影響度 $\sigma(v; p)$ と w の影響度 $\sigma(w; p)$ は，ともに最大値は 3

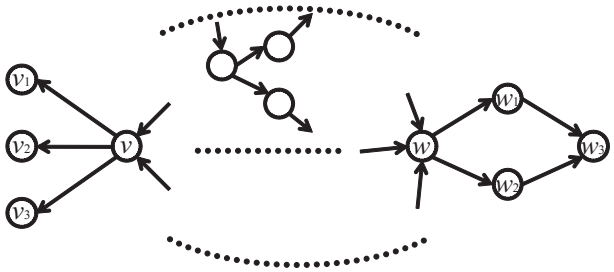

図 3: ネットワークの例.

であるが，それらは，

$$\begin{aligned}\sigma(v;p) &= 3p,\\ \sigma(w;p) &= 2p+(1-(1-p^2)^2) = 2p+2p^2-p^4\end{aligned}$$

と計算される [Kimura 06]．したがって，

$$\sigma(v;p)-\sigma(w;p) = p(1-p)(1-p-p^2)$$

が成り立つ．これより，

$$\begin{aligned}&\sigma(v;p) > \sigma(w;p) \quad \text{if } p < (-1+\sqrt{5})/2\\ &\sigma(v;p) \leq \sigma(w;p) \quad \text{otherwise}\end{aligned}$$

が成り立つ．直感的には，p の値が大きくなるにつれて，情報源ノードから 2 ステップで到達可能なノードがアクティブになる確率が，それから 1 ステップで到達可能なノードがアクティブになる確率よりも大きくなり，したがって，w がより大きな影響度をもつようになる．一般にネットワーク内にはこれらのような部分ネットワークが多く存在するので，拡散確率をできるだけ高精度に推定することは重要である．我々は，本論文で提案した手法が様々なタイプの社会ネットワーク分析に有用となりうると信じている．

さて，本論文で示した解析は，p がリンク集合 E のすべてのリンクに対して一つの値しかとらないという，最も単純な場合に対してのものであったが，しかしながら，我々の提案法は非常に一般的なものであることに注意しておく．より現実的な設定では，E を部分集合 $E_1, E_2, \cdots, E_N$ に分割して，各 E_n 内のすべてのリンクに対して異なる値 p_n を設定するということが可能である．例えば，ノード集合 V を，他に強い影響を及ぼすノード群とそうでないノード群の 2 つのグループに分割することが可能かもしれないし，または，他から影響を受けやすいノード群とそうでないノード群という別の 2 つのグループに分割することが可能かもしれない．さらには，ノード集合 V を複数のグループに分割することが可能であろう．もし，ノードのグループ化についての背景知識があれば，我々の手法はそれを最大限に利用することができる．これは人工知能アプローチの特徴の一つであるが，そのような背景知識を得ることは，社会ネットワークからの知識発見における重要な研究課題である．

5. まとめ

ネットワークトポロジーと情報拡散データが与えられたとき，その複雑な社会ネットワーク上で影響力が強いノードをランキングする手法を提案した．一般的な情報拡散モデルである IC モデルを用いて各リンクの拡散確率を，アクティブノード集合の時系列として観測された過去の情報拡散履歴から尤度最大化問題として推定する，EM アルゴリズムに基づいた効率的な手法を導いた．拡散確率がネットワーク上で一様であるという最も単純な設定の下で，二つの実ネットワークを用いた実験により提案法の有効性を実証した．まず，訓練データとして用いられる観測時系列データがある程度あるならば，提案法が拡散確率を高精度に推定できることを示した．次に，影響力が強いノードのランキングを，よく知られたヒューリスティクス（次数中心性，closeness 中心性，betweenness 中心性，authoritativeness）に基づいた手法よりも，提案法がはるかに高精度に予測できることを示した．

謝辞

本研究は，科学研究費補助金基盤研究 (C) (No. 20500147) の補助を受けた．

参考文献

[Backstrom 06] Backstrom, L., Huttenlocher, D., Kleinberg, J., and Lan, X.: Group formation in large social networks: Membership, growth, and evolution, in *KDD'06*, pp. 44–54 (2006)

[Brin 98] Brin, S. and Page, L.: The anatomy of a large-scale hypertextual web search engine, in *WWW'98*, pp. 107–117 (1998)

[Gruhl 04] Gruhl, D., Guha, R., Liben-Nowell, D., and Tomkins, A.: Information diffusion through blogspace, in *WWW'04*, pp. 107–117 (2004)

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

[Kimura 06] Kimura, M. and Saito, K.: Tractable models for information diffusion in social networks, in *PKDD'06*, pp. 259–271 (2006)

[Kimura 07] Kimura, M., Saito, K., and Nakano, R.: Extracting influential nodes for information diffusion on a social network, in *AAAI'07*, pp. 1371–1376 (2007)

[Kimura 09] Kimura, M., Saito, K., and Motoda, H.: Blocking links to minimize contamination spread in a social network, *ACM Transactions on Knowledge Discovery from Data*, Vol. 3, No. 2, Article 9 (2009)

[Leskovec 06] Leskovec, J., Adamic, L., and Huberman, B. A.: The dynamics of viral marketing, in *EC'06*, 228–237 (2006)

[Ng 01] Ng, A. Y., Zheng, A. X., and Jordan, M. I.: Link analysis, eigenvectors and stability, in *IJCAI'01*, pp. 903–910 (2001)

[Saito 08] Saito, K., Nakano, R., and Kimura, M.: Prediction of information diffusion probabilities for independent cascade model, in *KES'08*, pp. 67–75 (2008)

[Wasserman 94] Wasserman, S. and Faust, K.: Social network analysis, Cambridge University Press (1994)