原著論文

Original Paper

社会ネットワークにおける有力ノード抽出のための情報拡散モデルの学習

Learning Information Diffusion Model for Extracting Influential Nodes in a Social Network

木村 昌弘
Masahiro Kimura
龍谷大学理工学部電子情報学科
Department of Electronics and Informatics, Ryukoku University
kimura@rins.ryukoku.ac.jp

斉藤 和巳
Kazumi Saito
静岡県立大学経営情報学部
School of Administration and Informatics, University of Shizuoka
k-saito@u-shizuoka-ken.ac.jp

中野 良平
Ryohei Nakano
中部大学工学部情報工学科
Department of Computer Science, Chubu University
nakano@cs.chubu.ac.jp

元田 浩
Hiroshi Motoda
大阪大学産業科学研究所
Institute of Scientific and Industrial Research, Osaka University
motoda@ar.sanken.osaka-u.ac.jp

keywords: independent cascade model, diffusion probability estimation, learning algorithm, social network analysis, node ranking

Summary

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 when there is a reasonable amount of information diffusion data, the accuracy of the probability is outstandingly good, and 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

近年，複雑ネットワークの構造や機能に対する関心が高まっている [Newman 03a]．機能という観点では，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ので，社会ネットワークは情報拡散媒体として重要な役割を果たすと言える．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，情報を普及させる重要媒体として，最近，社会ネットワークの研究が注目されている [Domingos 05, Backstrom 06, Leskovec 06a, Watts 07].

ネットワーク空間における情報拡散の研究には，ブログ空間におけるトピック伝搬パターンの研究 [Adar 05, Gruhl 04] や，e メールネットワークを通じたコンピュータウィルスの拡散などのようなネットワークを通じた好ましくない情報の拡散に対して，その広がりを縮小するためのノード除去戦略の研究 [Albert 00, Newman 02] などがある．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 [Goldenberg 01, Kempe 03]．IC モデルを用いて，情報の広がりを最大にするには指定された数のどのノード群に最初にその情報を提供すればよいかという，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．また，最近，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広がりを最小化するという，影響最大化問題と反対の問題も，IC モデルを用いて研究されている [Kimura 09]．さらに，複雑ネットワークを IC モデルに従う情報拡散過程という動的な角度から理解するためのネットワーク可視化法も研究されている [Saito 08]．本論

文でも，与えられた社会ネットワークに対して，IC モデルに従う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有力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．我々は，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 モデルの下での "影響度" に関しノードをランキングすることにより，影響力が強い有力ノード群を抽出するという手法を提案する．ところで，IC モデルは前もって指定すべきパラメータをもっている．具体的には，ネットワーク上の各リンクに対し，前もって "拡散確率" を指定しなければならない．そこで，情報拡散データの観測集合を得る尤度を最大化することにより，IC モデルの拡散確率を推定する手法を提案する．十分な数の統計的に独立な情報拡散履歴が観測されているという場合に，二つの実ネットワークトポロジー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 IC モデルを有力ノード群抽出のためのランキングに用い，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分野の四つのヒューリスティクスによる結果とも比較する．

以下，論文の構成は次のとおりである．2 節では，IC モデルおよびノードの影響度の概念を定義する．3 節では，提案法を機械学習問題として定式化して詳述する．4 節では，実験設定および実験結果を述べる．5 節では，IC モデルの拡散確率推定の意義などについて議論する．6 節はまとめである．

2. 情報拡散モデル

本節では [Kempe 03] に従って，ネットワーク G 上の IC モデルを定義し，IC モデルに基づいたノードの影響度を定義する．

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．また，ノード v から w へのリンク (v,w) が存在するならば，w を v の "子ノード" と呼び，v を w の "親ノード" と呼ぶ．ノード v の子ノード全体の集合を，

$$F(v) \;=\; \{w \in V;\ (v,w) \in E\}$$

とし，ノード v の親ノード全体の集合を，

$$B(v) \;=\; \{u \in V;\ (u,v) \in E\}$$

とする．

IC モデルでは，各リンク $e=(v,w)$ 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．IC モデルに基づく情報拡散では，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は変化するが，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 S が与えられたとき，IC モデルの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，本拡散過程は終了する．

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，$\varphi(S;\Theta)$ は確率変数であることに注意する．我々は，$\varphi(S;\Theta)$ 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 $\sigma(v;\Theta)$ と記述し，$\sigma(v;\Theta)$ 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と言えることに注意する．

3. 提　案　法

3·1 有力ノード抽出法

本論文では，ネットワーク上での情報拡散の数理モデルとして IC モデルを仮定する．我々は，与えられたネットワーク $G=(V,E)$ に対し，IC モデルに従う情報拡散で影響力が強い有力ノード群を抽出する手法として，影響度 $\sigma(v;\Theta_0)$ に関するノードランキング法を提案する．ここに，Θ_0 は G における IC モデルの真の拡散確率ベクトルである．しかしながら，実際問題においては真の拡散確率ベクトルは入手不可能である．そこで，我々は，過去の独立な M 個の情報拡散履歴

$$\{D_m;\ m=1,\cdots,M\}$$

から推定した拡散確率ベクトル $\widehat{\Theta}$ を用いる手法を提案する．各情報拡散履歴 D_m はアクティブノード集合の時系列

$$D_m = \langle D_m(0), D_m(1), \cdots, D_m(T_m)\rangle$$

として観測される．ここに，$D_m(t)$ は時刻 t で初めてアクティブになったノード全体の集合であり，T_m は第 m 情報拡散履歴の最終時刻である．本問題設定は，確率過程のモデルが仮定されたときに，それから生成される観測時系列データに基づいてそのモデルパラメータを推定するというものであり，標準的なパラメトリック分布を仮定する統計モデリングと同じ設定である．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．

3·2 拡散確率推定法

§ 1 尤度関数

まず，拡散確率ベクトル $\Theta = (p_{v,w})$ に関する一つの情報拡散履歴 $D_m = \langle D_m(0), \cdots, D_m(T_m)\rangle$ の尤度関数 $\mathcal{L}(\Theta; D_m)$ を以下に定義する．ここに，$D_m(T_m+1) = \emptyset$ であることに注意する．

今，$v \in D_m(t), e = (v,w) \in E, w \in D_m(t+1) \cap F(v)$ 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，$(D_m(t) \cap B(w)) \setminus \{v\} \neq \emptyset$ ならば，他のノード $v' \in D_m(t) \cap B(w)$ 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 $t+1$ で初めてアクティブになる確率 $P_{m,t+1}(w;\Theta)$ は，

$$P_{m,t+1}(w;\Theta) = 1 - \prod_{v \in B(w)\cap D_m(t)} (1 - p_{v,w}) \quad (1)$$

で与えられる．ここに，$w \in D_m(t+1)$ ならば $D_m(t) \cap B(w) \neq \emptyset$ であることに注意する．さて，

$$C_m(t) = D_m(0) \cup \cdots \cup D_m(t)$$

と定義する．$C_m(t)$ 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．$v \in D_m(t)$ で $w \in F(v) \setminus C_m(t+1)$ 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，$v \in D_m(t)$ で $w \in F(v) \cap C_m(t)$ であるとき，または $v \notin C_m(T_m)$ 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，尤度関数 $\mathcal{L}(\Theta; D_m)$ は，式 (1) に基づいて，

$$\mathcal{L}(\Theta; D_m) = \left(\prod_{t=0}^{T_m-1} P_t^{+}(D_m;\Theta)\right)\left(\prod_{t=0}^{T_m} P_t^{-}(D_m;\Theta)\right) \quad (2)$$

と定義できる．ただし，

$$P_t^{+}(D_m;\Theta) = \prod_{w \in D_m(t+1)} P_{m,t+1}(w;\Theta)$$

$$P_t^{-}(D_m;\Theta) = \prod_{v \in D_m(t)} \prod_{w \in F(v)\setminus C_m(t+1)} (1 - p_{v,w})$$

である．

さて，$\{D_m; 1 \le m \le M\}$ を M 個の独立な情報拡散履歴とする．このとき，Θ に関する目的関数は，

$$\mathcal{J}(\Theta) = \sum_{m=1}^{M} \log \mathcal{L}(\Theta; D_m) \quad (3)$$

と定義できる．よって，我々の問題は，式 (3) を最大にする拡散確率ベクトル Θ を求めることである．

§ 2 学習アルゴリズムの導出

本節では，$\mathcal{J}(\Theta)$ を最大にする拡散確率ベクトル $\Theta = (p_{v,w})$ の推定法として，EM アルゴリズム [Dempster 77] における目的関数の最大化と類似した算法で逐次反復アルゴリズムを導出する．すなわち，まず，隠れ変数 $\{a_{v,w}\}$ を導入し，次に，$\bar{\Theta} = (\bar{p}_{v,w})$ を更新前の拡散確率ベクトル Θ の値として Q 関数 $Q(\Theta|\bar{\Theta})$ を定義し，そして，その最適化問題を解いてパラメータ更新則を導出する．

まず，第 m 情報拡散履歴 D_m の時刻 $t+1$ における，ノード $w \in D_m(t+1)$ を任意に固定して考える．

Part 1) 隠れ変数の導入: ノード $v \in B(w) \cap D_m(t)$ からリンク (v,w) を通じてノード w へ情報が拡散したならば $a_{v,w} = 1$ とし，そうでないならば $a_{v,w} = 0$ とすることにより，$B(w) \cap D_m(t)$ から $\{0,1\}$ への写像 a_w を定義する．ただし，$v \in B(w) \cap D_m(t)$ に対して，$a_w(v) = a_{v,w}$ である．以降，写像 a_w を，

$$a_w = (a_{v,w})_{v \in B(w)\cap D_m(t)}$$

とベクトルで表示することにする．各 $a_{v,w}$ は情報拡散履歴 D_m からは知ることができない隠れ変数であることに注意する．写像 $a_w = (a_{v,w})$ を用いると，式 (1) の確率 $P_{m,t+1}(w;\Theta)$ は，

$$P_{m,t+1}(w;\Theta) = \sum_{a_w \in \mathcal{A}_w} \prod_{v \in B(w)\cap D_m(t)} {p_{v,w}}^{a_{v,w}} (1 - p_{v,w})^{1-a_{v,w}} \quad (4)$$

と表現できる．ここに，$\mathcal{A}_w$ は少なくとも一つの $v \in B(w) \cap D_m(t)$ において 1 となるような $B(w) \cap D_m(t)$ から $\{0,1\}$ への写像全体の集合である．

Part 2) Q 関数の定義: ここで，$q(a_w \,|\, w, \Theta)$ を，

$$q(a_w \,|\, w, \Theta) = \frac{\prod_{v \in B(w)\cap D_m(t)} {p_{v,w}}^{a_{v,w}} (1 - p_{v,w})^{1-a_{v,w}}}{P_{m,t+1}(w;\Theta)} \quad (5)$$

で定義する．

$$q(a_w \,|\, w, \Theta) > 0, \quad \sum_{a_w \in \mathcal{A}_w} q(a_w \,|\, w, \Theta) = 1 \quad (6)$$

が成り立つことに注意する．このとき，式 (4), (5) より，

$$\log P_{m,t+1}(w;\Theta) - \log P_{m,t+1}(w;\bar{\Theta}) = \quad (7)$$

$$\log\left(\sum_{a_w \in \mathcal{A}_w} \left\{ q(a_w \,|\, w, \bar{\Theta}) \cdot \prod_{v \in B(w)\cap D_m(t)} \frac{{p_{v,w}}^{a_{v,w}} (1 - p_{v,w})^{1-a_{v,w}}}{{\bar{p}_{v,w}}^{a_{v,w}} (1 - \bar{p}_{v,w})^{1-a_{v,w}}} \right\}\right)$$

が成り立つ．また，式 (6) に注意すると，式 (7) において Jensen の不等式より，

$$\log P_{m,t+1}(w;\Theta) - \log P_{m,t+1}(w;\bar{\Theta}) \\ \geq Q_{m,t+1,w}(\Theta \,|\, \bar{\Theta}) \tag{8}$$

が成り立つ．ここに，

$$Q_{m,t+1,w}(\Theta \,|\, \bar{\Theta}) = \\ \sum_{a_w \in \mathcal{A}_w} q(a_w \,|\, w, \bar{\Theta})\; g_{m,t,w}(\Theta, \bar{\Theta}) \tag{9}$$

である．ただし，

$$g_{m,t,w}(\Theta, \bar{\Theta}) = \\ \sum_{v \in B(w) \cap D_m(t)} \{a_{v,w} \log p_{v,w} \\ + (1 - a_{v,w}) \log(1 - p_{v,w})\} \\ - \sum_{v \in B(w) \cap D_m(t)} \{a_{v,w} \log \bar{p}_{v,w} \\ + (1 - a_{v,w}) \log(1 - \bar{p}_{v,w})\}$$

である．$Q_{m,t+1,w}(\bar{\Theta} \,|\, \bar{\Theta}) = 0$ に注意しておく．式 (9) において，任意の $v \in B(w) \cap D_m(t)$ に対する，

$$\sum_{a_w \in \mathcal{A}_w} a_{v,w} q(a_w \,|\, w, \bar{\Theta})$$

を計算しよう．式 (5) を用いて計算すると，$a_{v,w} \in \{0, 1\}$ だから，

$$\sum_{a_w \in \mathcal{A}_w} a_{v,w} q(a_w \,|\, w, \bar{\Theta}) \\ = \sum_{a_w \in \mathcal{A}_w} \left\{ a_{v,w} \prod_{u \in B(w) \cap D_m(t)} \bar{p}_{u,w}{}^{a_{u,w}} \\ \cdot (1 - \bar{p}_{u,w})^{1 - a_{u,w}} \right\} \Big/ P_{m,t+1}(w;\bar{\Theta}) \\ = \sum_{\tilde{a}_w \in \mathcal{A}_w(v)} \left\{ \bar{p}_{v,w} \prod_{u \in (B(w) \cap D_m(t)) \setminus \{v\}} \bar{p}_{u,w}{}^{\tilde{a}_{u,w}} \\ \cdot (1 - \bar{p}_{u,w})^{1 - \tilde{a}_{u,w}} \right\} \Big/ P_{m,t+1}(w;\bar{\Theta}) \tag{10}$$

が成り立つことが容易にわかる．ここに，$\mathcal{A}_w(v)$ は，$(B(w) \cap D_m(t)) \setminus \{v\}$ から $\{0, 1\}$ への写像全体の集合であり，$\tilde{a}_w$ のベクトル表示は，

$$\tilde{a}_w = (\tilde{a}_{u,w})_{u \in (B(w) \cap D_m(t)) \setminus \{v\}}$$

である．すなわち，$\tilde{a}_{u,w} \in \{0, 1\}$ である．ところで，

$$\sum_{\tilde{a}_w \in \mathcal{A}_w(v)} \left(\prod_{u \in (B(w) \cap D_m(t)) \setminus \{v\}} \bar{p}_{u,w}{}^{\tilde{a}_{u,w}} \\ \cdot (1 - \bar{p}_{u,w})^{1 - \tilde{a}_{u,w}} \right) \\ = \prod_{u \in (B(w) \cap D_m(t)) \setminus \{v\}} \{\bar{p}_{u,w} + (1 - \bar{p}_{u,w})\} = 1$$

だから，式 (10) より，

$$\sum_{a_w \in \mathcal{A}_w} a_{v,w} q(a_w \,|\, w, \bar{\Theta}) = \frac{\bar{p}_{v,w}}{P_{m,t+1}(w;\bar{\Theta})} \tag{11}$$

が成り立つ．よって，式 (9), (11) より，

$$Q_{m,t+1,w}(\Theta \,|\, \bar{\Theta}) = \\ \sum_{v \in B(w) \cap D_m(t)} \left\{ \frac{\bar{p}_{v,w}}{P_{m,t+1}(w;\bar{\Theta})} \log p_{v,w} \\ + \left(1 - \frac{\bar{p}_{v,w}}{P_{m,t+1}(w;\bar{\Theta})} \right) \log(1 - p_{v,w}) \right\} + f_1 \tag{12}$$

が成り立つ．ここに，f_1 は Θ に依存しない定数である．

Part 3) パラメータ更新則の導出: さて，M 個の独立な情報拡散履歴全体 $\{D_m; 1 \leq m \leq M\}$ を考え，$\mathcal{J}(\Theta)$ を最大にする拡散確率ベクトル $\Theta = (p_{v,w})$ を推定する逐次反復アルゴリズムを導く．任意のリンク $(v, w) \in E$ に対して，

$$\mathcal{M}^+_{v,w} = \\ \{m \in \{1, \cdots, M\};\ \exists t \in [0, T_m - 1] \text{ such that } \\ w \in D_m(t+1), v \in B(w) \cap D_m(t)\}, \tag{13}$$

$$\mathcal{M}^-_{v,w} = \\ \{m \in \{1, \cdots, M\};\ \exists t \in [0, T_m] \text{ such that } \\ v \in D_m(t), w \in F(w) \setminus C_m(t+1)\} \tag{14}$$

を定義する．ここに，$\mathcal{M}^+_{v,w}$ はリンク (v, w) を通じて情報拡散が起こった可能性がある情報拡散履歴全体の集合を，$\mathcal{M}^-_{v,w}$ はリンク (v, w) を通じて情報拡散が起こらなかった情報拡散履歴全体の集合を，それぞれ表していることに注意する．任意のリンク $(v, w) \in E$ に対して $p_{v,w} > 0$ であるので，我々は，十分な数の独立な情報拡散履歴が観測されているとして，

$$\left|\mathcal{M}^+_{v,w}\right| : 十分大, \quad \forall (v, w) \in E \tag{15}$$

と仮定する．以下に $\Theta = (p_{v,w})$ の更新則を導く．

式 (2), (3) より，

$$\mathcal{J}(\Theta) - \mathcal{J}(\bar{\Theta}) = \\ \sum_{m=1}^{M} \left(\sum_{t=0}^{T_m - 1} h^+_{m,t}(\Theta, \bar{\Theta}) + \sum_{t=0}^{T_m} h^-_{m,t}(\Theta, \bar{\Theta}) \right)$$

である．ただし，

$$h^+_{m,t}(\Theta, \bar{\Theta}) = \sum_{w \in D_m(t+1)} \{\log P_{m,t+1}(w;\Theta) \\ - \log P_{m,t+1}(w;\bar{\Theta})\},$$

$$h^-_{m,t}(\Theta, \bar{\Theta}) = \\ \sum_{v \in D_m(t)} \sum_{w \in F(v) \setminus C_m(t+1)} \{\log(1 - p_{v,w}) \\ - \log(1 - \bar{p}_{v,w})\}$$

である．よって，式 (8) より，

$$\mathcal{J}(\Theta) - \mathcal{J}(\bar{\Theta}) \ \geq \ Q(\Theta \,|\, \bar{\Theta}) \tag{16}$$

が成り立つ．ただし，

$$Q(\Theta \,|\, \bar{\Theta}) = \tag{17}$$
$$f_2 + \sum_{m=1}^{M} \left\{ \sum_{t=0}^{T_m - 1} \sum_{w \in D_m(t+1)} Q_{m,t+1,w}(\Theta \,|\, \bar{\Theta}) \right.$$
$$\left. + \sum_{t=0}^{T_m} \sum_{v \in D_m(t)} \sum_{w \in F(v) \setminus C_m(t+1)} \log(1 - p_{v,w}) \right\}$$

であり，f_2 は Θ に依存しない定数である．また，

$$Q(\bar{\Theta} \,|\, \bar{\Theta}) \ = \ 0 \tag{18}$$

が成り立つことに注意しておく．よって，式 (16), (18) より，Q 関数 $Q(\Theta \,|\, \bar{\Theta})$ を最大にする Θ を求めることにより，目的関数 $\mathcal{J}(\Theta)$ を最大化するためのパラメータ更新則を導くことができる．$Q(\Theta \,|\, \bar{\Theta})$ は，式 (12), (17) より，

$$Q(\Theta \,|\, \bar{\Theta}) = \tag{19}$$
$$f_3 + \sum_{m=1}^{M} \left[\sum_{t=0}^{T_m - 1} \sum_{w \in D_m(t+1)} \right.$$
$$\sum_{v \in B(w) \cap D_m(t)} \left\{ \frac{\bar{p}_{v,w}}{P_{m,t+1}(w; \bar{\Theta})} \log p_{v,w} \right.$$
$$\left. + \left(1 - \frac{\bar{p}_{v,w}}{P_{m,t+1}(w; \bar{\Theta})} \right) \log(1 - p_{v,w}) \right\}$$
$$\left. + \sum_{t=0}^{T_m} \sum_{v \in D_m(t)} \sum_{w \in F(v) \setminus C_m(t+1)} \log(1 - p_{v,w}) \right]$$

となる．ただし，f_3 は Θ に依存しない定数である．式 (15), (19) より，関数 $Q(\Theta \,|\, \bar{\Theta})$ は，$\{\log p_{v,w}, \ \log(1 - p_{v,w});\ (v,w) \in E\}$ の線形結合でありかつそれらの係数は正であるので，唯一の極値をもちその極値は最大値でもある．したがって，$Q(\Theta \,|\, \bar{\Theta})$ の極値を与える Θ を求めれば良い．式 (13), (14), (19) より，任意の $(v,w) \in E$ に対して，

$$\frac{\partial Q}{\partial p_{v,w}} =$$
$$\sum_{m \in \mathcal{M}^{+}_{v,w}} \left\{ \frac{\bar{p}_{v,w}}{P_{m,t^{+}_{m}(v,w)+1}(w; \bar{\Theta})} \frac{1}{p_{v,w}} \right.$$
$$\left. - \left(1 - \frac{\bar{p}_{v,w}}{P_{m,t^{+}_{m}(v,w)+1}(w; \bar{\Theta})} \right) \frac{1}{1 - p_{v,w}} \right\}$$
$$- \left| \mathcal{M}^{-}_{v,w} \right| \frac{1}{1 - p_{v,w}} \tag{20}$$

である．ただし，$|S|$ は集合 S の要素数であり，$m \in \mathcal{M}^{+}_{v,w}$ に対して $t^{+}_{m}(v,w)$ は，

$$t^{+}_{m}(v,w) \in [0, T_m], \ w \in D_m(t^{+}_{m}(v,w) + 1),$$
$$v \in B(w) \cap D_m(t^{+}_{m}(v,w))$$

なるものである．ゆえに，目的関数 $\mathcal{J}(\Theta)$ を最大化するためのパラメータ更新則は，任意の $(v,w) \in E$ に対して，

$$p_{v,w} = \tag{21}$$
$$\frac{1}{\left|\mathcal{M}^{+}_{v,w}\right| + \left|\mathcal{M}^{-}_{v,w}\right|} \sum_{m \in \mathcal{M}^{+}_{v,w}} \frac{\bar{p}_{v,w}}{P_{m,t^{+}_{m}(v,w)+1}(w; \bar{\Theta})}$$

となる．ただし，式 (21) において $P_{m,t^{+}_{m}(v,w)+1}(w; \bar{\Theta})$ の値は式 (1) により計算する．式 (20) より，式 (21) の $\Theta = (p_{v,w})$ が関数 $Q(\Theta \,|\, \bar{\Theta})$ の極値を与えることは容易に示される．

§ 3　推定アルゴリズム

我々は，ネットワーク $G = (V, E)$ における IC モデルの拡散確率ベクトル $\Theta = (p_{v,w})$ を，M 個の情報拡散履歴 $\{D_m; m = 1, \cdots, M\}$ から，以下のようなアルゴリズムで推定する．

次の逐次反復アルゴリズムを K 回実行し，出力された K 個の拡散確率ベクトル Θ のうち，$\mathcal{J}(\Theta)$ の値が最大となるものを，推定拡散確率ベクトル $\hat{\Theta}$ として出力する．ここに，自然数 K は制御パラメータである．

1. 拡散確率ベクトル $\bar{\Theta} = (\bar{p}_{v,w}) \in (0,1)^{|E|}$ をランダムに生成する．
2. 式 (21) により，$\bar{\Theta}$ から更新拡散確率ベクトル $\Theta = (p_{v,w})$ を計算する．
3. $\|\Theta - \bar{\Theta}\| < \varepsilon$ ならば，反復を終了し，Θ を出力する．
4. $\bar{\Theta} \leftarrow \Theta$ としてステップ **2** に戻る．

ここに，ε は終了条件を制御する正数である．

提案法の基本性能を評価するために，本論文では以降，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．もちろん，提案法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りうるが，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]．また，このように拡散確率をすべてのリンクで同じ値に設定することにより，社会ネットワーク分析分野におけるネットワーク構造のみに基づいて有力ノードをランキングする代表的な既存法と，提案ランキング法を公平に比較することが可能になると考えられる．ところで，同一の情報拡散モデル（IC モデル）に従う，非常に多数の独立な情報拡散履歴 $\{D_m; m = 1, \cdots, M\}$ が観測されるという仮定は，現実的でないと考えられる．すなわち，式 (15) は大規模ネットワークにおいては一般には成立しないと考えられる．特に，ここでのように拡散確率を一様に設定することで，限られた数の独立な情報拡散履歴から，すなわち，式 (15) を仮定することなく，それらを生成した IC モデルを推定するという，より現実的な問題設定が可能となることにも注意しておく．拡散確率設定の一般化に関しては，5 章で議論する．

4. 実　　験

4·1 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トポロジーがもつ顕著な特徴 [Newman 03b] を持つ二つの大規模な実ネットワークトポロジー，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 $245,044$ 有向リンクをもつ．本実験では，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 [Kempe 03, Leskovec 06b] に従って，p の値を比較的小さい値に設定した．特に，p の値を $1/\bar{d}$ よりも小さい値に設定した．ここに，$\bar{d}$ はネットワークの平均出次数である．$1/\bar{d}$ の値は，ブログネットワークでは 0.15，ウィキペディアネットワークでは 0.03 であったので，我々は，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．

我々は，拡散確率 p の推定に M 個の情報拡散履歴を訓練サンプルとして用いた．ここに，M はパラメータである．各情報拡散履歴は，ランダムに選ばれた一つの初期アクティブノード $D_m(0) = \{v_m\}$ から真の IC モデルの拡散過程により生成される，アクティブノード集合の時系列 $\{D_m = \langle D_m(0), \cdots, D_m(T_m)\rangle\}$ とした．また，提案法の推定アルゴリズムにおける制御パラメータ K と ε は，それぞれ，$K = 5, \varepsilon = 10^{-12}$ と設定した．

4·2 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分野における四つのヒューリスティクスと比較した．

“出次数中心性”，“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 の出次数は v からのリンク数として定義され，ノード v の closeness は v とネットワーク内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も自然だと考えられる．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．

表 1　提案法における拡散確率の推定性能．

Results for the blog dataset

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 dataset

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)

4·3 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．

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，出次数法，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，丸印，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，出次数法，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．

5. 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．

次に，IC モデルに基づいた有力ノード群抽出において，拡散確率の値を知ることが重要であることを単純な解析により説明する．もし拡散確率の値がノードランキングに全く影響しないならば，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．ネットワーク $G=(V,E)$ 内には，図 3 に示すように次の条件を満たす二つのノード $v_0, w_0 \in V$ がある

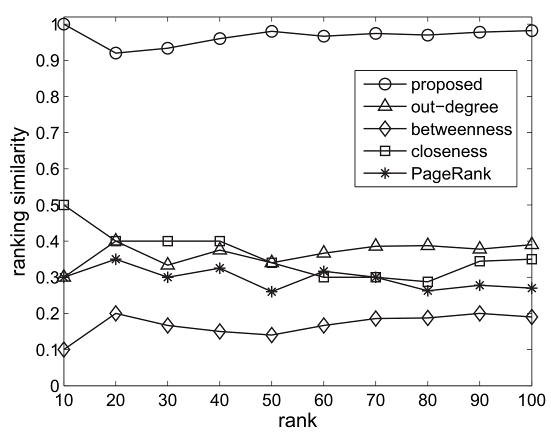


図 1　ブログネットワークでの有力ノード群抽出に関する性能比較.

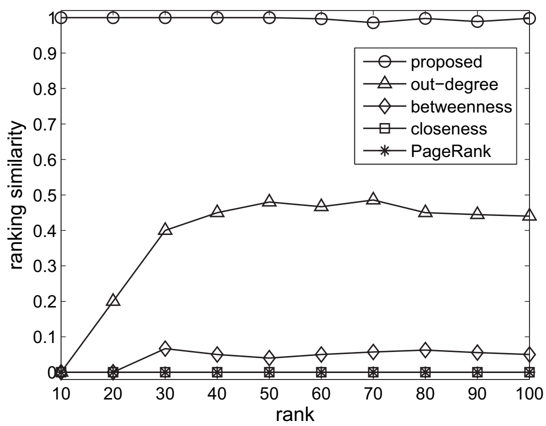


図 2　ウィキペディアネットワークでの有力ノード群抽出に関する性能比較.

と仮定する．

$$(v_0,v_1),(v_0,v_2),(v_0,v_3)\in E,$$
$$(w_0,w_1),(w_0,w_2),(w_1,w_3),(w_2,w_3)\in E.$$

このとき，v_0 の影響度 $\sigma(v_0;p)$ と w_0 の影響度 $\sigma(w_0;p)$ は，ともに最大値は 3 であるが，それらは，

$$\sigma(v_0;p) = 3p,$$
$$\sigma(w_0;p) = 2p+(1-(1-p^2)^2) = 2p+2p^2-p^4$$

と計算される [Kimura 06]．したがって，

$$\sigma(v_0;p)-\sigma(w_0;p)=p(1-p)(1-p-p^2)$$

が成り立つ．これより，

$$\sigma(v_0;p)>\sigma(w_0;p) \quad \text{if } p<(-1+\sqrt{5})/2$$
$$\sigma(v_0;p)\le\sigma(w_0;p) \quad \text{otherwise}$$

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_0 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

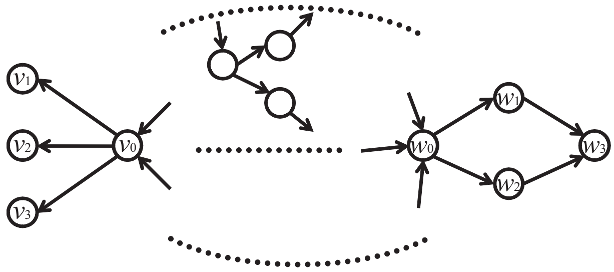


図 3　ネットワークの例.

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 ごとに異なる値 p_n を設定するということが考えられる．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二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二つのグループに分割することが可能かもしれない．類似の考えで，ノード集合 V を複数のグループに分割することも可能であろう．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．

6. 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（出次数中心性，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．

本研究では、社会ネットワークにおける情報拡散現象の最も基本的な数理モデルの一つであり，[Goldenberg 01, Kempe 03] など多様な研究が行われている IC モデルを，情報拡散モデルとして仮定した．もちろん，大規模社会ネットワークにおける現実の情報拡散現象は極めて複雑である．どのようなコミュニティが形成されるかにより，どのように拡散するかが決まる．IC モデルは，伝染病蔓延の基本確率モデルである SIR モデル [Newman 03a] と見なされるものであるが，現実の情報拡散現象を記述するには，「あるノード対の相互作用が特定時刻の 1 回に限られる」など仮定の厳しいモデルである．それらを緩和する研究として，時間遅れでの相互作用を許容するモデル [Gruhl 04] や複数回の相互作用を許容するモデル [Watts 07] など，多様なモデルが提案されているが，これらの研究でも基本的な IC モデルを土台としている．もちろん，情報拡散データからモデルパラメータを推定する我々の研究においても，上記のようなより現実的なモデルへの適用法を確立し，その有効性を示していくことは今後の重要な課題である．しかしながら，最も基本的な IC モデルでの手法を確立し評価することは，その第一歩として重要と考えられる．このようなステップは，複雑ネットワーク研究と機械学習研究を融合する新たな研究パラダイム構築のために必要不可欠であり，本研究によりその有望性が示せたことは十分に意義があると信じている．

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 ◇

[Adar 05] Adar, E. and Adamic, L.: Tracking information epidemics in blogspace, *Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence (WI'05)*, pp. 207–214 (2005)

[Albert 00] Albert, R., Jeong, H., and Barabási, A. L.: Error and attack tolerance of complex networks, *Nature*, Vol. 406, No. 6794, pp. 378–382 (2000)

[Backstrom 06] Backstrom, L., Huttenlocher, D., Kleinberg, J., and Lan, X.: Group formation in large social networks: Membership, growth, and evolution, *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2006)*, pp. 44–54 (2006)

[Brin 98] Brin, S. and Page, L.: The anatomy of a large-scale hypertextual web search engine, *Computer Networks and ISDN Systems*, Vol. 30, No. 1–7, pp. 107–117 (1998)

[Dempster 77] Dempster, A. P., Laird, N. M., and Rubin, D. B.: Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society B*, Vol. 39, No. 1, pp. 1–38 (1977)

[Domingos 05] Domingos, P.: Mining social networks for viral marketing, *IEEE Intelligent Systems*, Vol. 20, No. 1, pp. 80–82 (2005)

[Goldenberg 01] Goldenberg, J., Libai, B., and Muller, E.: Talk of the network: A complex systems look at the underlying process of word-of-mouth, *Marketing Letters*, Vol. 12, No. 3, pp. 211–223 (2001)

[Gruhl 04] Gruhl, D., Guha, R., Liben-Nowell, D., and Tomkins, A.: Information diffusion through blogspace, *Proceedings of the 13th International World Wide Web Conference (WWW 2004)*, pp. 107–117 (2004)

[Kempe 03] Kempe, D., Kleinberg, J., and 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)*, pp. 137–146 (2003)

[Kimura 06] Kimura, M. and Saito, K.: Tractable models for information diffusion in social networks, *Proceedings of the 10th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD 2006)*, LNAI 4213, pp. 259–271 (2006)

[Kimura 07] Kimura, M., Saito, K., and Nakano, R.: Extracting influential nodes for information diffusion on a social network, *Proceedings of the 22nd AAAI Conference on Artificial Intelligence (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, pp. 1–23 (2009)

[Leskovec 06a] Leskovec, J., Adamic, L., and Huberman, B. A.: The dynamics of viral marketing, *Proceedings of the 7th ACM Conference on Electronic Commerce (EC'06)*, pp. 228–237 (2006)

[Leskovec 06b] Leskovec, J., Singh, A., and Kleinberg, J.: Patterns of influence in a recommendation network, *Proceedings of the 10th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2006)*, LNAI 3918, pp. 380–389 (2006)

[Newman 02] Newman, M. E. J., Forrest, S., and Balthrop, J.: Email networks and the spread of computer viruses, *Physical Review E*, Vol. 66, No. 3, Article 035101, pp. 1–4 (2002)

[Newman 03a] Newman, M. E. J.: The structure and function of complex networks, *SIAM Review*, Vol. 45, No. 2, pp. 167–256 (2003)

[Newman 03b] Newman, M. E. J. and Park, J.: Why social networks are different from other types of networks, *Physical Review E*, Vol. 68, No. 3, Article 036122, pp. 1–8 (2003)

[Ng 01] Ng, A. Y., Zheng, A. X., and Jordan, M. I.: Link analysis, eigenvectors and stability, *Proceedings of the 17th International Joint Conference on Artificial Intelligence (IJCAI-01)*, pp. 903–910 (2001)

[Saito 08] Saito, K., Kimura, M., and Motoda, H.: Effective visualization of information diffusion process over complex networks, *Proceedings of the 2008 European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD 2008)*, LNAI 5212, pp. 326–341 (2008)

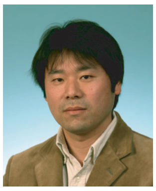
[Wasserman 94] Wasserman, S. and Faust, K.: Social network analysis, Cambridge University Press (1994)

[Watts 07] Watts, D. J. and Dodds, P. S.: Influence, networks, and public opinion formation, *Journal of Consumer Research*, Vol. 34, No. 4, pp. 441–458 (2007)

〔担当委員：栗原　聡〕

2009 年 5 月 21 日　受理

著　者　紹　介



木村　昌弘（正会員）

1987 年大阪大学理学部数学科卒業．1989 年同大学院理学研究科数学専攻修士課程修了．同年，日本電信電話株式会社入社．NTT コミュニケーション科学基礎研究所を経て，現在，龍谷大学理工学部電子情報学科准教授．複雑ネットワーク科学，データマイニングおよび機械学習の研究と教育に従事．博士（理学）．日本数学会，日本応用数理学会，日本神経回路学会，電子情報通信学会各会員．

斉藤　和巳（正会員）

1985 年慶応義塾大学理工学部数理科学科卒業．同年日本電信電話株式会社入社．1991 年より 1 年間オタワ大学客員研究員．2007 年より，静岡県立大学経営情報学部教授．機械学習，複雑ネットワーク等の研究に従事．博士（工学）．情報処理学会，電子情報通信学会，日本神経回路学会，日本応用数理学会各会員．

中野　良平（正会員）

1971 年東京大学工学部計数工学科卒業．同年，日本電信電話公社 (現 NTT) 入社．武蔵野電気通信研究所，情報通信研究所，コミュニケーション科学研究所を経て 1999 年退社．1999 年～2008 年名古屋工業大学教授，2008 年より中部大学教授．工学博士．人工知能，最適化，ニューラル情報処理の研究に従事．情報処理学会，日本神経回路学会，電子情報通信学会各会員．

元田　　浩（正会員）

1965 年東京大学工学部原子力工学科卒業．1967 年同大学院原子力工学専攻修士課程終了．同年，日立製作所に入社．同社中央研究所，原子力研究所，エネルギー研究所，基礎研究所を経て 1995 年退社．1996 年大阪大学産業科学研究所教授（知能システム科学研究部門，高次推論研究分野），2006 年定年退職し，現在米国空軍科学技術局アジア宇宙航空研究開発研究事務所 (AFOSR/AOARD) 科学顧問．大阪大学名誉教授．原子力システムの設計，運用，診断，制御に関する研究を経て，機械学習，知識獲得，知識発見，データマイニング，社会ネットワーク解析の研究に従事．工学博士．