

情報拡散モデルに基づいた社会ネットワークのリンク予測

Link Prediction for Social Networks Based on Information Diffusion Models

瀧上 晋太郎[†] 木村 昌弘^{†,††} 齊藤 和巳^{†††}

Shintaro Takigami Masahiro Kimura Kazumi Saito

1 はじめに

近年、社会ネットワークにおけるリンク予測問題が注目を集めている [1]. この問題は2つに大別されており、1つ目はネットワークで既知の部分から、残りのネットワークを予測する問題、2つ目は現在のネットワークから、ある期間後の将来のネットワークを予測する問題である。本論文では、後者の問題に取り組む。

Liben-Nowell ら [2] は、成長ネットワークにおいてノードの“proximity”に基づいたリンク予測法を提案している。ところで、現在はリンクで結ばれていないノードペアでも、一方のノードから他方のノードに情報が伝わりやすいならば、そのノード間は将来、リンクで結ばれる可能性が高いと推測される。したがって、情報拡散に基づいたリンク予測法の開発は重要と考えられる。本論文では、ネットワーク上の情報拡散モデルである Independent Cascade(IC) モデル [3,4] に基づいて、成長ネットワークにおけるリンクを予測する手法を提案する。実プログネットワークを用いた実験により、提案法の有効性を比較検証する。

2 リンク予測問題

時間的に成長する社会ネットワークのリンク予測問題を考える。無向ネットワーク $G = (V, E)$ および $G' = (V', E')$ を、それぞれ、この社会ネットワークのある時刻での観測データおよびその一定期間後の観測データとする。ここに、 V, V' はノード全体の集合、 E, E' はリンク全体の集合である。ネットワーク G のノード u において、 u の隣接ノード集合を $A(u) = \{w : \{u, w\} \in E\}$ で表す。共通の隣接ノードが存在するがリンクで結ばれていないノードペア全体の集合を、

$$S = \{\{v, w\} \notin E : v, w \in V, d(v, w) = 2\} \quad (1)$$

とする。ここで、 $d(v, w)$ はノード v とノード w のグラフ間距離である。 S に属するリンクを潜在リンクと呼ぶ。潜在リンクのうち新たに生成されたリンク全体の集合を

$$L = S \cap (E' - E)$$

とする。本論文では、 L に属するリンクを予測する問題を考察する。

3 提案法

3.1 リンク生成モデル

IC モデルに基づいた、次のようなリンク成長モデルを考える。 G の任意のノード u に対して、 $v \in A(u)$ 、 $w \notin A(u)$ を任意に選ぶ。そして、IC モデルに基づいてノード v からノード

w へ情報が伝播したとき、ノード u とノード w の間にリンク $\{u, w\}$ を生成する。

3.2 情報伝播確率の推定

本モデルに基づいて、ノード v, w 間の情報伝播確率を推定する。ノード u に対して、ノード集合 $G^+(u)$ 、 $G^-(u)$ を次のように定義する。

$$G^+(u) = \{\{v, w\} \in E : v, w \in A(u)\} \quad (2)$$

$$G^-(u) = \{\{v, w\} \in E : v \in A(u), w \notin A(u)\} \quad (3)$$

$\{v, w\} \in E$ に対して、 $p_{\{v, w\}}$ をリンク $\{v, w\}$ の情報伝播確率とする。最尤推定法に基づき、以下の目的関数の最大化問題として、 $p_{\{v, w\}}$ を推定する。

$$J = \log \prod_{u \in V} \left\{ \prod_{\{v, w\} \in G^+(u)} p_{\{v, w\}} \prod_{\{v, w\} \in G^-(u)} (1 - p_{\{v, w\}}) \right\} \quad (4)$$

ノード集合 $H^+(v, w)$ 、 $H^-(v, w)$ を次のように定義する。

$$H^+(v, w) = \{u \in V : \{v, w\} \in G^+(u)\} \quad (5)$$

$$H^-(v, w) = \{u \in V : \{v, w\} \in G^-(u)\} \quad (6)$$

式 (5) と (6) を用いると、式 (4) は以下のように書き表せる。

$$J = \sum_{\{v, w\} \in E} \{|H^+(v, w)| \log p_{\{v, w\}} + |H^-(v, w)| \log(1 - p_{\{v, w\}})\} \quad (7)$$

このとき、次の式が成り立つ。

$$|H^+(v, w)| = |A(v) \cap A(w)|, \quad (8)$$

$$|H^-(v, w)| = |A(v) \cup A(w)| - |A(v) \cap A(w)| - 2 \quad (9)$$

よって式 (7), (8), (9) より、最尤推定値 $\hat{p}_{\{v, w\}}$ は、

$$\hat{p}_{\{v, w\}} = \frac{|A(v) \cap A(w)|}{|A(v) \cup A(w)| - 2} \quad (10)$$

となる。我々は、Laplace smoothing を適用して、 $\hat{p}_{\{v, w\}}$ を

$$\hat{p}_{\{v, w\}} = \frac{|A(v) \cap A(w)| + 1}{|A(v) \cup A(w)|} \quad (11)$$

と推定する。

3.3 リンク予測法

我々は、潜在リンク $\{v, w\} \in S$ が実リンクに変化する確率 $q_{\{v, w\}}$ を

$$q_{\{v, w\}} = 1 - \prod_{u \in A(v) \cap A(w)} (1 - p_{\{u, v\}})(1 - p_{\{u, w\}}) \quad (12)$$

と推定する。

与えられた正の整数 k に対して、予測する k リンク集合を $B(k)$ ($\subset S$) とする。提案法では、 $B(k)$ を次のように計算する。

[†] 龍谷大学大学院 理工学研究科 電子情報学専攻

^{††} 龍谷大学 理工学部

^{†††} 静岡県立大学 経営情報学部

- step.1 任意のリンク $\{v, w\} \in E$ に関して、式 (11) を用い、情報伝播確率 $\hat{p}_{\{v, w\}}$ を推定する。
- step.2 任意の潜在リンク $\{v, w\} \in S$ に関して、式 (12) を用い、潜在リンクが実リンクに変換する確率 $q_{\{v, w\}}$ を計算する。
- step.3 変換確率 q の値に関して潜在リンクをランキングすることにより、予測リンク集合 $B(k)$ を求める。

4 実験評価

4.1 実験データ

日本のブログサービスプロバイダーが提供する実際のブロググローバルネットワークを用いた。まず、2006年5月にブロググローバルネットワーク $G = (V, E)$ を収集した。 $|V| = 56,894$, $|E| = 535,734$, $|S| = 156,874,190$ であった。1ヶ月後に再び、同じ部分のブロググローバルネットワーク $G' = (V', E')$ を収集した。 $|E' - E| = 41,220$, $|L| = 30,849$ であった。すなわち、新たに生成されたリンクのうち潜在リンクであったものの割合 $|L|/|E' - E|$ は75%であった。

4.2 比較法

提案法を Liben-Nowell ら [2] による従来法と比較した。我々のリンク予測問題においては、彼らの手法は、任意の $\{v, w\} \in S$ の “proximity” $score(v, w)$ を定義し、その値に従って潜在リンクをランキングすることにより、予測リンク集合 $B(k)$ を求めるものとなる。Liben-Nowell ら [2] は共著ネットワークを用いた実験において、以下の Adamic/Adar “proximity” が最も高性能であることを示した。我々は特に Adamic/Adar “proximity” を含む以下の3種類の “proximity” について調べた。

- Common Neighbors (CN):

$$score(v, w) = |A(v) \cap A(w)|$$

- Adamic/Adar (A/A):

$$score(v, w) = \sum_{z \in A(v) \cap A(w)} \frac{1}{\log |A(z)|}$$

- Preferential Attachment (PA):

$$score(v, w) = |A(v)| \cdot |A(w)|$$

また、ベースラインとして Random 法を調べた。ここに、Random 法とは、潜在リンク集合 S から一様ランダムに k 本のリンクを抽出することにより $B(k)$ を求めるものである。

4.3 評価尺度

各手法の予測性能を、ランク k に対して、求めた予測リンク集合 $B(k)$ の F 値

$$F(k) = \frac{2|L \cap B(k)|}{|L| + k} \quad (13)$$

で評価する。

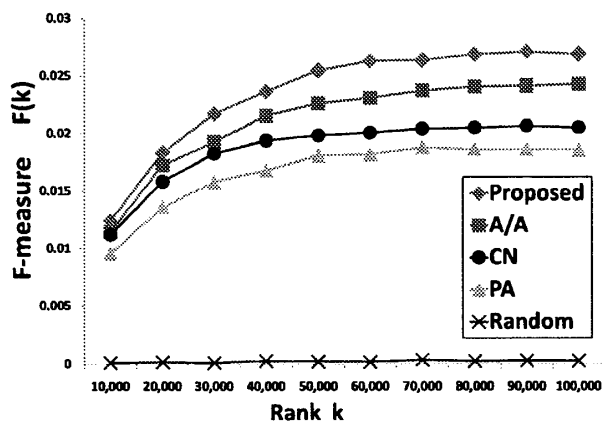


図1 予測性能の比較

表1 Random 法に対する予測性能の相対値 ($k = 30,000$)

Proposed	A/A	CN	PA
219	195	185	159

(単位 倍)

4.4 実験結果

図1は、ランク k における各手法の予測性能 $F(k)$ を表示している。提案法は従来法に比べて予測性能が高いことが観察される。

表1は、 $k = 30,000$ に関して、Random 法の予測性能 $F_{random}(k)$ に対する各手法の予測性能 $F(k)$ の相対値 $F(k)/F_{random}(k)$ を表示している。提案法は、従来法と同様、Random 法に比べて非常に予測性能が高いことが観察される。

5 まとめ

情報伝播モデルに基づく、成長する社会ネットワークのリンク予測法を提案した。大規模な実ブロググローバルネットワークを用いた実験により、提案法は、ノードの “proximity” に基づく従来法よりも、予測性能が高いことを実証した。

謝辞

本研究は科学研究費補助基盤研究 (C) (No.20500147) の補助を受けた。

参考文献

- [1] Getoor, L. and Diehl, C. P.: Link minig: a survey, *SIGKDD Explorations*, Vol. 7, Issue 2, pp. 84-89 (2005).
- [2] Liben-Nowell, D. and Kleinberg, J.: The link prediction problem for social networks, *Proc. CIKM'03*, pp. 556-559 (2003).
- [3] Kempe, D., Kleinberg, J., and Tardos, E.: Maximizing the spread of influence through a social network, *Proc. KDD-03*, pp. 137-146 (2003).
- [4] Kimura, M., Saito, K., and Nakano, R.: Extracting influential nodes for information diffusion on a social network, *Proc. AAAI-07*, pp. 1371-1376 (2007).