確率モデルを利用したリンク予測手法の提案

Link Prediction based on Network Evolving Model

志村海 [1]　大原剛三 [2*]　豊田哲也 [2]
Kai Shimura[1]　Kouzou Ohara[2]　Tetsuya Toyota[2]

[1] 青山学院大学大学院理工学研究科
[1] Graduate School of Science and Engineering, Aoyama Gakuin University
[2] 青山学院大学理工学部
[2] College of Science and Engineering, Aoyama Gakuin University

Abstract: In this work, we proposes a link prediction method that can predict new links under the situation where new nodes can be added to the network. To this end, the proposed method incorporates conventional link prediction measures into a network growth model. More specifically, the proposed method is a stochastic model based on GLP (Generalized Linear Preferential Attachment) model that considers the degree of inactiveness of nodes. A link prediction measure is combined into the calculation of the link selection probability of the model. A link is selected based on the results of a certain number of trials based on that probability, instead of selecting a link only with one trial. In addition, the proposed model considers the similarity between node attributes and the time decay function to improve the link prediction accuracy. Through experiments using co-authorship networks, we evaluate the accuracy of the proposed method

1　はじめに

近年，注目を集めている研究分野の 1 つに複雑ネットワークがある [7, 2]．その研究対象は，実世界における多数の実体間の関係を表現する巨大，かつ複雑なネットワークであり，それらに共通する諸性質を明らかにすることが主要な目的となっている．たとえば，人間関係，インターネット，食物網，神経ネットワークなど実世界における様々な関係性が複雑ネットワークとして表現され得る．これら個々のネットワークの性質は，従来からそれぞれの応用分野で研究されてきたが，1990 年代後半におけるスモールワールドモデル [7] やスケールフリーネットワークモデル [2] の提案を機に，それらの複雑ネットワークに共通する性質への関心が高まり，その研究はこの 20 年ほどの間に急速な発展を遂げてきた．

そのような複雑ネットワークに関する研究の 1 つにリンク予測がある [5]．これは，現時点のネットワークに対して，将来，新たに発生するリンクを予測することを目的としたものである．たとえば，研究者をノード，論文の共著関係をリンクとした共著者ネットワークを対象とした場合，将来のリンクを予測することは新たな共同研究関係を推定することに相当する．また，SNS（Social Networking Service）における友達推薦も，同じ問題と捉えることができる．従来のリンク予測手法の多くは，既存ノードのペアに対してネットワーク構造に基づくリンク予測指標値を計算し，その値がノードペア間に新たなリンクが発生すると予測する．これに対して，Chaturvedi らは，実ネットワークで利用可能なノード属性の類似度をリンク予測指標に加味することで予測性能が向上することを示している [3]．しかし，実世界のネットワークでは，リンクだけではなくノードも新たに発生し，それら新規ノードと既存ノードの間，もしくは新規ノード間にもリンクは生成され得るが，これらの既存手法ではそのようなリンクは予測できない．

一方，ノードの追加とリンクの発生を同時に考慮するネットワーク成長モデルも研究されている [7, 2, 4, 6]．これらの研究では，新規ノードとのリンク接続にはその時点で次数が高いノードほど選択される確率が高くなるという優先的選択（Preferential attachment） [2] や，ある程度次数が高いノードはそれ以上選択されなくなるようにノードに対する活性・非活性の状態を導入する [4] など，確率に基づいたネットワーク成長モデルが提案されている．このようなネットワーク成長モデルと前述のリンク予測をつなげる研究 [1, 6] もある．本稿では，その中でも Wang らの研究 [6] に着目す

*連絡先：青山学院大学大学院理工学研究科
〒 252-5258 神奈川県相模原市中央区淵野辺 5-10-1
E-mail: ohara@it.aoyama.ac.jp

る．この研究は，複数のネットワーク成長モデルを実際のネットワークの成長過程をどれだけうまく再現し得るかという観点から比較するものであり，その比較には，ある 2 つの時刻間で対象ネットワークで実際に発生したリンクに対する各モデルでの尤度を計算して利用している．複数の対象リンクに対する尤度の積が大きいモデルほど，現実のネットワークの成長をより忠実に再現し得ると考えられ，彼らの実験では，GLP（Generalised Linear Preference）モデルが最も高い評価を得る結果となっている．ここで，尤度の計算対象となるリンクはリンク予測問題における予測対象リンクであり，その点を考慮すると，ここで計算する尤度はリンク予測指標として利用できる可能性がある．

本稿では以上のような背景の下，Wang らの尤度計算の考えに基づいたリンク予測手法を提案し，その性能を評価する．具体的には，GLP モデルと同様に新規ノード間，新規ノードと既存ノード間，および既存ノード間すべての組合せでのリンクの発生を考慮するネットワーク成長モデルを新たに提案する．そのモデルにおけるリンク発生確率は，ネットワーク構造に基づく既存のリンク予測指標，ノード属性の類似度，既存リンクの時間減衰効果を考慮して決定する．また，提案モデルでは，Klemm [4] らの研究と同様に，既存ノードを非活性化するメカニズムを導入する．リンク予測では，通常のネットワーク成長モデルでは考慮されない実際のネットワークの特徴をより反映させた提案ネットワーク成長モデルにおけるリンク発生確率（尤度）を手がかりに発生リンクを予測する．共著ネットワークを用いた評価実験では，従来のネットワーク予測指標や Wang らの尤度をそのまま利用するよりも，提案手法の方がより高い精度でリンク予測が可能であることを示す．

2 GLP モデルと尤度の計算

本節では，本稿で提案するネットワーク成長モデルの基礎となる GLP モデルと，Wang らの尤度計算の考え方について概説する．

2.1 GLP モデル

GLP モデルでは，各単位時刻ごとに確率 p で既存のノード間に m 本のリンクを生成し，確率 $1-p$ で新規ノードを生成し，新規ノードを生成し，新規ノードと既存ノード間に m 本のリンクを生成する．各リンク生成時におけるノード i の選択は，次式により与えられる確率に従う．

$$\Pi_i = \frac{k_i - \beta}{\sum_j (k_j - \beta)} \tag{1}$$

ここで，k_i はノード i の次数であり，β は $\beta \in (-\infty, 1)$ となるパラメータである．明らかに次数が高いノードほどリンクの端点として選択される確率は高くなることから，GLP も優先的選択の考えに従ったものといえる．既存ノード間にリンクを生成する場合は，上記の確率に従い 2 つの既存ノードを選択し，新規ノードと既存ノード間にリンクを生成する場合は，基準となる新規ノードは固定し，m 個の既存ノードを上記の確率に従って選択する．

2.2 尤度の計算

Wang らは，対象ネットワーク G の時刻 t_0 と t_1（$t_0 < t_1$）におけるスナップショット $G_0 = (V_0, E_0)$ と $G_1 = (V_1, E_1)$ を考え，この時刻間で新たに発生したリンク $E_\delta = E_1 \setminus E_0$ に対して，対象ネットワーク成長モデルの下での尤度を計算している．各リンク $e \in E_\delta$ の尤度を $L(e)$ としたとき，各 e の発生が独立であると仮定することで，E_δ 全体の尤度 $L(E_\delta)$ を $\Pi_{e \in E_\delta} L(e)$ としている．BA モデルのような次数に基づく優先的選択を考えた場合，より遅い時刻で発生するリンクはそれよりも前に発生したリンクの影響を受けるため，この仮定は実際には必ずしも成り立たないことに注意されたい．

GLP モデルを対象とした場合，式 (1) から，既存ノード $x, y \in V_0$ 間にリンクが発生する場合，新規ノード $v \in V_1$ と既存ノード $x \in V_0$ 間にリンクが発生する場合，新規ノード $u, v \in E_1$ 間にリンクが発生する場合のリンク発生尤度は，各ノードがリンクの端点として選択される確率に基づき，それぞれ以下のように与えられる．

$$L(x, y) = \frac{k_x - \beta}{\sum_{j \in V_0}(k_j - \beta)} \frac{k_y - \beta}{\sum_{j \in V_0}(k_j - \beta)} \tag{2}$$

$$L(x, v) = \frac{k_x - \beta}{\sum_{j \in V_0}(k_j - \beta)} \tag{3}$$

$$L(u, v) = 1.0 \tag{4}$$

ここで，新規ノードに関してはその選択確率は常に 1.0 としている．

3 提案手法

ここでは，無向ネットワーク中のノード v は 1 つ以上の実数値属性をもつものと仮定する．その過程の下，GLP モデルを基礎とし，ノード属性，リンク予測指標，既存リンクに対する時間減衰重みを考慮したネットワーク成長モデルを提案する．そして，そのモデルにおけるリンク発生確率に基づき，新たに発生するリンクを候補から選択する．

3.1 提案ネットワーク成長モデル

提案するネットワーク成長モデルでは，GLP モデルと同様に，確率 p で既存ノード間にリンクを生成し，確率 $p-1$ で新規ノードを生成し，新規ノード間，および新規ノードと既存ノード間にリンクを生成する．ただし，非活性状態の既存ノードにはリンクを生成しないものとする．以下，各場合のリンク生成手順を示す．

1. 確率 p で既存ノード間に m 本のリンクを生成する場合，活性状態の既存ノード間に確率的にリンクを生成する．ここで，活性状態の既存ノード x，y 間にリンクが発生する確率 $p(x,y)$ は次式で与える．

$$p(x,y) = \frac{S_{xy}^{cw} - \beta}{\sum_i \sum_j (S_{ij}^{cw} - \beta)} \tag{5}$$

ここで，β は $\beta \in (-\infty, 1)$ を満たすパラメータであり，S_{xy}^{cw} は時間減衰重みとノード属性の類似度重みを考慮したリンク予測指標値である．S_{xy}^{cw} が高いノードペア間ほどリンク発生確率が高くなる．

2. 確率 $1-p$ で新規ノードを生成しリンクを生成する場合，新規ノード個数 n が 1 つの場合（$n=1$）と，2 以上の場合（$n>1$）それぞれについて，以下のようにリンクを生成する．

$n=1$ の場合

新規ノードが 1 個の場合，新規ノード z と活性状態の m 個の既存ノード間にリンクを生成する．1 本目のリンク生成では，次式で定義される確率に従って既存ノード x を選択する．

$$p(x,z) = \frac{k_x \times c_{xz} - \beta}{\sum_j (k_j \times c_{jz} - \beta)} \tag{6}$$

ここで，k_j はノード j の次数，c_{xz} はノード x，z 間の属性のコサイン類似度と定義する．この式は，次数と属性の類似度が高い既存ノードにリンクがつながる確率が高くなることを意味している．

2 本目以降のリンク生成では，次式により定義される確率に従って既存ノード x を選択する．

$$p(x,z) = \frac{S_{xz}^{cw} - \beta}{\sum_j (S_{jz}^{cw} - \beta)} \tag{7}$$

$n>1$ の場合

新規ノードが 2 個以上追加される場合，新規ノード z_1，z_2 間にリンク次式で定義する確率に従いリンクを生成する．

$$P(\langle z_1, z_2 \rangle) = c_{z_1 z_2} \tag{8}$$

新規ノードはリンクをもたないため，ノードのもつ属性のみに依存して確率が決まるものとする．具体的には，新ノード間のノード属性の類似度が高いほどそれらの間にリンクが発生しやすくなる．また，発生したそれぞれの新規ノードは，活性状態の m 個の既存ノードにリンクを m 本張る．新規ノード z，活性状態の既存ノード x 間にリンクが発生する確率 $P(x,z)$ は，式 (6) により与える．

また，新規ノードを生成した場合，生成した新規ノード数だけ活性状態のノードを非活性化する．活性状態の既存ノード x が非活性状態となるノード選択確率を次式で定義する．

$$p(x) = \frac{1}{S_{xz}^{cw} + a} \cdot \left(\sum_i \sum_j \frac{1}{S_{ij}^{cw} + a}\right)^{-1} \tag{9}$$

ここで，a は正の定数であり，z は新規ノードとする．この式は，リンク予測指標 S_{xz}^{cw} が低い新規ノードと既存ノードペアに関して，その既存ノードが非活性化される確率が高いことを意味している．

上記で用いるリンク予測指標 S_{ij}^{cw} は，対象ネットワーク G に対する隣接行列 A の (i,j) 成分 a_{ij} に対して，以下のように定義する．

$$S_{ij}^{cw} = f_{ij} \times m_{ij} \times c_{ij} \times w(\Delta t_{ij}) \tag{10}$$

ここで，f_{ij} はネットワーク構造に基づくリンク予測指標，c_{ij} はノード i，j の属性ベクトル $\vec{i}$，$\vec{j}$ を $[0,1]$ の範囲に正規化をした上で計算したコサイン類似度であり，$w(\Delta t_{ij})$ は次式で定義される時間減衰重みである．

$$w(\Delta t_{ij}) = (\Delta t)^{-\lambda} \tag{11}$$

t_{ij} はリンク (i,j) の生成時刻と現在時刻の差分であり，λ は $\lambda > 0$ となるパラメータである．

3.2 リンク予測

提案リンク予測手法では，前述の提案ネットワーク成長モデルを考え，そのリンク発生確率に基づき一定回数のリンク選択試行の結果，最も選択回数の多くなるリンクを予測結果とする．これは，Wang らの研究では予測対象となる正解リンクの再現性を尤度を用いて評価するのに対し，実際のリンク予測では正解データがないため，すべての可能な候補のリンクの尤度を考えなければならないためである．リンク発生確率が最も高いリンク候補が実際のネットワークで生成されるとは限らないため，ここでは確率に基づく多様性を担保するために一定回数のリンク選択試行を導入する．

表 1: DBLP データセット (1962-1967)

期間	著者	共著関係	平均次数	平均最短パス長	クラスタ係数
1962-1967	2,399	2,429	2.681	1.108	0.710

トレーニング期間			テスト期間		
期間	ノード数	リンク数	期間	新規ノード数	新規リンク数
1962-1964	1,252	1,339	1965	533	458
			1965-1966	734	718
			1965-1967	1,143	1,090

4 評価実験

4.1 実験設定

本研究では，Michael らが Web 上で公開している DBLP の共著者ネットワークのデータを用いる [1]．このデータセットには 1936 年から 2016 年の期間に出版された計 3,673,934 本の論文データが含まれており，著者をノード，共著関係をリンクとしたときのノード数は 1,851,761，リンク数は 5,926,951 である．

本実験では，従来のリンク予測に関する研究と同様に，対象データをトレーニング期間とテスト期間に分割することを考える．具体的には，トレーニング期間の最初にリンクが発生した時刻を t_0，トレーニング期間の最後にリンクが発生した時刻を t'_0 としたネットワークを $G[t_0, t'_0]$ と表す．同様に，テスト期間の最初にリンクが発生した時刻を t_1，テスト期間の最後にリンクが発生した時刻を t'_1 としたネットワークを $G[t_1, t'_1]$ と表す．ここで，ネットワーク $G[t_0, t'_1]$ はトレーニング期間，テスト期間におけるすべてのノードとリンクが含まれているネットワークとする．本実験では，対象ネットワークの大きさを制限するために，テスト期間のネットワーク $G[t_1, t'_1]$ で発生するリンクの端点となる著者のみから構成されるネットワークを考える．本実験では，トレーニング期間 [1962,1964] に対して，3 種類のテスト期間 [1965]，[1965-1966]，[1965-1967] を考え，テスト期間の長さに対する予測性能を評価する．トレーニング期間とテスト期間の組み合わせを表 1 に記す．

提案手法で用いるネットワーク構造に基づくリンク予測指標としては，共通隣接ノード，Adamic/Adar，Katz$_\beta$ の 3 つを用いた．各指標値の定義については文献 [5] を参照されたい．また，今回の実験では，テスト期間中に新規に発生したノードはランダムに決定した順序で発生したと仮定し，確率 p での新規ノードの発生は考慮していない．一度に生成するリンク数 m は 3，提案手法におけるリンク選択試行の回数は 100，リンク選択確率における β は -1.0，katz$_\beta$ のパラメータ値は 0.05 とした．評価指標は，正解リンクに対する予測結果の F 値の 10 試行平均を用いた．

[1] http://dblp.uni-trier.de/

4.2 実験結果

まず，提案手法と単純に GLP モデルを用いた場合のリンク予測精度の比較を図 1 に示す．GLP モデルを用いたリンク予測では，提案手法同様に確率に基づいたリンク選択選択試行 100 回の下での選択回数の多いリンクを予測結果とした．この結果から，提案手法は，3 つのネットワーク構造に基づくリンク予測指標との組み合わせいずれにおいても，GLP モデルを単純に適用した場合と比べて高い予測精度を達成していることがわかる．特に，共通隣接ノードを用いた場合の精度が高くなっている．

次に，Nowell らの実験 [5] と同様に，既存ノード間にだけ新規リンクが発生すると仮定した場合の，提案手法と各リンク予測指標をそのまま用いた場合のリンク予測結果を比較した．結果を図 2〜4 に示す．図中の Score が各リンク予測指標の結果を示しており，Probability が提案手法の結果を示している．この結果から，従来のリンク予測指標をそのまま用いるよりも，提案手法のほうが高い精度でリンクを予測できていることが分かる．この場合も，やはり共通隣接ノードがもっとも高い精度を達成する結果となった．これは，提案手法で考慮したノード属性の類似性との親和性が高いためであると考えられる．

5 おわりに

本稿では，ネットワーク成長モデルと従来のネットワーク構造に基づくリンク予測手法を組み合わせたリンク予測手法を提案し，その予測性能を実験的に評価した．現状では，限定的な実験設定での結果しか得られていないが，両者を組み合わせることで，より高い

予測精度が得られることがわかった．今後，より多様なデータでの検証を進める予定である．

参考文献

[1] Ahn, M. W., and Jung, W. S.: Accuracy Test for Link Prediction in terms of Similarity Index: The Case of WS and BA Models, Preprint submitted to Physica A, pp. 1-12 (2015).

[2] Barabasi, A., and Albert, R.: Emergence of Scaling in Random Networks, Science, Vol. 286, pp. 509-512 (1999).

[3] Chaturvedi, A., and Acharjee, T.: An Efficient Modified Common Neighbor Approach for Link Prediction in Social Networks, Proceedings of the Journal of Computer Engineering, Vol. 12, pp. 25-34 (2013).

[4] Klemm, K., and Eguiluz, V. M.: Highly Clustered Scale-free Networks, Physical Review E, Vol. 65, pp. 1-6 (2002).

[5] Nowell, D., and Kleinberg, J.: The Link Prediction Problem for Social Networks, Proceedings of the Conference on Information and Knowledge Management 2003, pp. 556-559 (2003).

[6] Wang, W. Q., Zhang, Qian. M., and Zhou, T.: Evaluating Network Models: A likelihood analysis, Europhysics Letters, Vol. 98, No. 2, pp. 1-6 (2012).

[7] Watts, D. J., and Storogatz, S. H.: Collective Dynamics of Small-World Networks, Nature, Vol. 393, pp. 440-442 (1998).

[8] 志村海，白川真一，大原剛三：リンク予測における時間減衰の効果について，人工知能学会知識ベースシステム研究会資料，Vol. 102, pp. 19-25 (2014).

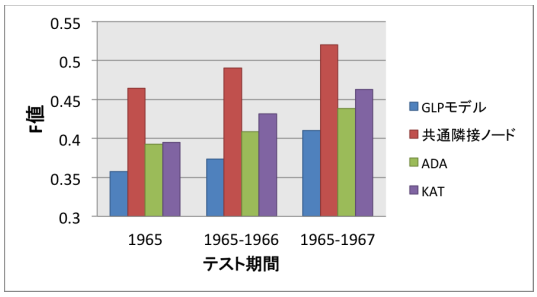

図 1: GLP モデル優先的選択と提案手法の比較

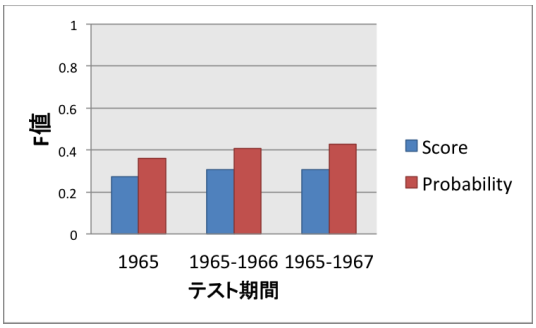

図 2: リンク予測指標ごとの比較（共通隣接ノード）

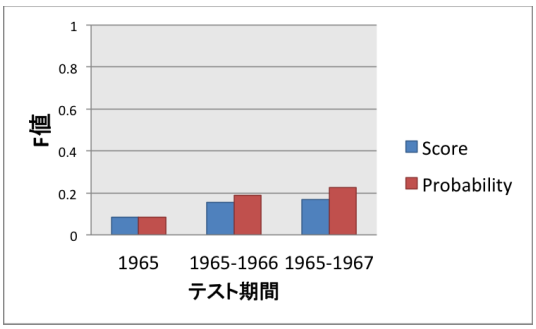

図 3: リンク予測指標ごとの比較（Adamic/Adar）

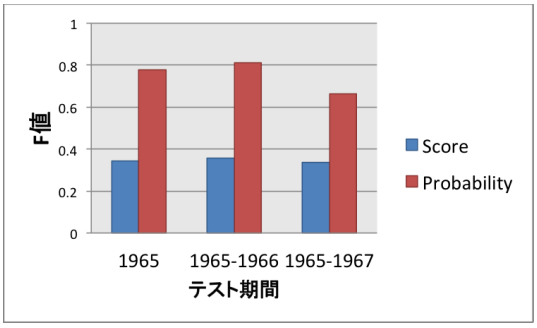

図 4: リンク予測指標ごとの比較（Katz_{β}）