

リンク予測を基にした時系列ネットワーク中でのオブジェクトランキング

Object Ranking in Evolutional Networks via Link Prediction

宮西 大樹
Taiki Miyanishi

神戸大学大学院システム情報学研究科
Graduate School of System Informatics, Kobe University
miyanishi@ai.cs.kobe-u.ac.jp

関 和広
Kazuhiro Seki

(同 上)
seki@cs.kobe-u.ac.jp

上原 邦昭
Kuniaki Uehara

(同 上)
uehara@kobe-u.ac.jp

keywords: link analysis, rankboost, link prediction, object ranking

Summary

This paper proposes a framework to predict future significance or importance of nodes of a network through link prediction. The network can be of any kind, such as a co-authorship network where nodes are authors and co-authors are linked by edges. In this example, predicting significant nodes means to discover influential authors in the future. There are existing approaches to predicting such significant nodes in a future network and they typically rely on existing relationships between nodes. However, since such relationships are dynamic and would naturally change over time (e.g., new co-authorship continues to emerge), approaches based only on the current status of the network would have limited potentiality to predict the future. In contrast, our proposed approach first predicts future links between nodes by multiple supervised classifiers and applies the RankBoost algorithm for combining the predictions such that the links would lead to more precise predictions of a centrality (significance) measure of our choice. To demonstrate the effectiveness of our proposed approach, a series of experiments are carried out on the arXiv (HEP-Th) citation data set.

1. はじめに

近年, Web を通じて World Wide Web のリンク構造やたんぱく質の相互作用, ソーシャルネットワークサービス (SNS) 上での友達関係, 論文の引用関係, 道路の交通網など, ネットワーク構造を持ったデータが容易に取得できるようになっている. ネットワークデータでは, 化学物質, 人物, 論文などをネットワーク中のノード, 互いの関係をリンクと考え, ノードを個別に見るのではなく, ノード間の関係全体を見ることで, ネットワーク特有の性質を見いだすことができる. 例えば, ノードの次数がべき分布に従うスケールフリー性 [Barabási 99] や, ノード間の経路が短くグラフが凝集された状態になるスモールワールド性 [Watts 98] は様々な種類のネットワークに共通して見られる性質である [Broder 00, Jeong 00, Newman 01b]. また, ネットワークの普遍的な原理を抽出するだけでなく, ネットワークの性質を生かして, 現実世界の問題を解決しようとする試みもある. 観測されたネットワークから未知の部分を予測するリンク予測や, ネットワークの構造を基にしてノードの重要度や影響力を量る

研究は, リンクマイニング [Getoor 05] と呼ばれている. リンク予測は, タンパク質の相互作用の予測 [Ben-Hur 05] や論文の共著関係の予測 [Liben-Nowell 07], SNS での友達の推薦 [Backstrom 11] に応用されている. 一方, ノードの重要度予測の例としては, 情報検索での Web 文書の順位付け [Brin 98, Kleinberg 99] や, 論文の著者が持つ重要度の推定 [Liu 05] や, Twitter^{*1} 上での影響力を持ったユーザーの特定 [Weng 10] などがある.

これまでのネットワークを対象とした研究の多くは, ノード間に新たな関係が出現したり既存の関係が消滅せず, 時間と共にネットワークの構造が変化しない場合を想定してきた. しかし, ネットワークデータの中には時間と共に構造が変化する時系列ネットワークの性質を持つデータも存在する. 時間ごとにネットワーク構造が変化すれば, 新たにリンクが出現すると同時にネットワーク中のノードの重要度も時間と共に変化すると考えられる. よって, 現在において重要とみなされていた化学物質, 人物, 論文などのノードが将来的には重要でなくなり, 逆に注目されていなかったノードが未来では重要に

*1 <https://twitter.com/>

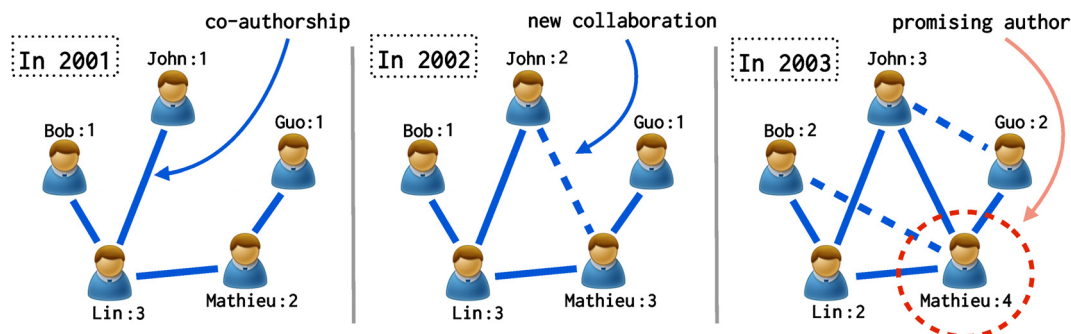


図1 リンク予測に基づく時系列ネットワーク中のノードの順位付け

なる可能性がある。未来のノードの重要度を予測できれば、有望な化学物質・薬物の早期発見や、現在または未来に注目されそうな論文を早い段階で見つけることができたり、Web ページやマイクロブログに対して即時に評価が行えるため、意思決定を行う際に有益な情報となる。

図1にノードを論文の著者、リンクを共著関係とした共著ネットワークを例にして、時系列ネットワークの特徴について述べる。この図は2001年から2003年に存在する5人の研究者 John, Bob, Guo, Lin, Mathieu の共著関係の遷移を描いており、共著の数が研究者の重要度となっている。時間が経つにつれて新たな共著関係が発生しているため、研究者の重要度が年代ごとに変化していることがわかる。特に Mathieu は2001年の時点であまり重要ではないものの、2003年の時点で共著数が2倍に増え、重要度が増していることが分かる。このように、現在では重要でないものの将来的に重要になるようなオブジェクトを特定することが本研究の目的である。

本稿では、時系列ネットワークの中から将来的に重要になるノードの順位付け手法を提案する。提案手法は大きく2つの段階に分かれる。

- (i) ノード間の周辺構造を基にリンク予測を行い、将来のネットワークを予測
- (ii) 予測したネットワークから得たノードの順位を既知のノードの順位を考慮して重み付けすることで、ノードの順位を間接的に予測

前者は、未来のネットワークのリンク構造を予測できれば、間接的にノードの順位も予測できるという仮説に基づいている。後者は、リンク予測とノードの順位を組み合わせることで、より精度良くノードの順位を予測できるという仮説に基づいている。そこで、実際の時系列ネットワークデータを用いて上記2つの仮説の検証を行う。使用する時系列ネットワークデータには、高エネルギー物理理論の論文データベース arXiv (hep-th) ^{*2}にある共著関係をもとに構築した共著ネットワークを用いる。ちなみに共著ネットワークでの共著関係は、一度できると消滅しないので本稿では関係の消滅は扱わない。

本論文では、まず2章で時系列ネットワーク中の著者の重要度の推移と、共著関係の発生が将来の著者の重要度に与える影響について検証する。3章では、将来的に重要になるノードを予測する方法として、リンク予測とノードの順位予測について述べる。そして4章では共著ネットワーク上の重要な著者の予測を行い、実験結果を示す。5章で本研究と関連する研究について紹介し、我々の研究との違いについて述べる。最後に6章でまとめと今後の課題について議論する。

2. 仮説の検証

本研究の目的は将来的に影響を持つまたは重要になるオブジェクト(ネットワーク上ではノードに対応する)を同定することである。この目的を達成するために、本章ではまず以下の3つの仮説を検証する。

- 【仮説1】 時系列ネットワーク中のノードの重要度は時間と共に変化する。
- 【仮説2】 正しいリンク予測を行うことで、精度良くノードの重要度の予測ができる。
- 【仮説3】 リンク予測とノードの重要度を組み合わせることで、より精度良くノードの重要度を予測できる。

ここで、重要なオブジェクトとは、ネットワーク上で要となるノードを指す。今回は共著ネットワークを用いるので、ノードの重要度は研究者が他の研究者と比べて協力的であることを表したり、異分野をつなぐ要となる研究者を示したり、ネットワーク上での波及的な影響力を表す。このように、重要度はどのようなネットワークを対象とするのか、どのような情報を取り出したいのかによって定義が異なる。本研究では、ネットワーク中のノードに対して具体的な意味を持つ重要度として古くからネットワーク分析で用いられてきたネットワークの中心性 [Freeman 79] という指標を用いる。

^{*2} <http://kdl.cs.umass.edu/data/hep-th/hep-th-info.html>

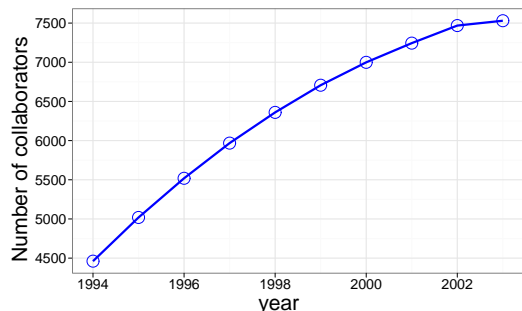


図2 arXiv (hep-th) から抽出した 1994 ~ 2003 年に存在する著者の共著関係数の推移

2.1 ノードの重要度

ネットワークの中心性とは、ネットワーク中のノードがどの程度の影響力を持っているか、中心的な役割を果たしているかをネットワークのリンク構造を基に定義した指標である。中心性には複数の定義があり、対象とするデータ・目的によって意義のある指標は異なる。以下に代表的な中心性を列挙する。ここでの E と V はそれぞれネットワークを構成するリンクの集合とノードの集合とする。

● 媒介中心性 (Betweenness)

媒介中心性はネットワーク中のクラスタ同士を結ぶノードを重要とみなす指標であり、自身を経路として通る任意のノード間の最短パスの数で定義する。これは全てのノードの組み合わせ j, k ($j, k \in V, j \neq i, k \neq i$) の最短パス $path_{jk}$ にノード i が含まれる数に等しい。

● 近接中心性 (Closeness)

近接中心性はあるノードと他のノードがどれだけ近い距離にあるかを示す指標であり、任意のノードに到達するための最短パスの平均で定義される。ノード i からノード j ($j \in V, i \neq j$) への距離 d_{ij} の平均の逆数 $\frac{1}{\sum_{j \in V, j \neq i} \frac{1}{d_{ij}}}$ が中心性の値となる。

● 次数中心性 (Degree)

次数中心性は、ネットワーク内のノードが他のノードとどの程度つながっているかを表す指標であり、自身に接続するノード数で定義される。つまり、ノード i がノード j に接続するリンク $e_{ij} \in E$ の数に等しい。

● PageRank

PageRank [Brin 98] は、ネットワーク中のノード間を確率的に遷移するランダムサーファァーを考えた場合、ランダムサーファァーがあるノードに定常的に訪問する確率を表している。この指標は、高いPageRankの値を持つノードに多く隣接するノードは高いPageRankの値を持つという再帰的な性質を持つ。各ノードのPageRankをベクトル表現したPageRankベクトルは $\vec{x} = \alpha \mathbf{P}^T + (1 - \alpha) \frac{\vec{1}}{n}$ として表現することができる。

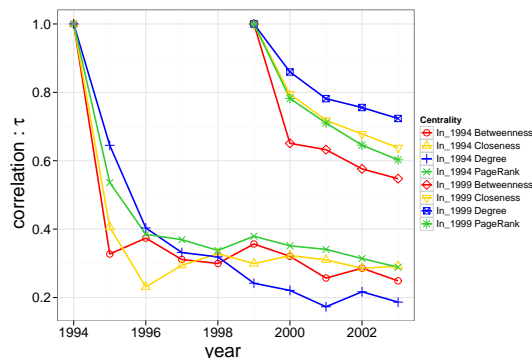


図3 各年代毎のネットワーク中心性による著者の重要度の変化

ここで、 \mathbf{P} は遷移行列、 n はノード数、 α は、ランダムサーファァーのテレポテーションを制御するパラメータである。

論文の著者をノード、論文の共著関係をリンクとした共著ネットワークを例にして、上述した中心性と共著ネットワーク上で要となる研究者との関係を説明する。媒介中心性はノード間の最短パス上にあるほど値が高くなるので、この指標を用いて複数の分野にまたがる研究者を探すことができる。近接中心性は他の研究者とつながりが近いほど値が大きくなるので、複数の研究者を介して効率的に多くの研究者と知り合える人物を示す指標となる。次数中心性は共同研究を行い協力関係にあった研究者の数を表すので、協力的な人物を表す指標となる。PageRankは自分の中心性が大きければ、自身につながる研究者の中心性も大きくなるので、波及的な影響力を持つ人物を表す指標といえる。

ネットワークの中心性は定義から分かるように、リンク構造の変化するネットワークを考慮していない。しかし、本研究で用いる共著ネットワークは時間が経つごとに新たな共著関係が発生するため、ネットワークの構造が絶えず変化する。上述した中心性の値も時間の経過と共に変化するため、現時点の重要度から未来のノードの重要度を知ることはできないと考えられる。

2.2 時間ごとの重要度の変化

本節では、2.1節で紹介したネットワークの中心性によって著者を順位付けし、各年代ごとの著者の順位の遷移を見ていくことで、本章の冒頭で述べた【仮説1】を検証する。時系列ネットワークには、arXiv (hep-th) を基に作成した共著ネットワークを用いる。このデータは2003年までに8,392人の著者が存在し、総共著数は87,794個、同著者間の共著回数を除くと20,387個の共著関係が存在する。図2は縦軸が共著数、横軸が年代を表しており、年代ごとの共著数の変化を示している。この図から、1994~2003年までに新たな共著関係が時間の経過と共に増加し続け、年代を経るごとに共著ネットワークの構造が変化していくことが分かる。

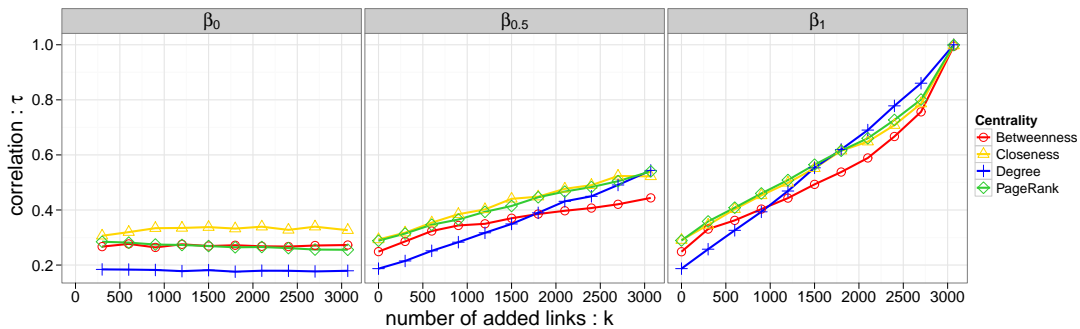


図4 正しいリンクを加えた数と予測したノードの重要度との関係

著者の重要度の変化を確認するため、データセット中の1994年までに存在する全著者2,950人を対象として、2003年までの変化を見ていく。重要度の変化は、1994年の著者のネットワーク中心性の値と1994～2003年の値の上位100件に対してケンドールの順位相関係数 τ [Kendall 38]を用いて測る。相関係数 τ の値が1に近いほど、互いの重要度によるノードの並びは似ており、0に近いほど似ていないことを示している。なお、以降よりPageRank [Brin 98]のテレポテーションを制御するパラメータ α の値はBrinらにならい $\alpha = 0.85$ に統一する。

図3に、ある年代と他の年代との中心性に基づくノードの並びの相関を示す。縦軸が重要度の並びの上位100件の相関係数、横軸が各ネットワークの年代を示している。図中の左の折れ線は、1994年のノード重要度の並びとその他の年代の並びを比較した結果であり、右は1999年とその他を比較した結果である。図より、年代を経るごとにすべての中心性について相関係数の値が減少していることが分かる。また、比較する年代の始点(1994年と1999年)によって変化の度合いが異なることも分かる。これは1994年以降の関係の増加数に比べて1999年の増加数が少ないためだと考えられる(図2参照)。ただし、両者とも年代を経るごとにすべての中心性について相関係数の値が減少している。このことから、時間の経過により共著ネットワークの構造が変化することで、著者の重要度は変化すると結論づけることができる。すなわち、【仮説1】の正しさを検証できた。

2.3 リンクの発生過程とノードの重要度の関係

前節において、ネットワークの構造が変化すると共にノードの重要度も変化することが分かった。次に、時間が経つごとに発生するリンクの数とノードの重要度との関係を明らかにする。今回、ネットワークの中心性をノードの重要度として用いるため、その値はネットワークのリンク構造によって決まる。すなわち、将来のリンクの発生を予測できれば、未来のノードの重要度を予測できると考えられる。このことから、【仮説2】「正しいリンク予測を行うことで、精度良くノードの重要度の予測ができる」かどうかについて以下の手順で検証する。

- (1) 1995～2003年の間にリンクが発生するノードペアを正リンク、そうでないペアを負リンクとし、両者の割合を β ($0 \leq \beta \leq 1$)とする。
- (2) 正リンクと負リンクを $\beta : 1 - \beta$ の割合になるように k 本のリンクを無作為に選択する。
- (3) 選択した k 本のリンクを1994年のネットワークに加えて2003年のネットワークを予測する。
- (4) 予測したネットワーク中の各ノードに対して所望のネットワーク中心性をノードの重要度として計算する。
- (5) 実際の2003年のネットワークのノードと予測したノードの重要度上位100件を順位相関係数 τ で比較する。
- (6) 上記の手順を30回繰り返し、 τ の平均を計算する。

図4は上記の手順で求めた順位相関係数 τ と加えたリンクの本数 k の関係を各 β ごと($\beta = 0, 0.5, 1$)に示したものである。縦軸が重要度の並びの相関係数、横軸が加えたリンクの本数となる。図から、実際に発生しないリンクを加えていった場合($\beta = 0$)、相関係数はほぼ変化しない。つまり、誤ったリンクを予測しても将来のノードの重要度は予測できないことが分かった。一方、実際に発生するリンクを加えた場合($\beta = 1$)、加えるリンクの数を多くするほど線形的に相関係数が増加しており、正しいリンクを予測することでノードの重要度を精度良く予測できることが分かった。また、正リンクの数と負リンクの数の割合を等しくした場合($\beta = 0.5$)、リンクを加えたときの相関係数の上昇率は、 $\beta = 1$ のときの半分程度になっていることが分かる。このことから、リンク予測の結果がすべて正しくなければ($\beta < 1$)、将来の順位を完全に予測することはできず($\tau < 1$)、その一方でリンク予測の精度を上げる($\beta > 0$)ことで、ノードの重要度予測の精度を上げることができると分かった。【仮説2】の正当性を確認できた。

2.4 ノードの重要度を考慮したリンクの発生過程

前節で、将来的に発生するリンクを正しく予測することで、将来のノードの重要度を正しく予測できることが分かった。前節のリンク予測の方法は将来的にリンクが

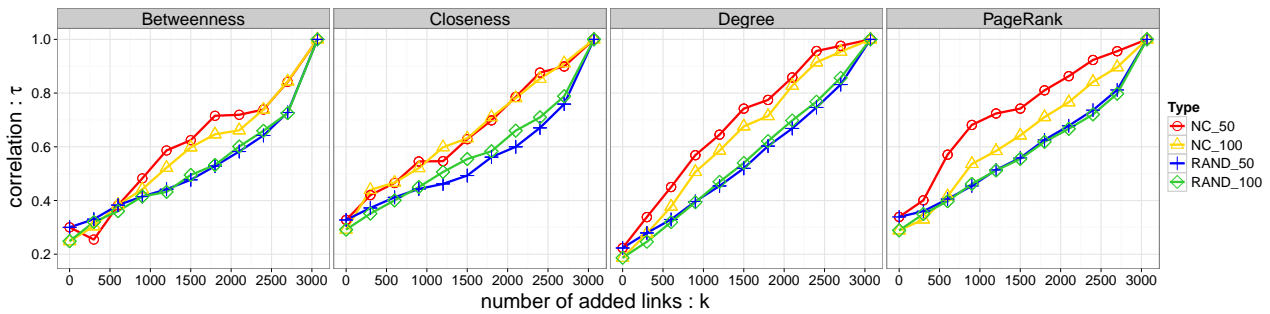


図5 ノードの重要度を考慮してリンク予測を行いノードの重要度を予測

あると分かっているノードのペアをリンク有りとみなし、そのリンクを既知のネットワークに無作為に付与してノードの重要度を間接的に予測していた。ここで、無作為にリンクを付与するのではなく、ノードの重要度に関連するリンクを優先的に付与すれば、正解のリンクが少ない場合でもノードの重要度を正しく予測できるのではないかと考えられる。そこで、【仮説3】（リンク予測をノードの重要度と組み合わせた方が精度良くノードの重要度を予測できる）を検証する。

このために、ノードの重要度を考慮したリンクの発生過程の手順を以下に示す。

- (1) 1995～2003年 で実際に発生するリンクに対して、1994年のノードの中心性の値をもとにリンク発生の優先度を決める。1つのリンクには2つのノードが接続するので、高い方の中心性の値をリンクの優先度として、リンクを降順に順位付けする。
- (2) 順位付けされたリンクの上位 k 本を 1994 年に付与し 2003 年のネットワークを予測する。
- (3) リンクを付与したネットワーク上で各ノードの重要度を計算する。
- (4) 実際の 2003 年のネットワークのノードと予測したノードの重要度上位 100 件を順位相関係数で比較し、この相関係数を τ とする。
- (5) 上記の手順を 30 回繰り返し、 τ の平均を計算する。

上記の手順による相関係数と加えたリンクの本数の関係を図5に示す。縦軸が重要度の並びの相関係数、横軸が加えたリンクの本数となる。RAND_50とRAND_100は2.3節で紹介した無作為に真のリンクを加えてノードの重要度を予測し、それぞれ上位50件と100件について真の重要度との相関を見た結果である。一方、NC_50とNC_100は上で述べたノードの重要度を考慮して真のリンクを加え、それぞれ上位50件と100件のノードの順位を予測した結果である。図5より、すべてのリンク予測手法でリンクを加えるごとに順位相関係数 τ が増加していることが分かった。また、正しいリンクを1000本程度の少ない本数で予測した場合でも、ノードの重要度を考慮したリンク予測の方が無作為なリンク予測より精度良くノードの重要度を予測できていることが分かった。

これは上位50件、100件のノードどちらにも言える。このことから、ノードの重要度を考慮したリンク予測を行うことで、精度良くノードの重要度を予測できることが分かった。【仮説3】が正しいことを確認できた。

本章では、予備実験としてあらかじめリンクの有無を知った上でリンク予測を行った。しかし、実際の場面ではリンクの有無をあらかじめ知る事はできないため、何らかの方法でリンクの有無を予測する必要がある。次章では、以上の知見を基に、過去のデータを用いてリンク予測を行う方法とその結果を利用してノードの重要度を予測する方法について述べる。

3. 提案手法

3.1 リンク予測

未来のネットワークを予測することは、既知のネットワーク中にある全ノードペアのリンクの有無を判別することと同等である。これをリンク予測と呼び、ノードの持つ情報から予測する方法とネットワークの構造から予測する方法の2種類に分けられる。今回はリンク予測の汎用性を考えて、ノードが持つデータの性質が明示されておらず、ネットワークのリンク構造だけが観測できる場合を想定する。そして、各ノードペアのリンクの有無を独立に予測していく。リンク構造から予測する方法は、ノードペアに対してリンク構造に基づく特徴を定義し、機械学習の手法を用いてリンクが発生するノードペアの特徴を学習する。ノードペアの特徴としては、2ノード間の周辺のリンク構造を定量化したリンク指標の値を用いる。以下に代表的なリンク指標を示す。

● 共通隣接ノード指標 (COM)

共通隣接ノードは、あるノードペアに共通して隣接するノードの数を表す指標である [Newman 01a]。Newmanは研究者ネットワークでこの指標を計算し、2人の研究者に対して関係のある研究者の数と将来的にその2人の間に新たな関係ができる確率には相関があることを示した。ノード n と隣接するノードの集合を $\Gamma(n)$ とすると、2ノード間 $\langle v^{(i)}, v^{(j)} \rangle$ のスコアは以下のように定義できる。

$$\text{COM}(v^{(i)}, v^{(j)}) = |\Gamma(v^{(i)}) \cap \Gamma(v^{(j)})| \quad (1)$$

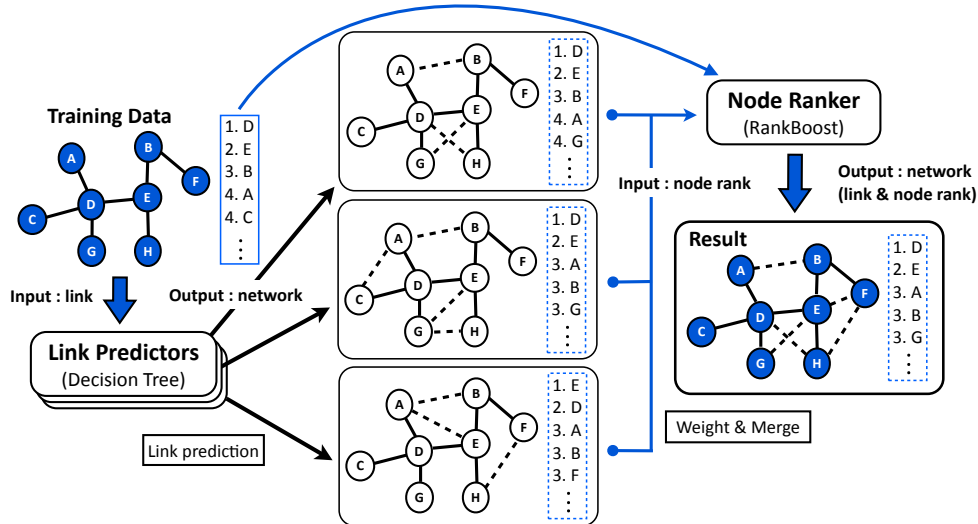


図 6 RankBoost を用いてノードの順位を求める手法の概念図

● Jaccard 係数 (JAC)

Jaccard 係数は 2 つのノードに隣接するノード集合の内、互いに隣接するノードの数の割合を示した指標である。この指標は情報検索でベクトル間の類似度を表すためにも用いられている [Manning 99]。Jaccard 係数は以下のように定義できる。

$$\text{JAC}(v^{(i)}, v^{(j)}) = \frac{|\Gamma(v^{(i)}) \cap \Gamma(v^{(j)})|}{|\Gamma(v^{(i)}) \cup \Gamma(v^{(j)})|} \quad (2)$$

● Adamic/Adar (ADA)

この指標は共通隣接指標を改良した尺度であり、共通に隣接するノードの中でも他のノードとあまり隣接しない稀なノードを重く評価する [Adamic 03]。Adamic/Adar は以下のように定義できる。

$$\text{ADA}(v^{(i)}, v^{(j)}) = \sum_{k \in |\Gamma(v^{(i)}) \cap \Gamma(v^{(j)})|} \frac{1}{\log |\Gamma(v^{(k)})|} \quad (3)$$

● $Katz_\gamma$ 指標 (KAT)

$Katz_\beta$ は、2 つのノードを結ぶパス数の重みつき和で表され、共通隣接ノード指標の一般形であるといえる [Katz 53]。ノード $v^{(i)}$ と $v^{(j)}$ を結ぶ長さ l のパス数を $\text{paths}_{v^{(i)}, v^{(j)}}^{(l)}$ と表すと、以下のように定義できる。

$$\text{KAT}(v^{(i)}, v^{(j)}) = \sum_{l=1}^{\infty} \gamma^l |\text{paths}_{v^{(i)}, v^{(j)}}^{(l)}| \quad (4)$$

● 優先的選択指標 (PRE)

優先的選択指標はスケールフリーネットワークの生成モデルに基づいた指標である [Barabási 99]。ノード v に接続するリンクができる確率は v に隣接するノードの数 $|\Gamma(v)|$ に比例すると考える。言い換えれば、ノードが他の多くのノードと隣接していればいほど、そのノードは将来的に新しいノードを獲得し易いとした指標であり、下記のように定義できる。

$$\text{PRE}(v^{(i)}, v^{(j)}) = |\Gamma(v^{(i)})| \cdot |\Gamma(v^{(j)})| \quad (5)$$

以後、 $Katz_\gamma$ にあるノード間のパスの重みを表す減衰パラメータ γ の値を 0.05 とする。リンク予測では、ネットワークの性質ごとに有効なリンク指標が異なる。そこで、本研究で用いるリンク予測手法では、ノードペアの全種類のリンク指標を特徴として使い、リンクの有無をクラスラベルに見立て、C4.5 [Quinlan 93] を用いてリンクができるノードペアの特徴を学習する。C4.5 は予測した事例がクラスに属する確率を与えることができるので、このリンクの有無の確率を大きい順に並べて、上位 l 件のノードペアに対してリンクが有ると判定する。なお、リンク予測問題は正例（リンク有り）と負例（リンク無し）の数に大きな差がある不均衡データであるため、バギングを行うことでこの不均衡データの問題に対処する [Chawla 04]。具体的には、負例だけを復元抽出して、正例と負例の数が同数の訓練データセットを複数作成し、各々のデータセットを用いて分類器を作成する。そして、作成した分類器の出力結果の平均を最終的なリンク予測の結果とする。

3.2 ノードの順位学習と予測

リンク予測を正しく行うことでノードの重要度を正しく予測できることが 2.3 節で判明している。また、実際に発生するリンクの内、重要度の高いノードに接続するリンクはよりノードの重要度予測に貢献することが明らかになっている (2.4 節参照)。このことから、重要度の高いノードの情報をリンク予測と組み合わせれば、より精度良くノードの重要度予測ができるのではないかと考えられる。しかし、ノードの重要度とリンク予測の結果のどちらを重視するか、またどのように組み合わせるかは自明ではない。また、リンク予測からノードの重要度予測を間接的に行うのではなく、ノードの重要度予測を直接行う方が妥当である。そこで、ノードの重要度予測に有効なリンク予測の結果を使い、ノードの重要度（順位）を直接予測するためノードの順位学習を用いる。

順位予測を行う方法として、以下の3つの方法を試みる．ここで、C4.5を10個バギングした結果を1つのリンク予測器とする．

リンク予測器単体 (ONE)

リンク予測器を1つだけ用いてリンク予測を行い、予測したネットワークから各ノードの中心性を求め、ノードの順位を間接的に予測する．

リンク予測器複数 (MUL)

リンク予測器を複数用いてリンク予測を行い、それぞれのリンク予測器を用いてノードの順位を複数予測し、その平均を最終的なノードの順位とする．なお、リンク予測の結果については全リンク予測器の出力の平均を用いる．

RankBoost (RB)

ノードの順位を教師情報として用い、リンク予測器によって予測されるノードの順位を事例としてRankBoost [Freund 03] を適用する．各事例の重みを利用してノードの順位を加重平均する．

いずれの手法も、リンク予測から得られたネットワークからノードの順位を間接的に予測している．各々の違いは、リンク予測の仕方とリンク予測の重み付けの仕方である．ONEとMULは、リンク予測を単体で行うか複数で行うかで異なり、リンク予測の重み付けは両者とも行わない．一方、RBのリンク予測の方法はMULと同じであるものの、リンク予測の重み付けにノードの順位情報を使う点でMULと異なる．図6に、RankBoostを用いたノードの順位予測手法の模式図を示す．図にあるように、RankBoostはノードの順位予測を上手く行うことのできたリンク予測に大きい重みを与える仕組みになっている．次節でRankboostについて説明する．

3.3 RankBoost

RankBoostは順位学習アルゴリズムの一つで、情報検索 [Liu 07]、タグ推薦 [Wu 09]、表情認識 [Yang 09] の分野で、文書、タグ、画像などのオブジェクトの再順位付けに使用されている．RankBoostの学習では、順序が定義された事例に対して順序を与える関数を弱学習器として用い、予測した順序と訓練データ中の順序の不一致を損失として定義する．この損失を逐次的に最小化することで、事例のペアの順序関係を学習する．学習の段階では、正しい順序を持つペアの重みは小さくしつつ、誤った順序を持つペアの重みを大きくすることで、事例のペアの重みを更新する．RankBoostの損失関数は式(6)のように定義される．

$$rloss_D(H) = \sum_{x_0, x_1} D(x_0, x_1) \delta(H(x_1) \leq H(x_0)) \quad (6)$$

$$D(x_0, x_1) = \max(0, \Phi(x_0, x_1)) \quad (7)$$

ここで、 x_0 は x_1 よりも順位が低い事例を示しており、 H は事例に対して順位を与える弱学習器の線形和(強学

Algorithm1 RankBoost によるノードの順位付けの流れ

入力: ★ 教師情報となるネットワーク中心性 c に基づく

上位 p 件のノード $\{v_i\}$ ($i = 1, \dots, p$) の順位

★ $\mathcal{V} \times \mathcal{V}$ 上のノードペアの重みの分布 D

★ 時刻 t のネットワーク g^t

ここで、 \mathcal{V} はノードの集合を表す

○ 分布の初期化 $D_1 = D$

○ ネットワーク g^t から K 個のリンク予測器 $\{L\}_K$ を作成

○ $\{L\}_K$ から時刻 $t+1$ の K 個のネットワーク g^{t+1} を予測

for $k = 1, \dots, K$ **do**

• ノードペア (v_i, v_j) を分布 D から選択

• ネットワーク g_m^{t+1} から弱順位付け h_k を取得

• 更新: $\alpha_k = \frac{1}{2} \ln(\frac{1+r}{1-r})$,

ここで、 $r = \sum_{v_i, v_j} D_k(h_k(v_j) - h_k(v_i))$

• 更新: $D_{k+1}(v_i, v_j) = \frac{D_k(v_i, v_j) \exp(\alpha_k(h_k(v_i) - h_k(v_j)))}{Z_k}$

ここで、 Z_k は正規化項

$Z_k = \sum_{v_i, v_j} D_k(v_i, v_j) \exp(\alpha_k(h_k(v_i) - h_k(v_j)))$

end for

最終的な順位を出力: $H(v) = \sum_{k=1}^M \alpha_k h_k(v)$

習器) である．また、 $\delta(\pi)$ は π が成立するとき1の値をとり、それ以外の場合は0の値をとる関数である．式(7)の $D(\cdot)$ は事例ペアの重みの分布を表した関数であり、初期値を任意に設計できる．今回はフィードバック関数 $\Phi(\cdot)$ を $\Phi: \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ として、 x_1 が x_0 より順位が高ければ、 $\Phi(x_0, x_1) > 0$ となるように設定する．つまり、 x_0 と x_1 の順序を逆に予測する順位学習器 H には大きい損失が与えられる．また、 h_k を k 番目の弱学習器からの出力(今回はノードの順位)とした場合、順位損失の上限 Z_k を

$$Z_k = \sum_{x_0, x_1} D_k(x_0, x_1) \exp(\alpha_k(h_k(x_0) - h_k(x_1))) \quad (8)$$

と定義することができる．ここで、 α_k は各弱学習器の結果 h_k の重みであり、順位損失 Z_k を最小化するように選択することで、正しい順位を出力する関数を学習する．

3.4 リンク予測器の重み付け

本節では、前節で紹介したRankBoostを用いたリンク予測器の重み付け方法について述べる．まず、ある時刻 t におけるネットワーク g^t を訓練データとして与え、時刻 $t + \Delta t$ のネットワーク $g^{t+\Delta t}$ の予測を行う K 個のリンク予測器 $\{L\}_K$ を作成する．これらのリンク予測器は、ノードの集合 \mathcal{V} にあるすべてのノードペアについて、リンクが発生する確率を与える．次に、作成したリンク予測器によって得られたネットワークから各ノードの中心性に基づく重要度を求める．このリンク予測器に対応するノードの重要度の並び(順位)が現在の重要度の並びに近づくようにRankBoostを用いて学習する．まず、分布 D をフィードバック関数 $\Phi(\cdot)$ で初期化する．今回、各中心性に

基づいてノードに順位を与える関数を $rank(\cdot)$ とし、順位差が開いたノードペアに対して重みが大きくなるように、フィードバック関数を $\Phi(v_i, v_j) = rank(v_j) - rank(v_i)$ と定義する。次に、以下の手順を K 回繰り返すことで、各リンク予測器に対する重みを求める。分布 D からノードペア (v_i, v_j) を選び、 k 番目のリンク予測器 L_k によって予測されたネットワーク $g_k^{t+\Delta t}$ から各中心性によるノードの順位の結果 $h_k(\cdot)$ を取得する。順位の一致度を分布 D_k で重み付けした総和 r を取得した後に、 r を使い k 番目のリンク予測器の重み α_k を決定する。そして、分布 D を α_k と予測によるノードペアの順序で更新する。 K 個すべてのリンク予測器の結果に対して学習を行って得られた重み $\alpha_1, \alpha_2, \dots, \alpha_K$ を用いて、各リンク予測器から得たノードの順位 $h_k(\cdot)$ を重み付けし、最終的なノードの順位の結果を得る。本手法では、各リンク予測器が出力したノードの順位の並びが RankBoost の事例となる。ここで、リンク予測器が予測したリンク有りの確率が高い上位 l 本を加えてネットワークを予測し、そのネットワーク中のノードの重要度を中心性 c の定義に従って求める。そして、重要度の高い上位 p 件のノードの順位を事例の特徴として学習することで、上位のノードを正しく予測できるリンク予測器により大きい重みを与える。Algorithm1 に RankBoost を用いたリンク予測器の重み付けを行う擬似コードを示す。ちなみに、本稿では提案手法をリンクに方向がない無向グラフに対して適用するが、リンク予測の手法を有向の関係に拡張するだけで、容易に有向グラフに適用できる。

4. 評価実験

本章では、ノードの重要度を予測する方法として、リンク予測を行ってから間接的にノードの重要度を予測する方法と RankBoost を用いた予測方法との比較を行う。まず、リンク予測の精度について評価し、その次にノードの重要度予測について評価する。これらの実験から、リンク予測がノードの重要度予測に有効であり、ノードの重要度とリンク予測を組み合わせたことがノードの順位予測に有効であることを示す。

4.1 実験設定

本節では、arXiv (hep-th) のデータセットを用いた将来の共著関係と著者の重要度予測に関する実験設定を述べる。まず、著者の重要度予測を行うために著者間の関係をリンク予測により推定する。リンク予測に用いるデータは 1994 年までに存在する著者 2,950 人に限り、新たに出現する著者は考慮せず、訓練データとテストデータを作成する。訓練データは 1994 ~ 1998 年の全著者間のリンク指標 (3.1 節) を特徴として抽出し、1999 年の共著関係をクラスラベルとする。共著関係があれば、その著者間の関係を正例、共著関係がなければその著者間の関係を負例とする。また、RankBoost を使う場合は、1999

年時点での任意の中心性による著者の順位を教師データとして用いる。なお、各リンク予測の結果について上位何件を使用したかは、ノードの順位予測の結果が最良になるように調整した。これらの実験設定で、3.1 節で紹介したリンク指標を用いたリンク予測手法と、ONE, MUL, RB によって得られた結果を比較し、リンク予測そのものの精度とリンク予測に基づくノードの重要度予測の精度について評価する。

4.2 リンク予測の精度

リンク予測に基づくノードの重要度予測では、2.3 節で示したように、リンク予測の精度がノードの重要度予測と密接に関係があることが分かっている。そこで、まず各リンク予測手法のリンク予測精度について観察する。3.1 節で紹介したリンク指標を用いたリンク予測手法と ONE, MUL, RB のそれぞれのリンク予測の精度について評価する。リンク予測の精度は 2000 年から 2003 年までに発生した共著関係を正例とし、リンクが発生しなかった著者の組を負例とする。評価指標としては、ROC 曲線下の面積 (AUC: Area Under the ROC Curve) を用いる。AUC は 0 から 1 の値をとり、予測器が正例を負例よりも上位に順位付けできる度合いを表している。正例と負例をランダムに分類する予測器の AUC は 0.5、正例と負例を完璧に予測できる場合は 1 の値をとる。AUC の値が 1 に近いほど、そのリンク予測手法は正リンクを負リンクよりも優先的に予測していると見なせる。

MUL のリンク予測では、複数のリンク予測器から取得したリンク有りの確率の平均をリンク予測の結果とする。RB に関しては、RankBoost で得られた弱学習器の重み α をリンク予測の結果の重みとして、加重平均をリンク予測の結果とする。ここで、ネットワークを予測するために追加したリンク数 (l)、ノードの順位付けに用いたネットワークの中心性の種類 (c)、上位何件のノードを用いるか (p)、で結果が異なる。今回は RankBoost がリンク予測およびノードの重要度予測に有効であるかどうかを検証したいので、結果が一番良いパラメータ ($c = \text{Betweenness}$, $l = 1150$, $p = 10$) を用いる。

各リンク予測手法の AUC を表 1 に示す。表中の JAC は Jaccard 係数, COM は共通隣接指標, ADA は Adamic/Adar, PRE は優先的選選指標, KAT は Katz $_{\gamma}$ 指標, ONE はリンク予測器単体, MUL はリンク予測器複数, RB は RankBoost を用いたリンク予測手法である。AUC の高かった順にリンク指標を並べると、KAT > PRE > RB = MUL > ONE > ADA > COM > JAC となり、Katz $_{\gamma}$ 指標が最良のリンク予測精度を示した。ONE と MUL と RB が PRE や KAT と比べて劣っているのは、これらの手法がすべてのリンク指標を特徴として使っているため、COM, JAC, ADA の影響を受けて予測精度が落ちたためと考えられる。

次に、学習に用いた特徴が同じであり、学習方法が異なる 3 つの方法 ONE, MUL, RB を比較する。リンク予

表1 リンク予測手法の AUC

JAC	COM	ADA	PRE	KAT	ONE	MUL	RB
0.6561	0.6564	0.6564	0.8435	0.8475	0.7646	0.8066	0.8069

表2 各リンク予測手法と RankBoost によるノードの重要度予測の結果

Top	Centrality	BASE	JAC	COM	ADA	PRE	KAT	ONE	MUL	RB
5	Betweenness	0.0000	0.0000	0.0000	0.2000	0.2000	0.2000	0.0067	0.0000	0.9933*
5	Closeness	0.2000	0.6000	0.6000	0.8000	0.7379	0.8000	0.5133	0.8000	0.8000
5	Degree	0.4444	0.5893	0.5893	0.6804	0.4444	0.4444	0.5404	0.7379	0.9487*
5	PageRank	0.2000	0.4000	0.2000	0.4000	0.2000	0.2000	0.3133	0.4000	0.8200*
10	Betweenness	0.2889	0.3778	0.4222	0.4667	0.4667	0.5111	0.3956	0.4222	0.6326*
10	Closeness	0.5111	0.6000	0.5111	0.6444	0.6293	0.6444	0.5111	0.5556	0.6919*
10	Degree	0.7383	0.7916	0.8236	0.8837	0.7621	0.8098	0.7944	0.8866	0.9290*
10	PageRank	0.4667	0.5111	0.6000	0.6444	0.4667	0.5111	0.5541	0.6889	0.8000*
20	Betweenness	0.3474	0.3895	0.3789	0.4526	0.4526	0.4421	0.3828	0.4000	0.5382*
20	Closeness	0.3747	0.4656	0.4274	0.4697	0.4380	0.4274	0.3811	0.4063	0.5439*
20	Degree	0.6093	0.6223	0.6407	0.6486	0.6278	0.6748	0.6225	0.6319	0.6815*
20	PageRank	0.6526	0.6947	0.7579	0.7263	0.6526	0.6947	0.6712	0.7368	0.7632*

測器単体 (ONE) と比べて、複数のリンク予測器の出力を平均した結果 (MUL) と RankBoost を用いたリンク予測 (RB) は AUC の値が向上した。この結果に対して t 検定を行ったところ p 値が 0.05 未満となり統計的に有意であった。また、MUL と RB の結果を比較したところ有意差は認められなかった。これより、複数のリンク予測器を用いることでリンク予測の精度を向上させることができる一方で、ノードの重要度を考慮したリンク予測を行ってもリンク予測の精度に影響を与えているとは言えないことが分かった。

4.3 順位予測の精度

リンク予測を用いたノードの重要度予測を評価するため真のノードの重要度と予測した重要度のケンドールの順位相関係数で評価する。表2に、各ノード予測手法による相関係数を各ネットワーク中心性、および上位5, 10, 20 ごとに示す。表中の BASE は、リンク予測を行わなかった場合であり、1999 年と 2003 年のノードの重要を順位相関係数により比較した結果である。太字の部分が各手法の内、一番良い結果を示した値である。対応するパラメータを7章に示す。また、*の印がついた部分は、二番目に良かった結果に対して Wilcoxon 検定を用いた結果、有意 ($p < 0.05$) な差があったことを示している。

予測を行わない BASE の結果と他のリンク予測を行った結果を比べると、すべての場合においてリンク予測を行うことにより相関係数の値が同等かそれ以上に高くなっている。このことから、リンク指標を用いたリンク予測を行うことで、未来のノードの順位を予測できることが分かった。

全体的な傾向として、各手法の予測結果は何も予測しなかった時の予測結果 (BASE) を反映している。例えば、

RB の Degree と PageRank は、Top10 では前者の方が良いにもかかわらず、Top20 では逆転している。BASE の結果でも同じことが起きており、他のリンク指標の結果からも同様の現象を観測できる。しかし、RB の Betweenness の Top5 は BASE での値が低くても高い値を示している。これは RB によるノードの重要度を考慮した順位の重み付けが有効に働いたためと考えられる。事実、Top5 の予測精度は Betweenness 以外の中心性についても精度が良い。

続いて、各中心性の定義とリンク予測との関係を考察する。Degree や PageRank はノードに接続するリンク数が多いほど値が大きくなるので、これらの重要度を予測するためには、将来的に多くのノードと隣接するノードへのリンクを予測することが必要である。一方、Betweenness や Closeness の値が高いノードは任意のノードペア間の最短パス上にあるか、他のノードと距離が近くあることを示している。そのため、これらの重要度の予測に必要なことはノード間の最短パスを予測することである。

今回のリンク予測では、3.1 節で説明したリンク指標を用いている。各リンク指標の値は、ノードペアの周辺リンク構造によって決まるので、ノードに接続するリンク数や他のノードからの近さなどが指標の値に影響する。つまり、リンク指標を用いたリンク予測にはノードの性質が考慮されることになる。例えば、共通隣接指標 (COM)、Jaccard 係数 (JAC)、Adamic/Adar (ADA)、Katz $_{\gamma}$ (KAT) はノード間に共通して隣接するノードが多いほど高い値を示すリンク指標である。隣接するノードが多いということは、既に多くのノードと関係を持っており、Degree や PageRank が高いことを示している。BASE と比べて COM, JAC, ADA, KAT によってノードの重要度予測の精度が向上したのは、既に Degree や

PageRank の高いノード間にリンクがあると判定したためと考えられる。これは 2.4 節の結果と一致する。一方で、Betweenness や Closeness といった最短パスの予測が必要な中心性の予測では、2 つ以上のノードを介しているか、または間接的なつながりを持たないノードペアに対してもリンク予測を行わなければなりません。そのため、COM, JAC, ADA と違いノード間の間接的なつながりを考慮しない優先的選択指標 (PRE) や 2 つ以上のノードを介したノード間のつながりを考慮できる KAT が COM, JAC, ADA と同等かそれ以上の予測結果を示したと考えられる。

次に、リンク予測の精度の差によるノードの順位予測の精度差について観察する。表 1 と表 2 から、リンク指標を用いた場合において、リンク予測の精度が良いからといって、必ずしもノードの順位予測の精度が良いわけではないことが分かった。例えば、COM と JAC と ADA のリンク予測精度はほぼ同じであったのに対して、ノードの順位予測精度は ADA がほとんどの中心性において良い結果を示している。これは、COM や JAC と違い ADA は次数が大きい隣接ノードに対して罰則を与えていることが原因として考えられる。このことから、リンク予測を用いたノードの順位予測ではノードペアの各次数だけでなく、ノードペアに隣接するノードの次数も影響することが分かった。すなわち、リンク指標が単に正リンクになる可能性の高いノードペアを選んでいるのではなく、リンクに接続するノードの性質も同時に考慮してリンク予測を行っている結論付けることができる。そのため、2.3 節で示した「正しいリンク予測を行えばノードの順位予測精度が上がる」という観測とは異なる結果となった。

ONE と MUL と RB は同じ特徴を使用しており、学習方法だけが違う。ONE と MUL を比べた場合、すべての順位、ネットワーク中心性において MUL が一番良い結果を示している。これは MUL は ONE と比べて、リンク予測の精度が高いためであり、2.3 節の結果と一致する。MUL と RB を比較すると、順位に関わらずほぼすべて中心性において RB が MUL よりも精度良くノードの順位を予測できている。RB と MUL は同じリンク予測器を用いて、違いとしては、RB がノードの重要度を用いてリンク予測器の出力したノードの順位を利用してリンク予測器を重み付けしているところである。つまり、ノードの重要度を教師情報として用いることで、より正確にノードの順位を予測できたのである。この結果は 2.4 節の実験結果に合致する。

5. 関連研究

従来から、ある時点のネットワークを対象として、何らかの組織やグループの中から中心的な存在、他者に影響を与える存在をネットワーク的な視点からとらえようとする研究が盛んに行われてきた。その中でもオブジェクト同士の関係をネットワークとみなし、オブジェクトの重要

度を再帰的に計算する方法、例えば、PageRank [Brin 98] や HITS [Kleinberg 99] などは、ネットワークデータであれば種類を問わず適用できることから、独自に改良され応用されてきた [Balmin 04, Haveliwala 03, Lempel 00]。人物の重要度に基づく順位付けに関する研究としては、Q & A フォーラム [Zhang 07] や、共著ネットワーク上の著者の順位付け [Liu 05]、企業の順位付け [Jin 08] などがある。しかし、これらの研究はある時間のオブジェクトの重要度や影響度だけに注目しており、時間が経つごとに構造が変化するネットワーク中のノードに対して、直接適用できない。一方、時系列ネットワーク中で、中心的な存在を順位付けする手法もある [O'Madadhain 05, Sayyadi 09]。O'Madadhain らは、過去に起こったオブジェクト間の関係を考慮して、再帰的にオブジェクトの重要度を求めている。また、Sayyadi が提案した FutureRank は、論文と著者、論文の発行年に応じた論文の重要度を定義し、逐次的に重要度を更新することで、重要な論文・著者の予測を行っている。しかし、ネットワーク上の重要度の指標は、ネットワークからどのような情報を取り出すかによって変わる。従来の手法は特定の影響・重要度の予測を目的としているため、ネットワークの構造によって決まる任意の重要度の指標 (ネットワーク中心性) を予測できない。本論文で提案する手法は、ネットワーク全体を予測するため、リンク構造により定義されるオブジェクトの将来の重要度を予測することができる。

6. おわりに

本論文では、リンク予測の問題を解くことでノードの順位予測を行うモデルを提案し、共著ネットワークから、将来的に重要または影響力を持つ著者を推定した。arXiv (hep-th) のデータセットから作成した共著者ネットワークを基に、ネットワーク中心性に基づく著者の順位の変化を各年代ごとに調べた。その結果、著者の重要度は変化していることが分かった。さらに著者をノード、著者間の関係をリンクと見なした場合、将来的に発生するリンクが著者の重要度に与える影響についても確認を行った。結果、将来的に発生するリンクを正しく予測できるほど、ノードの重要度を正確に予測できることが分かった。これらの結果から、著者間の関係予測を将来の著者の順位予測に適応させる手法と、さらに、RankBoost を用いてノードの重要度とリンク予測を組み合わせたノード順位の予測手法を提案した。評価実験を行ったところ、リンク予測単体の問題を解くよりも、著者の将来的な重要度をより精度良く推定することができた。今後の課題として、大規模なデータに対する実験および論文の引用関係や World Wide Web のリンク構造に代表される有向グラフに対する本手法の適用が挙げられる。また、今回は共著関係の回数や不活性化といったリンクの強度を考慮していないので、リンクの強度を考慮した重要度の予測についても取り組んでいきたい。

7. 付 録

本稿で提案した RankBoost のパラメータを表 2 に対応して表 3 に示す。

表 3 RankBoost のパラメータ

Top	Centrality	l	p	c
5	Betweenness	900	5	Bet
5	Closeness	1000	5	Clo
5	Degree	1350	10	Bet
5	PageRank	1050	10	Bet
10	Betweenness	950	10	Bet
10	Closeness	850	10	Bet
10	Degree	1400	10	Bet
10	PageRank	1150	10	Bet
20	Betweenness	1150	10	Bet
20	Closeness	1100	5	Bet
20	Degree	1350	5	Clo
20	PageRank	1300	5	Clo

◇ 参 考 文 献 ◇

- [Adamic 03] Adamic, L. and Adar, E.: Friends and neighbors on the web, *Social Networks*, Vol. 25, No. 3, pp. 211–230 (2003)
- [Backstrom 11] Backstrom, L. and Leskovec, J.: Supervised random walks: predicting and recommending links in social networks, in *Proceedings of the 4th ACM international conference on Web Search and Data Mining*, pp. 635–644 (2011)
- [Balmin 04] Balmin, A., Hristidis, V., and Papakonstantinou, Y.: Objectrank: Authority-based keyword search in databases, in *Proceedings of the 30th international conference on Very large data bases*, pp. 564–575 (2004)
- [Barabási 99] Barabási, A. and Albert, R.: Emergence of scaling in random networks, *Science*, Vol. 286, No. 5439, pp. 509–512 (1999)
- [Ben-Hur 05] Ben-Hur, A. and Noble, W.: Kernel methods for predicting protein–protein interactions, *Bioinformatics*, Vol. 21, No. 1, pp. 38–46 (2005)
- [Brin 98] Brin, S. and Page, L.: The Anatomy of a Large-Scale Hypertextual Web Search Engine, in *Proceedings of the 7th international conference on World Wide Web* (1998)
- [Broder 00] Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A., and Wiener, J.: Graph structure in the web, *Computer networks*, Vol. 33, No. 1–6, pp. 309–320 (2000)
- [Chawla 04] Chawla, N. V., Japkowicz, N., and Kotcz, A.: Editorial: special issue on learning from imbalanced data sets, *ACM SIGKDD Explorations Newsletter*, Vol. 6, pp. 1–6 (2004)
- [Freeman 79] Freeman, L.: Centrality in Social Networks Conceptual Clarification, *Social Networks*, Vol. 1, No. 3, pp. 215–239 (1979)
- [Freund 03] Freund, Y., Iyer, R., Schapire, R., and Singer, Y.: An efficient boosting algorithm for combining preferences, *The Journal of Machine Learning Research*, Vol. 4, pp. 933–969 (2003)
- [Getoor 05] Getoor, L. and Diehl, C.: Link mining: a survey, *ACM SIGKDD Explorations Newsletter*, Vol. 7, No. 2, pp. 3–12 (2005)
- [Haveliwala 03] Haveliwala, T.: Topic-sensitive pagerank: A context-sensitive ranking algorithm for web search, *IEEE transactions on knowledge and data engineering*, Vol. 15, No. 4, pp. 784–796 (2003)
- [Jeong 00] Jeong, H., Tombor, B., Albert, R., Oltvai, Z., and Barabási, A.: The large-scale organization of metabolic networks, *Nature*, Vol. 407, No. 6804, pp. 651–654 (2000)
- [Jin 08] Jin, Y., Matsuo, Y., and Ishizuka, M.: Ranking entities on the web using social network mining and ranking learning, in *Proceedings of the WWW 2008 workshop on Social Web Search and Mining*, pp. 1349–1352 (2008)
- [Katz 53] Katz, L.: A new status index derived from sociometric analysis, *Psychometrika*, Vol. 18, No. 1, pp. 39–43 (1953)
- [Kendall 38] Kendall, M.: A new measure of rank correlation, *Biometrika*, Vol. 30, No. 1/2, pp. 81–93 (1938)
- [Kleinberg 99] Kleinberg, J.: Authoritative sources in a hyperlinked environment, *Journal of the ACM (JACM)*, Vol. 46, No. 5, pp. 604–632 (1999)
- [Lempel 00] Lempel, R. and Moran, S.: The stochastic approach for link-structure analysis (SALSA) and the TKC effect1, *Computer Networks*, Vol. 33, No. 1–6, pp. 387–401 (2000)
- [Liben-Nowell 07] Liben-Nowell, D. and Kleinberg, J.: The link-prediction problem for social networks, *Journal of the American Society for Information Science and Technology*, Vol. 58, No. 7, pp. 1019–1031 (2007)
- [Liu 05] Liu, X., Bollen, J., Nelson, M., and Sompel, Van de H.: Co-authorship networks in the digital library research community, *Information Processing & Management*, Vol. 41, No. 6, pp. 1462–1480 (2005)
- [Liu 07] Liu, T., Xu, J., Qin, T., Xiong, W., and Li, H.: Letor: Benchmark dataset for research on learning to rank for information retrieval, in *Proceedings of SIGIR 2007 workshop on learning to rank for information retrieval*, pp. 3–10 (2007)
- [Manning 99] Manning, C. and Schuetze, H.: *Foundations of statistical natural language processing*, MIT Press, Hayward Street Cambridge, MA, USA (1999)
- [Newman 01a] Newman, M.: Clustering and preferential attachment in growing networks, *Physical Review E*, Vol. 64, No. 2, p. 025102 (2001)
- [Newman 01b] Newman, M.: Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality, *Physical review E*, Vol. 64, No. 1, p. 016132 (2001)
- [O'Madadhain 05] O'Madadhain, J., Hutchins, J., and Smyth, P.: Prediction and ranking algorithms for event-based network data, *ACM SIGKDD Explorations Newsletter*, Vol. 7, No. 2, pp. 23–30 (2005)
- [Quinlan 93] Quinlan, J.: *C4.5: programs for machine learning*, Morgan Kaufmann, Waltham, MA, USA (1993)
- [Sayyadi 09] Sayyadi, H. and Getoor, L.: Future Rank: Ranking Scientific Articles by Predicting their Future PageRank, in *Proceedings of the 9th SIAM International Conference on Data Mining*, pp. 533–544 (2009)
- [Watts 98] Watts, D. and Strogatz, S.: Collective dynamics of 'small-world' networks, *Nature*, Vol. 393, No. 6684, pp. 440–442 (1998)
- [Weng 10] Weng, J., Lim, E.-P., Jiang, J., and He, Q.: TwitterRank: finding topic-sensitive influential twitterers, in *Proceedings of the third ACM international conference on Web search and data mining*, pp. 261–270 (2010)
- [Wu 09] Wu, L., Yang, L., Yu, N., and Hua, X.: Learning to tag, in *Proceedings of the 18th international conference on World Wide Web*, pp. 361–370 (2009)
- [Yang 09] Yang, P., Liu, Q., and Metaxas, D.: Rankboost with l1 regularization for facial expression recognition and intensity estimation, in *Proceedings of the 12th International Conference on Computer Vision*, pp. 1018–1025 (2009)
- [Zhang 07] Zhang, J., Ackerman, M., and Adamic, L.: Expertise networks in online communities: structure and algorithms, in *Proceedings of the 16th international conference on World Wide Web*, pp. 221–230 (2007)

〔担当委員：矢田 勝俊〕

2011 年 11 月 25 日 受理

著 者 紹 介



宮西 大樹(学生会員)

平成 21 年神戸大学大学院工学研究科情報知能学専攻博士前期課程修了。同年，同大学院システム情報学研究科計算科学専攻博士後期課程進学。情報検索，データマイニングの研究に従事。情報処理学会，学生会員。



関 和広

平成 14 年図書館情報大学情報メディア研究科修士課程修了。平成 18 年インディアナ大学図書館情報学研究科博士課程修了。神戸大学助手，助教を経て現在同大学院システム情報学研究科講師。情報検索，自然言語処理，機械学習の研究に従事。Ph.D. 電子情報通信学会，自然言語処理学会，ACM SIGIR 各会員。



上原 邦昭(正会員)

昭和 53 年大阪大学基礎工学部情報工学科卒業。昭和 58 年同大学院博士後期課程単位取得退学。同産業科学研究所助手，講師，神戸大学工学部情報知能工学科助教授，同都市安全研究センター教授等を経て，現在同大学院システム情報学研究科教授。工学博士。人工知能，特に機械学習，マルチメディア処理の研究に従事。電子情報通信学会，計量国語学会，日本ソフトウェア科学会，AAAI 各会員。