Vol.2011-MPS-82 No.3
2011/3/7

ネットワークの構造解析に基づく有望ノードの同定

宮西大樹[†1]　関　和広[†2]　上原邦昭[†3]

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（**有望エンティティ**）を同定する．従来では，ある時点における著者をノード，著者同士の共著関係をエッジとした共著ネットワークから，構造的な特徴を基に重要度や影響力の大きな著者の同定を行ってきた．しかし，著者同士の関係は年を追うごとに変化しており，著者の最新の重要度や影響力を把握するためには，現時点における著者間の関係を見るだけでは不十分である．そこで，本論文では，時間とともに変化するネットワークデータを対象として，ネットワークの構造によって決定された各ノードの将来的な重要度・影響力（ネットワークの中心性）をリンク予測と RankBoost を用いることでノードの順位を予測する手法を提案する．この手法を共著ネットワークに適用することで，将来の主要な著者を予測する．arXiv（hep-th）データセットから抽出した共著ネットワークを用いた実験により，リンク予測をノードの順位予測に適用させることで AUC の高いリンク予測行うことができ，将来的なノードの順位をより正しく予測できた．

Promising Entities Discovery Based on Network Analysis

Taiki Miyanishi ,[†1] Kazuhiro Seki [†2]
and Kuniaki Uehara [†3]

This paper proposes a framework to predict future significance or importance of nodes of a network through link prediction. The network can be any kind, such as a co-authorship network where nodes are authors and co-authors are linked by edges. In this example, prediciting significant nodes may mean 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 predicitions 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，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可能となっている．ネットワークデータは化学物質，人物，論文などのエンティティ同士の関係を集約したものであるため，エンティティの相対的な関係から，影響力を持つエンティティや重要な役割を果たす関係が自然と決まってくる．データマイニングの分野では，こういった重要で影響力を持ったエンティティ，またはエンティティ間の関係をネットワークの観点から解析・予測する技術（リンクマイニング）が注目されている．Getoor[8] によると，リンクマイニングのタスクは，リンク構造を用いたノードの分類・順位付け・クラスタリング，リンク予測，構造パターンの発見に分けることができる．本研究では，著者をノード，著者間の共著関係をエッジとした共著ネットワークに対して，リンクマイニングの手法を適用することで，個別に見るだけでは分からない有益な情報の取得を目的とする．

現在まで，研究者・産学官の連携支援を目的とした研究者ネットワーク検索エンジン「Polyphonet」[15] や，将来的に発生する共著関係を既知の共著関係から予測する研究[14]，また，共著ネットワークにおける相対的な関係から，論文の著者が持つ影響力に応じて順位付けを行う研究がなされてきた．いずれも，個人を取り巻くネットワークから有益な情報を取り出すことを目的としており，我々の目的と近い．しかし，これらの研究はいずれもある時間における人間関係や人物の重要性に注目しており，また将来的なネットワークを考える上でも人物の重要度と人物の関係を個別に考えている．本来，人間関係に基づくネットワークは，

†1 神戸大学大学院工学研究科
Graduate School of Engineering, Kobe University
†2 神戸大学自然科学系先端融合研究環
Organization of Advanced Science and Technology, Kobe University
†3 神戸大学大学院システム情報学研究科
Graduate School of Systems Informatics, Kobe University

時間が経つにつれて変化し，個人の重要度や影響力は個人間の関係と相互に影響し合い変化するはずである．

そこで，本研究では，将来的に発生する著者間の関係の予測を著者の影響力・重要度に基づく順位の予測に適応させる手法を提案し，主要な著者の予測を行う．なお，本研究で提案する手法は，共著ネットワークに限らず，ネットワーク構造を持つデータならばどのようなデータに対しても適用可能である．

本論文では，まず 2 章で関連研究について紹介し，3 章において有望エンティティを定義，およびその具体例として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述べる．4 章では，有望エンティティを同定するためのリンク予測とノードの順位予測の方法について述べる．そして，5 章では共著ネットワークにおける重要な著者の予測について，arXiv(hep-th)[1] のデータセットを用いた実験を行い，6 章でまとめと今後の課題について述べる．

2. 関 連 研 究

従来から，ある時間におけるネットワークを対象として，何らかの組織やグループにおいて中心的な存在，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．その中でもネットワーク中のエンティティの重要性を再帰的に計算する方法，例えば，PageRank[5] や HITS[12] 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[3),9),13)]．しかし，これらはある時間のエンティティの重要性や影響度だけに注目しており，時間が経ごとに変化するネットワーク（動的ネットワーク）中のエンティティに対しては直接適用できない．一方，動的ネットワークの中から，中心的な役割を果たす存在の順位付けを行う手法として，O'Madahain らの手法[17]，Sayyadi らによる FutureRank[19] がある．O'Madahain らは，過去に起こったエンティティ間の関係を考慮して，再帰的にエンティティの重要性を求めていくため，ネットワークの時間的な変化を考慮した重要性を求めることができる．また，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，ネットワークにおける重要性の指標は，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性の指標を予測できない．本論文で提案する手法は，ネットワーク自体を予測するため，予測したい指標がネットワークの構造によって決まりさえすれば，将来のエンティティの重要度を予測できる．

3. 有望エンティティ

本研究の目的は将来的に影響・重要度を持つ著者の同定である．このような著者を「有望エンティティ」と呼ぶ．ここでいう有望さは，将来的に著者が，どれだけ重要または影響力を持つかを定量化した値である．この指標は，共著ネットワークを研究者同士のつながりと考えると，研究者同士の集まりの中で，研究者個人の立ち位置や他の研究者に与える影響度などを表す．本論文ではネットワーク中心性をエンティティの重要度と見なす．

3.1 ネットワークの中心性

ネットワークの中心性とは，ネットワークの構造から，ネットワークを構成するノードがネットワーク中で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表す指標である．この指標は社会学で古くから使用されており，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示す．ここで，E と V はそれぞれネットワークを構成するエッジの集合とノードの集合を表す．

- **次数中心性**[6]
  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，自身に隣接するノード数で定義される．つまり，ノード i が与えられたときのエッジ $e_{ij} \in E$ の数に相当する．
- **近接中心性**[6]
  ネットワーク内におけるコミュニケーションの効率を表し，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 $\frac{1}{\sum_{j \in V, j \neq i} \frac{d_{ij}}{n-1}}$ に相当する．
- **媒介中心性**[6]
  情報伝達におけるフローのコントロール可能性を表し，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される．全てのノードの組 j, k $(j, k \in V, j \neq i, k \neq i)$ の最短パス $path_{jk}$ にノード i が含まれる数に相当する．
- **PageRank**[5]
  ネットワーク中においてノード間を確率的に遷移するランダムサーファーを考えた場合，ランダムサーファーがあるノードを訪ねる確率を表す指標．重要なノードに多く隣接するノードは重要という考えのもと定義されている．各ページの PageRank をベク

Vol.2011-MPS-82 No.3
2011/3/7

トル表現した PageRank ベクトルは $\vec{x} = \alpha \mathbf{P}^T + (1-\alpha)\frac{\vec{1}}{n}$ として表すことができる．ここで，$\mathbf{P}$ は遷移行列，n はノード数．α は，ランダムサーファーのテレポーテーションを制御するパラメータであり，本論文では Brin[5] にならい $\alpha = 0.85$ とする．

論文の著者をノード，論文の共著関係をエッジとした共著ネットワークを考えた場合，次数中心性は今までに共同研究を行い協力関係にあった研究者の数を表している．共著関係が研究者同士のつながりを表すと考えると，近接中心性は研究者を通してどれだけ効率的に多くの研究者と知り合えるかを表す指標となっている．また，媒介中心性は，共著関係を情報伝達の経路として考えた場合に，情報伝達の要所となる人物を同定するために役立つ．PageRank は研究者コミュニティーの中で顔の広い人物を表す指標となっている．

本研究で用いる共著ネットワークでは，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も変化すると考えられる．そのため，上記に示した中心性の値も変化することが予想される．

3.2 エンティティの順位の変化

本節では，前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，順位の予測に意味があるかを知るため，各年代ごとの著者の順位の遷移を見ていく．データには，arXiv（hep-th）[1] 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 8392 人の著者が存在し，総共著数は 87794 回，同著者間の共著回数を除くと 20387 個の共著関係が存在する．また，この共著ネットワークは時間の経過と共に共著が増加し続け，ネットワーク構造が年を経るごとに変化していく．

著者の順位の変化を見るために，1994 年までデータベース中に存在する全著者 2950 人を対象として，2003 年までの順位の変化を見ていく．順位の変化は，1994 年の順位と 1994〜2003 年における著者の順位について，ケンドールの順位相関係数 τ[11] を見ていくことで確認する．相関係数 τ の値が 1 に近いほど，互いの順位の並びは似ており，0 に近いほど似ていないことを示している．順位を相関係数を用いて比較する際，上位何件までを見るかで値が異なるため，今回は上位 100，2000 件について相関係数の変化を見ていくことにする．図 1 に，ネットワーク中心性に基づく順位の変化を上位別，年代ごとに示す．グラフから，上位，下位に関わらず年代を経るごとに相関係数が変化していることが分かる．

3.3 エンティティの順位とリンク発生の関係

次にノードの順位がリンクの発生とどのような関係があるかを確認する．前述したように，ノードの順位はネットワークの中心性によって決まる．ネットワークの中心性はネットワークの構造，つまりノード間のリンクのつながり具合によって決定される．よって，リンクの発生を予測できさえすれば，ネットワークの中心性により決まるノードの順位を予測できると考えられる．また，リンク予測の精度によってノード順位の予測精度の上限が決まるかも確認する．この考えを確かめるため，以下の予備実験を行う．まず，1994〜2003 年の間にリンクが発生するノードの組を正リンク，リンクが発生しないノードの組を負リンクとする．次に，1994 年のネットワーク対して，正リンクと負リンクを $\alpha : 1-\alpha$ $(0 \leq \alpha \leq 1)$ の割合で合計 K 本づつ加えていく．$\alpha = 1$ であれば，実際に発生するリンクを K 本づつ加えることになり，$\alpha = 0$ であれば，実際には発生しないリンクを K 本づつ加えることになる．α の値が大きいほど正しいリンク予測を行うことになる．

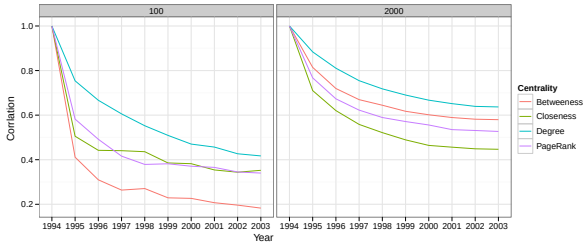

図 1 年代ごとのネットワーク中心性に基づく順位

図 2 は，1994 年の共著ネットワークに正リンクと負リンクを加えていき，ケンドールの順位相関係数を用いて，予測したネットワークの各中心性ごとの順位と，実際の 2003 年の順位とを比較したものである．順位の比較には上位 100 件までを用い，α の値は上 3 つが $0, 0.2, 0, 4$ となり，下 3 つが $0.6, 0.8, 1$ となっている．加えるリンクは正リンクと負リンクが $\alpha : 1-\alpha$ $(0 \leq \alpha \leq 1)$ の割合になるように，10 本ずつランダムに選び 1994〜2003 年の間に発生するリンク数まで加えてネットワークを予測する．そして，予測したネットワークを基にノードの順位を間接的に予測する．なお，リンクはランダムに選ばれるため，図示した相関係数は 10 回の試行の平均である．

図 2 より，実際に発生しないリンクを加えていった場合（$\alpha = 0$），相関係数はほぼ変化しない．誤ったリンクを予測しても，将来のノードの順位は予測できないことがわかる．一方，実際に発生するリンクを加えた場合（$\alpha = 1$），リンクを加えれば加えるほど，線形的に相関係数が増加し，正しい順位を予測できていることがわかる．また，正リンクの割合が負リンクに比べて大きくなればなるほど（α の値が 1 に近いほど），リンクを加えたときの

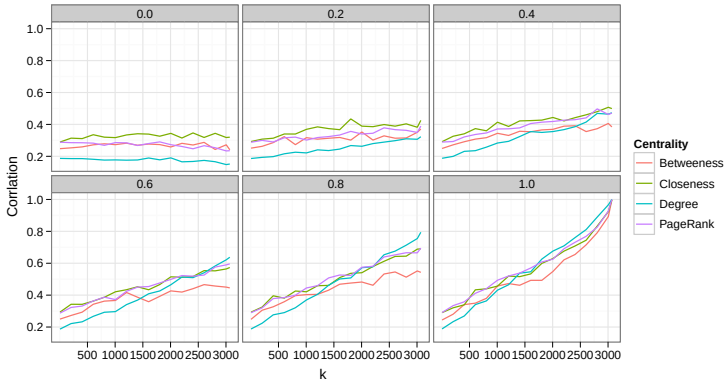

図 2 正例と負例のリンク加えて予測した順位と真の順位の相関

相関係数の上昇率も大きくなることが分かる．また，α の値によって最終的な順位予測精度の上限も決まり，リンク予測の結果がすべて正しくなければ（$\alpha < 1$），将来の順位を完全に予測することはできない（$\tau < 1$）ことが分かる．

これらの実験から，共著ネットワークにおいて著者の順位は年代を経ることに変化しており，新たに発生する共著関係の予測を正しく予測する（$\alpha > 0$）ことにより，ネットワークの中心性に基づく将来の著者の順位を正しく予測できることが確認できた．本節では，あらかじめリンクの有る無しを知った上でリンク予測を行っている．次節では，過去のデータを用いてリンク予測を行う方法について述べる．

3.4 リンク予測

未来のネットワークを予測することは，局所的にみると，任意の 2 ノード間のリンクの有無を予測することに等しい．これをリンク予測といい，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．また，ネットワークの構造からリンク予測を行う場合，各リンクを独立に予測するか，ネットワーク全体を予測するかでアプローチが変わる．今回は，ノードが持つデータの性質が明示されておらず，ネットワークの構造だけが観測できる場合を考え，各リンクを独立に予測する．

ネットワークの構造から予測する方法では，ノードの組に対してネットワークの構造を用いた特徴を定義し，従来の機械学習のアプローチを用いてリンクの有る・無しを予測する．ノードの組の特徴には，リンク指標と呼ばれる 2 ノード間の周辺のリンク構造を定量化した値を用いる．代表的なリンク指標を以下に示す．

- **共通隣接ノード（CN）指標**[16]
  ノード同士が共通の隣接ノードを多く持っているほど，2 つのノードの間にリンクが現れやすいという考えに基づく．ノード $v^{(i)}$，$v^{(j)}$ 間の共通隣接ノード指標は $\mathrm{CN}(v^{(i)}, v^{(j)}) = |\Gamma(v^{(i)}) \cap \Gamma(v^{(j)})|$ として定義される．ここで，$\Gamma(v)$ はノード v と隣接するノードの集合を表す．
- **Jaccard 係数（JAC）**[14]
  共通の隣接ノードが, 2 つのノードの隣接ノード集合に占める割合が大きいほどリンクが現れやすいことを示した指標．$\mathrm{JAC}(v^{(i)}, v^{(j)}) = \frac{|\Gamma(v^{(i)}) \cap \Gamma(v^{(j)})|}{|\Gamma(v^{(i)}) \cup \Gamma(v^{(j)})|}$ として定義される．
- **Adamic/Adar（ADA）**[2]
  重みつきの共通隣接ノード指標であり，人付き合いの良くない人を共通の友人としてもつ 2 人は，友達である可能性が高いとした指標．$\mathrm{ADA}(v^{(i)}, v^{(j)}) = \sum_{k \in |\Gamma(v^{(i)}) \cap \Gamma(v^{(j)})|} \frac{1}{\log |\Gamma(v^{(k)})|}$ として定義される．
- ***Katz*$_\beta$ 指標（KB）**[10]
  2 つのノードを結ぶパスの集合の数の重みつき和であり，共通隣接ノード指標の一般形．$\mathrm{KB}(v^{(i)}, v^{(j)}) = \sum_{l=1}^{l} |\mathrm{paths}^{(l)}_{v^{(i)}, v^{(j)}}|$ として定義される．ここで，$\mathrm{paths}^{(l)}_{v^{(i)}, v^{(j)}}$ はノード $v^{(i)}$ とノード $v^{(j)}$ を結ぶ長さ l のパス数を表す．
- **優先的選択指標（PA）**[4]
  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．隣接ノードが多いほど新たなリンクを得やすいとしたもの．$\mathrm{PA}(v^{(i)}, v^{(j)}) = |\Gamma(v^{(i)})| \cdot |\Gamma(v^{(j)})|$ として定義される．

本研究で行うリンク予測は，上記のリンク指標全てを全ノードの組に対する特徴とし，リンクの有る無しをクラスラベルとし，C4.5[18] をリンクの分類器として作成する．分類器の出力したリンクの有る無しの確率を大きい順に並べて, 適当な閾値以上のノードの組にリンクが有ると判定する．また，出力されたリンク有りの確率が高ければ高いほど実際に早期にリンクが出現すると考える．なお，リンク予測問題では正例（リンク有り）と負例（リンク無し）の数に大きな差がある（不均衡データの問題）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．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．

Vol.2011-MPS-82 No.3
2011/3/7

4. ノード順位の学習と予測

本節では，リンク予測の結果を用いてノードの順位学習を行い，将来的なノードの順位を予測する方法について述べる．順位予測を行う方法として以下の 3 つの方法を挙げる．ここでは C4.5 を 10 個バギングしたリンクの分類器をリンク予測器と呼ぶことにする．

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．

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の予測結果の平均からノードの順位を予測．

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[7] を適用する．各事例の重みを利用してリンク予測器の結果を加重平均し，ノードの順位を予測．

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 は，リンク予測の仕方は MUL と同じであるものの，リンク予測の重み付けにノード順位の情報を使う点で MUL と異なる．つまり，RB はリンク予測をノードの順位予測に適応させるモデルになっている．図 3 は RankBoost を用いて共著関係の予測を著者の順位予測に適応させる仕組みを表した図である．

4.1 RankBoost

RankBoost[7] は順序が定義された事例に対して，事例間の順序を与える関数を弱学習器とし，予測した順序と訓練データ中の順序の不一致を最小化するように学習するモデルである．順位付けを学習するために順位付け損失を

$$rloss_D(H) = \sum_{x_0,x_1} D(x_0,x_1)\delta(H(x_1) \leq H(x_0)) \quad (1)$$

$$D_k(x_0,x_1) = max(0, \Phi(x_0,x_1)) \quad (2)$$

と定義する．H_t は弱学習器の線形和 (強学習器) であり，$\delta(\pi)$ 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．$\Phi(\cdot)$ はフィードバック関数 $\Phi : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ であり，x_1 が x_0 より順位が高ければ，$\Phi(x_0, x1) > 0$ とする．また，順位損失の上限 Z_t は

$$Z_k = \sum_{x_0,x_1} D_t(x_0,x_1) \exp(\alpha_k(h_k(x_0) - h_k(x_1))) \quad (3)$$

である．ここで，h_k は k 番目の弱学習器からの出力であり，今回はリンク予測から得た著者の順位に対応する．α_k は各弱学習器の結果 h_k の重みであり，Z_k を最小化するように選ばれる．

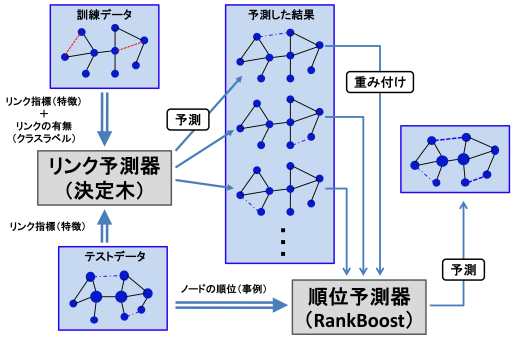

図 3 RankBoost を用いた著者の順位予測の流れ図

4.2 リンク予測器の重み付け

RankBoost を用いてリンク予測器の重み付けを行う方法について述べる．まず，ある時刻 t における共著ネットワーク g^t を与えたときに，時刻 $t + \Delta t$ の共著ネットワーク $g^{t+\Delta t}$ の予測を目的とした 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著者の集合 $\mathcal{E}$ にあるすべての著者の組に対して共著関係が発生する確率を与える．次に作成したリンク予測器を RankBoost を用いて順位付けを行う．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著者に順位を与える関数を $rank(\cdot)$ とし，フィードバック関数を順位差が開いた著者の組に対して重みが大きくなるように，$\Phi(x_0,x_1) = rank(x_1) - rank(x_0)$ 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 から著者の組 $(e_i,\ e_j)$ 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著者の順位の結果 h_k 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，r を基に 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 と予測による著者の組の順序で更新する．K 個すべての事例に対して行った結果得られた重み $\alpha_1, \alpha_2, \ldots, \alpha_K$ を用いて，各リンク予測の結果 h_k を重み付けし，最終的なリンク予測器 (SL) の結果を得る．最後に SL から共著ネットワークを構築し，間接的に著者の順位を求める．以下に擬似コードを掲載する．

Algorithm Rankboost training process

Input: Given entities $e_1, \cdots, e_m \in \mathcal{E}$,
distribution D over $\mathcal{E} \times \mathcal{E}$ and a graph g^t at time t
where
$\mathcal{E}$ is the set of the entities
Initialize $D_1 = D$
Generating link predictors $\{L\}_K$ from the graph g^t
where
$\{L\}_K$ is the function giving link occurrence probabilities to all entity pairs
Predicting K-th graph g^{t+1} from $\{L\}_K$
for $k = 1, \cdots, K$ **do**
- Select pair $(e_i, e_j) \in \mathcal{E} \times \mathcal{E}$ with dictribution D
- Get weak ranking h_k from the graph g_k^{t+1}
- Update: $\alpha_k = \frac{1}{2}\ln(\frac{1+r}{1-r})$,
  where $r = \sum_{e_i,e_j} D_k(h_k(e_i) - h_k(e_j))$
- Update: $D_{k+1}(e_i, e_j) = \frac{D_k(e_i,e_j)\exp(\alpha_k(h_k(e_i)-h_k(e_j)))}{Z_k}$
  where Z_k is a normalization factor.
  $Z_k = \sum_{e_i,e_j} D_k(e_i, e_j)\exp(\alpha_k(h_k(e_i) - h_k(e_j)))$

end for
Output the final ranking: $H(e)$ using the strong predictor $\mathbf{L} = \sum_{k=1}^{K} \alpha_k L_k$

5. 実　　験

本章では，リンク予測およびリンク予測を用いたノードの順位予測に対する RankBoost の有効性を評価する．

5.1 実験設定

本節では，arXiv（hep-th）[1] のデータセット用いた将来の共著関係と著者の順位予測に関する実験設定を紹介する．まず，ノードの順位予測を行うためにリンク予測を行う．リンク予測に用いるデータは 1994 年までに存在する著者 2950 人に限り，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（3.4 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．なお，ネットワーク構造の変化と関係のない過去に生じた共著関係については予測の対象外とする．テストデータには，1995 年〜1999 年までのリンク指標を特徴として用い，2000〜2003 年に新しくできる共著関係をクラスラベルとする．以上がリンク予測単体を用いる場合 (ONE) と，複数のリンク予測器を用いる場合 (MUL) の訓練データとテストデータの設定である．リンク予測器の結果の重み付き平均を求める場合 (RB) は，2000 年のノードの順位を使ってリンク予測器の重みを求める．2000 年の訓練データはノードの順位だけを用い，リンクに関する情報は使用しない．両方の実験設定で比較することにより，リンク予測をノードの順位情報に適応させることで，ノードの順位予測をより精度良く行えることを示す．

5.2 リンク予測の精度

ONE，MUL，RB のそれぞれのリンク予測の精度について紹介する．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 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おり，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．ONE，MUL，RB の AUC はそれぞれ，0.765，0.816，0.820 であった．RB に関して，教師として使う中心性や上位何件のノードを使うかで結果が異なる．今回は RankBoost が順位情報を上手く取り入れることができるかを見るため，結果が一番良かった中心性と上位のノードをパラメータとして用いている．

リンク予測器単体 (ONE) と比べて，リンク予測器を 30 個バギングした (MUL) が AUC の値で約 8%向上しており，複数のリンク予測器を用いることでリンク予測の精度が上がることが分かる．次に，MUL と RankBoost による加重平均 (RB) を比較する．MUL と RB は，いずれも同じリンク予測器を用いており，RB は順位情報だけを余分に保持している．結果，AUC の値は RB の方が MUL より僅かながら上回っている．このことから，順位情報がリンク予測においても有効に機能することが分かり，RankBoost が順位情報を有効に

表 1 ONE：リンク予測器単体，MUL：リンク予測器を 30 個バギング，RB：RankBoost によるリンク予測器の加重平均，それぞれの cor_{max}

手法	上位	Degree	PageRank	Closeness	Betweeness
ONE	5	1.572	3.786	12.45	0.013
ONE	10	5.382	3.947	0.003	3.621
ONE	30	0.298	0.336	0.0269	0.016
MUL	5	0.482	5.000	18.64	0.000
MUL	10	2.050	3.866	0.097	5.155
MUL	30	0.294	0.262	0.035	0.023
RB	5	1.811	4.296	15.85	0.000
RB	10	5.029	4.106	0.129	5.218
RB	30	0.322	0.306	0.046	0.039

活用できることが分かる．

5.3 順位予測の精度

リンク予測を用いたエンティティの順位予測の評価には次の指標を用いる．

$$cor_{max} = \max\{\sum_{k}^{K}(cor(k) - cor_{base})\} \quad (4)$$

ここで，$cor(k)$ はリンクを k 本予測して得たノードの順位と実際のノードの順位とのケンドールの相関係数である．また，cor_{base} は予測したリンクが 0 本の場合のノードの順位間の相関係数である．つまり，ある時刻 t のノードの順位と予測したい時刻 $t + \Delta t$ のノードの順位の相関を見ることになる．cor_{max} を用いることで，リンク有りと予測したリンクの本数にかかわらず，ノードの順位の予測性能を知ることができる．

表 1 に，2003 年の共著ネットワークを対象とした ONE，MUL，RB の各中心性に基づく，上位 5，10，30 ごとの cor_{max} を示す．ONE と MUL を比較すると，次数中心性 (Degree) においては ONE が勝っているが，それ以外の中心性においては，ONE と MUL は同等か，MUL が良い結果を残している．リンク予測器を複数使用することで，ノードの順位をより精度良く予測できている．MUL と RB の比較では，RB の方がほぼすべてにおいて勝っており，ノードの順位情報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が予測できていることがわかる．

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)[1] 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して，ネットワーク中心性に基づく著者の順位の年代ごとの変化を調べた．その結果，著者の順位は上位・下位をとわず変化していることが分かった．さらに，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順位に与える影響についても実験を行った．結果，将来的に発生するリンクを予測できるほど，著者の順位を正確に予測できることが分かった．これらの事実から，既知のネットワークを基にして，著者間の関係予測の結果に RankBoost を適用することで，将来の著者の順位を予測する手法を提案した．その結果，リンク予測単体の問題を解くよりも著者の順位予測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と，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ネットワークについても本手法を適用しいきたい．また，リンク予測とノードの順位予測を独立ではなくマルチタスク学習の枠組みを用いて同時に推定していくつもりである．

参 考 文 献

1) http://kdl.cs.umass.edu/data/hepth/hepth-info.html.
2) Lada A. Adamic and Eytan Adar. Friends and neighbors on the web. *Social Networks*, Vol.25, No.3, pp. 211–230, jul 2003.
3) A. Balmin, V. Hristidis, and Y. Papakonstantinou. Objectrank: Authority-based keyword search in databases. In *Proceedings of the Thirtieth international conference on Very large data bases-Volume 30*, pp. 564–575. VLDB Endowment, 2004.
4) A.L. Barabási, H.Jeonga, Z.Néda, E.Ravasza, A.Schubertd, and T.Vicsek. Evolution of the social network of scientific collaborations. *Physica A: Statistical Mechanics and its Applications*, Vol. 311, No. 3-4, pp. 590–614, aug 2002.
5) Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, Vol.30, pp. 107–117, apr 1998.
6) LintonC. Freeman. Centrality in social networks conceptual clarification. *Social Networks*, Vol.1, pp. 215–239, 1979.
7) Y.Freund, R.Iyer, R.E. Schapire, and Y.Singer. An efficient boosting algorithm for combining preferences. *The Journal of Machine Learning Research*, Vol.4, pp. 933–969, 2003.
8) Lise Getoor and Christopher P. Diehl. Link mining: a survey. *ACM SIGKDD Explorations Newsletter*, Vol.7, No.2, pp. 3–12, dec 2005.
9) T.H. Haveliwala. Topic-sensitive pagerank: A context-sensitive ranking algorithm

for web search. *IEEE transactions on knowledge and data engineering*, pp. 784–796, 2003.

10) Leo Katz. A new status index derived from sociometric analysis. *Psycometrika*, Vol.18, No.1, pp. 39–43, mar 1953.

11) MGKendall. A New Measure of Rank Correlation. *Biometrika*, pp. 81–93, 1938.

12) J.M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, Vol.46, No.5, pp. 604–632, 1999.

13) R.Lempel and S.Moran. SALSA: the stochastic approach for link-structure analysis. *ACM Transactions on Information Systems (TOIS)*, Vol.19, No.2, pp. 131–160, 2001.

14) David Liben-Nowell and Jon Kleinberg. The link prediction problem for social networks. *Proceedings of the twelfth international conference on Information and knowledge management*, pp. 556–559, 2003.

15) Y. Matsuo, J. Mori, M. Hamasaki, T. Nishimura, H. Takeda, K. Hasida, and M.Ishizuka. POLYPHONET: an advanced social network extraction system from the web. *Web Semantics: Science, Services and Agents on the World Wide Web*, Vol.5, No.4, pp. 262–278, 2007.

16) M.E.J. Newman. Clustering and preferential attachment in growing networks. *Physical Review E*, Vol.64, No.2, p. 025102, jul 2001.

17) Joshua O'Madadhain, Jon Hutchins, and Padhraic Smyth. Prediction and ranking algorithms for event-based network data. *ACM SIGKDD Explorations Newsletter*, Vol.7, No.2, pp. 23–30, dec 2005.

18) J.R. Quinlan. *C4. 5: programs for machine learning*. Morgan Kaufmann, 1993.

19) H.Sayyadi and L.Getoor. FutureRank: Ranking scientific articles by predicting their future PageRank. In *Proc. of the 9th SIAM International Conference on Data Mining*, pp. 533–544. Citeseer, 2009.