リンク予測に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

清水　豪士

富山県立大学　情報基盤工学講座
t715038@st.pu-toyama.ac.jp

July 7, 2021

はじめに

背景

- ネットワークデータでは人物，論文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出すことができる．
- 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し，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ノードの重要性も時間と共に変化すると考えられる．

目的

- 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

ネットワーク中心性

ネットワーク中心性

- ネットワーク中のノードがどの程度の影響力持っているか，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.
- 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.

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.
- 自信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.

近接中心性（Closeness）

- 近接中心性は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.
- 任意のノードに到達するための最短パスの平均で定義する.

ネットワーク中心性

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のかを表す指標.
- 自身に接続するノード数で定義する.

PageRank

- PageRank 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ーを考えた場合，ランダムサーフ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.

ネットワークの中心性は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を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.

提案手法

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り，リンク予測と呼ばれる．

リンク予測

- ノードの持つ情報から未来のネットワークを予測する方法
- ネットワークの構造から予測する方法

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．

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うち，互いに隣接するノードの数の割合を示した指標．

提案手法

Adamic/Adar（ADA）

- 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．

$Katz_\gamma$（KAT）

- 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系であるといえる．

優先的選択指標（PRE）

- 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．
- ノードが他の多くのノードと隣接していればいるほど，そのノードは将来的に新しいノードを獲得しやすい指標．

提案手法

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．

本研究では，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 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

- 決定木を生成するためのアルゴリズム
- 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あると判定する．

提案手法

ノードの順位学習と予測

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，より精度よくノードの重要度予測ができるのではないかと考える．

ノードの重要度予測に有効なリンク予測の結果を使い，ノードの重要度の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．

- リンク予測器単体（ONE）
- リンク予測器複数（MUL）
- RankBoost（RB）

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．

RB 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．

提案手法

RankBoost

RankBoost は順位学習アルゴリズムの 1 つで，情報検索，タグ推薦，表情認識の分野で，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関数器として用い，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 の損失関数は以下の式のように定義される．

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

RankBoost

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

上の式は事例ペアの重みの分布を表した関数である．

また，次の式を h_k を k 番目の弱学習器からの出力とした場合，順位損失の上限として定義する．

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

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．

提案手法

RankBoost を用いたリンク予測器の重み付け

リンク予測器の重み付け

ある時刻 t におけるネットワーク g^t を空連データとして与え，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．
リンク予測器はノードの集合 γ 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．

続き

まず，分布 D をフィードバック関数 $\Phi()$ で初期化する．
各中心性に基づいてノードに順位を与える関数を $rank()$ 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．

提案手法

続き

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 $h_k()$ 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 $\alpha_1, \alpha 2, ..., \alpha_K$ を用いて，各リンク予測器から得たノードの順位 $h_k()$ を重み付けし，最終的なノードの順位の予測結果を得る．

評価実験

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.

実験設定

arXiv のデータセットを用いて，将来の共著関係と著者の重要度予測を行う.

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.

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する.

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.

実験設定

リンク予測の精度

評価指標として，ROC 曲線下の面積（AUC）を用いる．

AUC は 0 から 1 の値を取り，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．

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みなせる．

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無の確率の平均をリンク予測の結果とする．

RB 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，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，上位何件のノードを用いるかで結果が異なってくる．

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を用いる．

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

図 1: リンク予測手法の AUC

AUC の高い順に並べると KAT が一番高くなり，$Katz_\gamma$ 指標が最良のリンク予測精度を示した.

ONE，MUL，RB を比較すると，ONE と比べ MUL の結果と RB のリンク予測は AUC の値が向上した.

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れ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わかった.

順位予測の精度

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***

図 2: 各リンク手法と RankBoost によるノードの重要度予測の結果

BASE はリンク予測を行わなかった場合である．
太字が各手法のうち，一番良い結果を示した値である．

順位予測の精度

- BASE 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わかった．
- 図 2 の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ときの予測結（BASE）を反映している．

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．
これは，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．

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.

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に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.

分析

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，ノードに卒属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．

COM，JAC，ADA，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，すで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．

一方で Betweenness や Closeness といった最短パスの予測が必要な中心線の予測では，2 つ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いもリンク予測を行う必要がある．
そのため，ノード間の間接的なつながりを考慮しない 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ている．

おわりに

まとめ

- 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.
- 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課題として挙げられる.

TOYAMA Prefectural University

実験結果

TOYAMA Prefectural University

実験結果