

リンク予測に基づく有望ノードの同定

Promising Nodes Detection using Link Prediction

宮西大樹*1

Taiki Miyanishi

関和広*2

Kazuhiro Seki

上原邦昭*3

Kuniaki Uehara

*1*3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

本研究では、ネットワークの構造が時間と共に変化する動的なネットワークを対象にして、現在は重要とみなされていないが、将来において重要とみなされるようなノード（有望ノード）を同定する手法について述べる。有望ノードの同定には、ネットワークのリンク構造を用いたリンク予測と、ネットワークの中心性をノードの重要度とみなしたノードの順位付けの2つの手法を用いて行う。また、有望ノードを同定する際に、リンク予測の精度が有望ノードの同定に及ぼす影響についても考察する。

1. はじめに

近年、ソーシャルネットワークサービス（SNS）における人間関係、生物学におけるタンパク質間の相互作用など、ネットワーク構造で表現されるデータの解析が注目されている。特に、データマイニングの分野では、このようなネットワークで表現されるデータの解析をリンクマイニングと呼んでいる。この分野の第一人者である Getoor ら [5] は、このリンクマイニングのタスクを、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区分している。本稿では、この内、リンク予測とノードのランキングの手法を用いることにより、時間と共に構造が変化するネットワークにおいて、現在は重要でないが将来的に重要となるようなノード（有望ノード）を同定する手法について述べる。

ここでのリンク予測問題は、現在のネットワーク構造から将来のネットワーク構造を予測する問題である。そのため、対象とするデータは、時間が経過する中でネットワークを構成する任意の2ノード間に新たなつながりが生じる（あるいは、消滅する）データを用いる。また、そのような時間と共に構造が変化するネットワークを動的ネットワークとし、その動的ネットワークのある時点のスナップショットを静的ネットワークと呼ぶことにする。

この動的ネットワークの例には、論文の共著関係や E-mail の送受信関係などが挙げられる。両者のネットワークの性質は、共著の関係は時間と共に増え続け、消滅することはないが、E-mail の送受信関係は生じたり消滅したりする。ある時点（区間）における共著関係や E-mail の送受信を表すネットワークは静的ネットワークといえる。この他に、論文の引用関係をリンクとする論文の引用関係ネットワークや、SNS における友人関係をリンクとする友人関係ネットワークなど、時間と共に変化するネットワークは複数考えられる。これらの例において、本研究の目的である有望ノードの同定は、前者においては現在の被引用数は少ないが将来に被引用数が飛躍的に増える論文を探し出すことを、後者においては現在は注目されていないが将来的に人気を集める人物を発見することを意味する。

従来、主に静的な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 [4]。一方、動的ネットワークを対象として、中心的・影響を与える存在の予測に着目した研究はあまり行われていない。その中で、ネットワークのノード情報と構造情報に基づく特徴を、教師あり学習アルゴリズムに渡してリンク予測を行い、その後ノードのランキングを行う手法が O'Madhain らによって提案されている [7]。本稿では、彼らが提案した手法をもとに、リンク予測の精度と有望ノードの順位付けとの関係を明らかにする。

2. 有望ノードの同定

2.1 ネットワークの中心性

本節では、改めて有望ノードの同定について定義し、次に有望ノードを同定する方法について述べる。そのために、まず有望ノードを定義するために用いるネットワーク中心性について紹介する。

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内において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ための指標である。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ってくる。そのため、今回は複数の中心性について予測を試みる。以下に、本研究で使用する中心性を示す。これらはいずれもネットワーク解析において代表的な中心性である。

- 次数中心性 [4]
ネットワーク内のノードとどの程度つながっているかを表し、自身に隣接するノード数で定義される。
- 近接中心性 [4]
ネットワーク内での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。
- 媒介中心性 [4]
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。
- PageRank [3]
ネットワーク中においてノード間を遷移する状態が繰り返し起こる場合に、ネットワーク中を遷移する者がどの

連絡先: 宮西大樹, 神戸大学大学院工学研究科情報知能学専攻, 兵庫県神戸市灘区六甲台町 1-1, Tel: 078-803-6220, E-mail: miyanishi@ai.cs.kobe-u.ac.jp

程度ノードを訪れたかを表す指標。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。

2.2 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、2ノード間にリンクがある・なしを予測することである。このように、ネットワークの観測された部分から未だ観測されていない部分を予測する問題をリンク予測問題といい、このリンク予測をいかにして行なうかが重要である。良く用いられる方法としては、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二種類がある。ノードの持つ情報から予測する方法では、ノードの特徴を用いる。その際、各ノードの特徴を見るのではなく、ノードペアの特徴ベクトルを用いる。また、ネットワークの構造から予測する方法では、ノードペアに対してノードの周辺のリンク構造（リンク指標）に基づく特徴を用いる。両手法ともノードペアの特徴を用いて、従来の機械学習のアプローチを利用することによりリンクのある・なし（または強度）を予測する。代表的なリンク指標を以下に示す。

- 共通隣接ノード指標
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つのノードの間にはリンクが現れやすい
- Jaccard 係数
共通の隣接ノードの数が、2つのノードの隣接ノード集合に占める割合の大きいノードペアほどリンクが現れやすい
- Adamic/Adar [1]
重み付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もの
- $Katz_{\beta}$ 指標 [6]
2つのノードを結ぶパスの集合の大きさの重み付き和であり、共通隣接ノード指標の一般形
- 優先的選択指標 [2]
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

ネットワークの構造からリンク予測を行う場合でも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、ノードの持つ情報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。また、リンク予測についてネットワークごとに有効なリンク指標は異なるので、上記すべてのリンク指標をノードペアの特徴として用いる。リンク予測は、ノードペアの過去の情報を用いてロジスティック回帰を使用することで行う。そして、出力したリンクのある・なしの確率を大きい順に並べて、適当な閾値以上のペアの間にリンクがあるかどうかを予測する。ちなみに、出力されたリンクのある確率が高ければ高いほど実在するネットワークにおいても早期にリンクができることと考えることにする。

2.3 評価手法

本手法の有効性を示すために、新たに有望ノードという概念を定義する。有望ノードとは、動的ネットワークを一定時間間隔ごとに区切り、その中から、ある2つの時点を選定し、

各時点の静的ネットワークにおいて、ネットワークの中心性をもとに各ノードの順位付けを行い、両時点における順位の開きが大きく、新しい時点では高順位であり、古い時点においては低順位であるノードのことを言う。そして、上記の要件を満たすように、ノードの有望さを定義すると、有望さはノードの古い順位を新しい順位で割った値となる。この定義により、古い時点のネットワーク中での影響度は小さく、新しい時点での影響度が増すノードを高く評価（有望である）し、またその逆を低く評価する（有望でない）ことができる。

有望なノードを同定できるかどうかの評価は、高エネルギー物理学の論文データベースもとに共著関係ネットワークを構築し、そこから有望となるノード（ここでは、人間）を推定することで行う。また、将来の関係の予測精度と有望ノード同定との関係を明らかにする実験を現在行っている。

3. おわりに

本稿では、論文の共著関係をもとに、現在は重要でないが将来的に重要となるようなノードを同定する方法について述べた。実験結果の詳細については、発表時に口頭で報告する予定である。

参考文献

- [1] Lada A. Adamic and Eytan Adar. Friends and neighbors on the web. *Social Networks*, Vol. 25, No. 3, pp. 211–230, jul 2003.
- [2] A. L. Barabási, H. Jeonga, Z. Nédá, 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.
- [3] 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.
- [4] Linton C. Freeman. Centrality in social networks conceptual clarification. *Social Networks*, Vol. 1, pp. 215–239, 1979.
- [5] Lise Getoor and Christopher P. Diehl. Link mining: a survey. *ACM SIGKDD Explorations Newsletter*, Vol. 7, No. 2, pp. 3–12, dec 2005.
- [6] Leo Katz. A new status index derived from sociometric analysis. *Psychometrika*, Vol. 18, No. 1, pp. 39–43, mar 1953.
- [7] 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.