The Link Prediction Problem for Social Network (2003)

David Liben-Nowell and Jon Kleinberg
Cornell Univ. (USA)

the Twelfth Annual ACM International Conference on Information and Knowledge Management (CIKM'03)
http://www.cs.cornell.edu/home/kleinber/link-pred.pdf

1 本論文の目的

本論文ではソーシャルネットワーク中の欠けているリンクを推定する問題（Link Prediction Problem）を新たに定義する。これは、成長するネットワーク（例えば共著ネットワーク）のある時点でのスナップショットがある場合に、ノード間に新たにリンク（例えば新しい共著関係）が追加される可能性の高さを求める問題といえる。本論文ではこの問題を定義するとともに、ネットワーク構造そのものが新しいインタラクション（ネットワークに追加される新しいリンク）を推定するための情報を持っていること、そしてネットワーク構造を利用した推定指標の性能を、実際の論文の共著ネットワークを用いた実験を通して示す。

2 背景

近年、成長するネットワークに関する研究が多くなされている。その多くは現実にある様々なネットワークの成長モデルを作ることに取り組んでいる。

ネットワークはリンクの追加によって成長し変化するが、ネットワーク上においてリンクの追加がどのように起こるかについてはまだよくわかっていない。そこで本論文では追加されるリンクを推定する Link Prediction Problem を提案する。

3 手法・アルゴリズム

リンク追加にとって本質的な要素とは何かを考える必要が Link Prediction Problem にはある。共著関係を例に取ると、転職によって突然同じ職場に配属された二人が共著関係になるということを推定するのは不可能に近い。しかし共著ネットワークにおいて近いと感じられる二人、例えば共通の共著者を多く持っている二人は、この先共著になる可能性は高いだろうと多くの人は考えるだろう。本論文では、この直感的な測り方の確からしさを、複数の手法を比較することで評価する。

比較する手法は以下の通り。大きく分けると (1) 隣接ノードに基づく指標（Common neighbors、Jaccard 係数、Preferential attachment）、(2) ネットワーク全体に基づく指標（Katz、Hitting time、PageRank、SimRank）、(3) よりハイレベルなアプローチ（Low-rank、Unseen bigrams、Clustering）に分けられる。

- graph distance : 2 ノード間の最短パスの長さ
- common neighbors : 2 ノード間で共通する隣接ノードの数
- Jaccard's coefficient : $\frac{\Gamma(x)\cap\Gamma(y)}{\Gamma(x)\cup\Gamma(y)}$
- Adamic/Adar : 2 ノード間で共通する隣接ノードの隣接ノード数の逆数 （共通する知り合いが、知り合いの数の少ない人であるほどスコアが高い）
- preferential attachment : $|\Gamma(x)||\Gamma(y)|$ （x と y、それぞれが知り合いの多い人であるほどスコアが高い）
- Katzβ : 2 ノード間をつなぐパスの数の和
- hitting time, rooted PageRank$_\alpha$, SimRank$_\gamma$: いずれもランダムウォークをベースとした指標

推定対象とするネットワークには、5 つの分野（astro-ph:宇宙物理、cond-mat:物性、gr-pc:量子宇宙論、hep-ph:高エネルギー物理現象学、hep-th:高エネルギー物理理論）から取り出した共著関係を用いる。全データは physics e-Print arXiv [1]から取得した。トレーニングデータは 1994〜1995 年の期間に、テストデータは 1997〜1999 年の期間に出版された論文である。なお、性能の評価にはトレーニングデータ、テストデータそれぞれに少なくとも 3 本の論文が含まれている著者のみを対象とした。

4 評価

テストデータのうちコアノード（トレーニングデータ、テストデータにそれぞれ 3 本以上の論文があるノード）間の新しい（トレーニングデータに含まれない）エッジを推定できたときに、その推定が正解したとみなす。

推定指標は全ノードのペアのうち新たにエッジを追加できる（トレーニングデータ中にエッジのない）すべてのペアに対してスコアを付ける。そしてスコアを付けたペアのうち、コアノードによって構成されるペアの上位 N 件を推定結果とする（N は正解データの数、つまりテストデータのうちコアノードのペアで構成された数)。この推定結果のうち何%が正解であったかを評価指標とする。

分析を行うにあたっては、ランダムに推定した場合、raph distance、common neighbors の三つを基準として、これらに対する性能比較により行う。

[1] http://www.arxiv.org

5 知見

全ての指標およびデータセットにおいて、ランダムを上回る性能が出た。今回用いた指標はいずれもネットワークのトポロジー情報のみを用いたものである。つまり、ネットワークトポロジーからリンク推定が可能であることが示された。しかし、もっとも高い値が出た組み合わせでも正解する確率は16%程度であり、性能向上のための改善が必要である。

今回用いた指標には、(1) 隣接ノードに基づく指標（Common neighbors、Jaccard 係数、Preferential attachment）、(2) ネットワーク全体に基づく指標（Katz、Hitting time、PageRank、SimRank）、(3) よりハイレベルなアプローチ（Low-rank、Unseen bigrams、Clustering）の大きく三つに分けられる。(2) および (3) が比較的推定性能が高く、Adamic/Adar や Katz が良い結果を出した。

指標にはペアのスコアを求める際に共通する隣接ノードが必要なものがいくつかある。しかしデータを調べると、新たに生まれたエッジの 70〜80%は共通する隣接ノードを持たない（グラフ上にて 3 ホップ以上の関係にある）ペアのものであった。3 ホップ以上に対象を絞った場合を調べたところ、やはり全てランダム以上（3〜8 倍）の性能が出た。一番良かったのは unweighted Katz+unseen bigram であった。

（文責：濱崎）