

Web 上からソーシャルネットワークを抽出する汎用モデル*

金英子†
東京大学大学院松尾豊‡
産業技術総合研究所石塚満§
東京大学大学院

1 Abstract

人や組織のコミュニティの分析やオントロジーの抽出などにおいて、ソーシャルネットワークは重要な役割を果たしている。近年は、Web 上の情報から人や組織などのソーシャルネットワークを抽出して分析しようとする研究が盛んに行われている。これまでは、分析や応用のニーズに応じて、様々な種類のエンティティやネットワークの抽出手法が検討されてきたが、それぞれ個別の手法になっている。本稿では、ソーシャルネットワークを構築する汎用モデルを提案して、指標とパラメータを定義することで、様々な種類のソーシャルネットワークを構築できることを示す。そして、これまで行われた事例に適用して、各システムの抽出プロセスを説明する。

2 関係抽出の汎用モデル

ソーシャルネットワークは、目的の関係構造を適切に反映することが重要である。本研究で、ネットワークの抽出モデル（図 1）を、下記の数式に一般化する。

$$f(\mathbb{S}_r(X, Y), \Theta) \rightarrow \{0, 1\} \quad (1)$$

ここで、 $\mathbb{S}_r(X, Y)$ は m 次元のベクトル空間 $(S_r^{(1)}(X, Y), S_r^{(2)}(X, Y), \dots, S_r^{(m)}(X, Y))$ であり、エンティティ X と Y の関係 r における関係の強さのさまざまな指標を表す。例えば、 $S_r^{(1)}(X, Y)$ は、 $Cooc(X, Y) (= n_{X \cap Y}, \text{共起頻度})$ や、 $Jaccard(X, Y) (= n_{X \cap Y} / n_{X \cup Y}, \text{Jaccard 係数})$ 、或いは $Overlap(X, Y) (= n_{X \cap Y} / \min(n_X, n_Y), \text{Overlap 係数})$ を表すことができる。さらに、複数の指標を組み合わせたスコア関数であることも、関係を記述する文脈 D のスコアであることもある。パラメータ Θ は、 n 次元のベクトル空間 $(\theta^{(1)}, \theta^{(2)}, \dots, \theta^{(n)})$ であり、それぞれの関係の指標における閾値関数を示す。例えば、 Θ は、 T_{ov}, T_{co}, M_1 と M_2 の組み合わせで不均一なコミュニティにおける関係抽出を行う。

関数 f は、これらの複数の指標とパラメータに基づいて、ネットワーク上に X と Y の間にエッジでつながるか

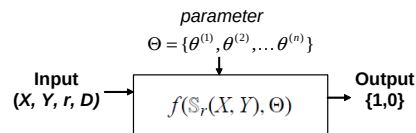


図 1: general extraction model of social networks

どうかを決める 2 値関数 $\{1, 0\}$ である。 f 関数の内部構造は、エンティティの種類や紐帯を示す関係、コミュニティの特徴により異なる。そのため、オフラインジョールが必要になり、訓練データを基に、適切な指標とパラメータの値を定めておく必要がある。そして、オンライン状態では、入力されたエンティティに対してネットワークを抽出する。ここでは、いかに効率的に Web 上からソーシャルネットワークを抽出することが重要になる。基本的には、**検索クエリの生成**、**Google 検索**、**ネットワークの構築**という 3 つのプロセスにより Web 上からソーシャルネットワークを抽出する。まず、入力されたエンティティを用いて検索クエリ $X \text{ AND } Y$ を生成する。特定の関係に絞るときには、その関係の関係語をクエリに加える。そして、生成されたクエリを検索エンジン（例えば、Google）に入力する。検索結果からヒット件数と上位の Web ページ（或いは、スニペット） D 取得する。最後に、定めた指標とパラメータにより関数 f の値を求めて（1 の場合エッジでつながり）ネットワークを構築する。

3 実例

3.1 研究者ネットワークの抽出

検索エンジンを用いて Web 上からソーシャルネットワークを抽出する既存の研究では、研究者を対象とすることが多い [2, 3]。研究者ネットワークでは、式 2 のように、関係の強さの指標（例えば、Overlap 係数 [2]。Mika らの研究では、Jaccard 係数にしている [3].）を決めて、その閾値をパラメータとする。

$$\begin{cases} \mathbb{S}_r(X, Y) = \text{Overlap}(X, Y) \\ \Theta = T_{ov} \end{cases} \quad (2)$$

システムの流れとして、まず、オフライン状態で、訓練データから T_{ov} の値を学習する。そして、オンライ

*General-purpose Model for Social Network Extraction from Web

†YingZi Jin, The University of Tokyo

‡Yutaka Matsuo, AIST

§Mitsuru Ishizuka, The University of Tokyo

ン状態で、 X AND Y のクエリを生成し、Google のヒット件数を求めて、 f 関数によりネットワークを構築する。 f 関数は、 $Overlap(X, Y) > T_{ov}$ の場合 1 になり、 X と Y の間に関係があるとしてエッジでつなぐ。

3.2 アーティストのネットワーク抽出

均一なコミュニティに対しては、一貫した閾値により関係の有無を判断することができるが、不均一なコミュニティに対しては一貫した閾値を決めることができない。ここで、不均一というのは、関係の出力が不安定であることを指す。例えば、異なる国や分野の人の関係は、同じ国や分野より Web 上に名前の共起が少ないので、関係が低く計算される。

我々は、ネットワーククエスチョンの考えに基づいて、それぞれの人から見た関係の強さも考慮する。ネットワークの抽出は簡単で、パラメータ Θ は、全体からみた関係の強さの閾値 T_{ov} とそれぞれの人からみる強い関係の人の数の閾値 M を両方用いる。

$$\begin{cases} \mathbb{S}_r(X, Y) = Overlap(X, Y) \\ \Theta = (T_{ov}, M) \end{cases} \quad (3)$$

ネットワークの抽出は、まず、オフライン状態で訓練データから各パラメータの値を学習する（例えば、訓練データを F 尺度最大になるように分類できるパラメータの値を取得する）。そしてオンライン状態で、名前のペア X AND Y で検索クエリを生成し、Google によりそれぞれのヒット件数を求める。そして、ネットワークを構築する段階では提案の手法（閾値調整）を用いる。まず、研究者ネットワークと同様に、絶対的な手法によりネットワークを構成する、すなわち、 $Overlap$ 係数が閾値 T_{ov} 以上のエンティティペアに対して $f = 1$ としネットワーク上でつなぐ。次に、相対的な手法により、エッジの数が M 以下のノードに対して、 $Overlap$ 係数が閾値を満たさなくてもエッジの数が M になるまで、そのノードとの関係の強い順につないでいく。実際のシステム (<http://release.nikkei.co.jp/>) [1] の中では、複数の指標 ($\mathbb{S}_r(X, Y) = (Overlap(X, Y), Cooc(X, Y))$) を用いて、各ノードの閾値を徐々に下げていく方法で、よりロバストに関係を抽出している。

3.3 企業間関係のネットワーク抽出

Web 上のメディア効果などにより、名前が頻出するエンティティ（例えば、企業や組織名）に対して、ヒット件数は関係の強さを適切に表すことができない。Web 上からこれらのエンティティ同士の関係を抽出するために、我々は X と Y について記述している文脈を集めて、その文脈から関係を表すスコアを計算して、ネッ

トワークを抽出している [4]。文脈のスコアは式 4 のように、その文脈に含まれる語（関係語と呼ぶ）のスコアを足し合わせて計算することができる。

$$\begin{cases} \mathbb{S}_r(X, Y) = \sum_{w \in (X, Y)} W^{(w)}(X, Y) \\ \Theta = score_{thre} \end{cases} \quad (4)$$

ネットワークの抽出は、オフライン状態で、関係ごとのパラメータの値 $score_{thre}$ と関係語 W のリストを取得しておく。関係語のスコアは、その関係の代表語（例えば、提携関係の場合は「提携」）との Jaccard 係数を用いて計算する。そして、オンライン状態で、関係語の上位 n_q 個を企業の名前のペア X と Y に加えることで検索クエリを生成する。そして検索クエリを検索エンジンに入力し、上位にヒットしたページをダウンロードして関係のページ集合を得る。最後に、 X と Y の周りの文脈から、2つの企業の関係の有無を判断する。上記の式 4 により、関係は、2つの企業の名前と関係語を含む文に対して、文に含まれる関係語のスコアを足し合わせたスコアを用いて閾値 $score_{thre}$ と比較して、 f 関数の値を決める。

4 まとめ

本稿では、Web 上からソーシャルネットワークを抽出する汎用モデルを提案し、3つの実例を通してその適応性を示すとともに、それぞれの具体的な指標、パラメータと抽出プロセスを説明した。今後は、より多くのエンティティのソーシャルネットワークを Web 上から抽出してみて、それぞれの抽出手法を検討してまとめていきたい。

参考文献

- [1] Y.Z. Jin, Y. Matsuo, and M. Ishizuka. Extracting a social network among entities by web mining. In *ISWC06 Workshop on Web Content Mining with Human Language Technologies*, 2006.
- [2] Y. Matsuo, J. Mori, M. Hamasaki, H. Takeda, T. Nishimura, K. Hasida, and M. Ishizuka. POLYPHONET: An advanced social network extraction system. In *Proc. WWW2006*, 2006.
- [3] P. Mika. Flink: Semantic web technology for the extraction and analysis of social networks. *Journal of Web Semantics*, Vol. 3, No. 2, 2005.
- [4] 金英子, 松尾豊, 石塚満. Web 上の情報を用いた企業間関係の抽出. *人工知能学会論文誌*, Vol. 22, No. 1, pp. 48–57, 2007.