3H1-1

Webから抽出した社会ネットワークに基づくエンティティランキングの学習

Ranking Entities on the Web using Social Network Mining and Ranking Learning

金英子[*1] 松尾豊[*2] 石塚満[*1]
Yingzi Jin Yutaka Matsuo Mitsuru Ishizuka

[*1]東京大学大学院 情報理工学系研究科
Graduate School of Information Science and Technology, The University of Tokyo

[*2]東京大学大学院 工学系研究科
School of Engineering, The University of Tokyo

Social networks have garnered much attention recently. Several studies have been undertaken to extract social networks among people, companies, and so on automatically from the web. In social sciences, social networks are used to analyze the performances and values of companies. This paper describes an attempt to learn ranking of named entities from a social network that has been mined from the web. For example, if we seek to rank companies, we can extract the social network of the company from the web and discern and subsequently learn a ranking model based on the social network. Consequently, we can predict the ranking of a new company by mining the relations to other companies. Using our algorithm, we first construct social networks using several relevance measures in addition to text analysis. Subsequently, the relations are integrated to maximize the ranking predictability. We also integrate several relations into a multiple relational network and use the latest ranking learning algorithm to obtain the ranking model. In addition, the ranking scores on each network as relational feature combined with entity. We conducted two experiments on a social network among corporations to learn the ranking of market prices, and on a social network among researchers for ranking of researchers' productivity.

1. はじめに

1930 年代から社会計量学の分野で始まった社会ネットワーク分析は，近年，Web に見られるような電子的な流通が増えるにつれて，研究者の注目を集めるようになった．Web 上には多様なエンティティが無秩序かつ大量に存在し，それらのエンティティ同士を結ぶ様々な関係を自動的に見つけ出し，社会ネットワークを構築して分析する研究が多く行われている [Kautz 97, Mika 05, Matsuo 06, Jin 07]．Staab の “Social Networks Applied” で述べているように，社会ネットワークは，コミュニティの発見やクチコミ・マーケティングの分析など，経済・経営分野での応用が期待されている [Staab 05].

社会ネットワーク分析では，行為者のもつ “影響力” を分析するために，ネットワーク中心性 [Scott 00] という指標がしばしば用いられている．例えば，より多くの人と関係をもつほど影響力が大きいという考えに基づく次数中心性や，他の人となるべく近い距離で繋がっているほど影響力が大きくなるという近接中心性，任意の２者間をつなぐ媒介的な位置にいるほど重要であるという媒介中心性などがある．一方，リンクマイニング分野においても，ネットワークにおけるノードをランキングすることは重要な課題であり [Getoor 05]，その目的は，与えたネットワークの関係構造に基づいてノードのランキングを学習することである [Agarwal 06, Chang 00].

Web からの社会ネットワークの抽出，およびネットワークのランキング学習という２つの研究に関連して，本研究では，Web から抽出された社会ネットワークに基づいて，エンティティのランキングを学習する試みを行う．エンティティのランキングは，それぞれの目的により異なる．これらの予測目的のランキングに影響するものは，一体なんだろうか．本研究では，従来の属性に基づく分析から離れて，関係が予測目的に影響を与えるという観点で，社会ネットワークによる予測目的のランキングへの影響の具合を調べる．

本研究で提案するランキング学習モデルでは，任意のエンティティのリストを与えた場合，まず，既存の手法により Web 上の情報からエンティティ同士の様々な関係を抽出する [Matsuo 06, Jin 07]．そして，中心性とランキングアルゴリズムを利用してネットワーク中のノードをランキングし，正解となるランキングを学習する．ここで，ネットワークから生成されたランキングを「内部ランキング」と呼ぶことにする．そして，外部から与えた予測目的のランキングを「外部ランキング」と呼ぶ．本研究では，内部ランキングを手がかりに外部ランキングを学習する３つの手法を提案する．

評価実験では，電気系企業 312 社のネットワークから企業の価値を予測することと，253 人の研究者のネットワークから研究者のパワーを予測する試みを行う．本論文では，まず２章でランキング学習モデルを概観し，３章で既存の Web 上から社会ネットワークを抽出する手法をまとめる．そして，４章では，提案のネットワークに基づいたランキング学習モデルを説明し，5 章で評価実験と結果を述べ，6 章で議論と結果を述べる．

2. ランキング学習モデル

本研究では,「無数に存在する関係の中で，予測したい対象があって始めて関係が見える」という仮説に基づいている．すなわち，関係と予測対象について考えると，２つのエンティティ，A と B の間には，数限りない関係が成立し得る．例えば，２人の研究者の場合は,「同じ研究室に所属している」「同じ研究分野を専攻としている」「共同研究をしている」「共著関係である」などが考えられる．もう少し関係の範囲を広げると，例えば,「同じアジア人である」「同じ人間である」とも言えるし,「名前が漢字３文字」なども関係と認められる．このような無数にある関係の中から，研究分野のコミュニティを抽出するた

連絡先: 金英子，東京大学大学院 情報理工学系研究科
〒 113-8656 文京区本郷 7-3-1 工学部新 2 号館 111C1 室
TEL: 03-5841-6774, FAX: 03-5841-8570
Email: eiko-kin@mi.ci.i.u-tokyo.ac.jp

めに，なぜ「共同研究」とか「共著」という関係が重要に思えるのか？それは，関係がそれ単独で世界に存在するものではなく，予測したい対象があって初めて着目すべき関係が明らかになるからである．現実世界において，予測したい対象は，人気のある日記，商品の売り上げ，企業の価値，有名人の知名度，論文の重要度など様々ある．このような対象を予測するのに寄与する関係はいったい何か？どのようなネットワークの内部ランキングが予測目的を最も反映するか？これらは，ネットワークの学習が様々なエンティティに適用できるようにするために必要な議論である．本研究では，予測対象として外部のランキングデータを学習データとして用い，ネットワークを自動的に学習して分析することにより，上記の問題を解明する．

提案手法は，２つのステップがある．

ステップ 1: 社会ネットワークの抽出 任意のエンティティのリストを与えると，Web 上から社会ネットワークを抽出する既存の手法により，様々なネットワークを構成する．

ステップ 2: ランキングの学習 予測対象のランキングを与えると，ステップ 1 で得られる社会ネットワークに基づいて，そのランキングを学習する．

従って，得られるモデルにより，任意のエンティティのランキングをその関係のネットワークから予測することができる．

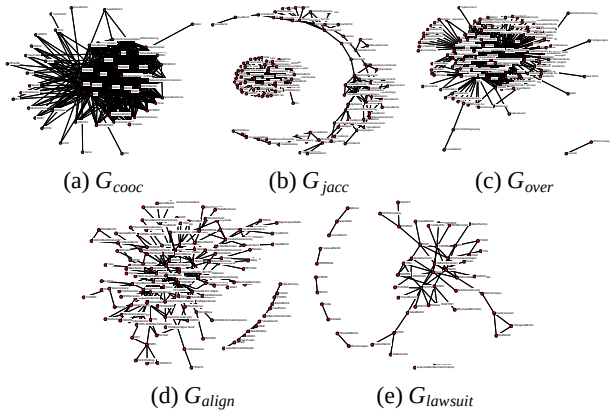
3. Web 上から社会ネットワークの抽出

この段階では，エンティティの名前のリストを与えれると，検索エンジンを用いて、Web 上からさまざまな関係の社会ネットワーク $G_i(V, E_i), i = 1, \ldots, m$ を抽出する．ここで，m は関係の種類を表し，V はエンティティの集合，E_i は i 番目の関係を表すエッジの集合を示す．ここでは，無向グラフを対象とする．

まず，検索エンジンを用いて，Web 全体における名前の共起情報を調べ，Web 共起のネットワークを構築する [Kautz 97, Mika 05, Matsuo 06]．すなわち，Web における x と y の関係を，“x AND y” というクエリを検索エンジンになげることで得られる共起情報を用いて計算する．そして，関係の強さの閾値を設定することで，x と y の共起指標がそれ以上であればエッジを張るというアプローチをとる．本研究では，よく使われる３つ指標により Web 共起のネットワークを構築する：「共起のネットワーク」(G_{cooc})，「Jaccard ネットワーク」(G_{jacc}) と「Overlap ネットワーク」(G_{over}). 共起の尺度は，それぞれ名前の共起頻度 $n_{x\wedge y}$, Jaccard 係数 $n_{x\wedge y}/n_{x\vee y}$ ([Mika 05, Kautz 97] で用いた尺度), および Overlap 係数 $n_{x\wedge y}/\min(n_x, n_y)$ ([Matsuo 06] で用いた尺度) を用いる．

Web 共起のネットワークに対し，松尾らは関係を表す属性を用いて，学習データから関係を判別するルールを生成することで,「共著」「同研究室」「同プロジェクト」「同発表」など関係のラベルを付与する [Matsuo 06]．同じアプローチで本研究では，研究者の Jaccard ネットワークに対して,「同プロジェクト」「同所属」の関係のラベルを付与し，それぞれのラベルのネットワークを「同所属関係のネットワーク」 (G_{affi}) と「同プロジェクト関係のネットワーク」(G_{proj}) とする．

金らは，関係の識別という手法を提案し，Web 上に名前が頻出するエンティティに対して，社会ネットワークを抽出する手法を提案している [Jin 07]．まず，関係語を検索クエリに加えることで，Web 上からエンティティ x と y のある関係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，ネットワークを抽出している．文脈のスコアは，その文脈に含まれる関係語のスコアを足し合わせて計算する．



(a) G_{cooc} (b) G_{jacc} (c) G_{over}

(d) G_{align} (e) $G_{lawsuit}$

図 1: Web から抽出した企業関係ネットワーク．

本研究では,「提携関係のネットワーク」 (G_{align}) と「訴訟関係のネットワーク」 ($G_{lawsuit}$) の２つの種類の企業間関係のネットワークを構築する．

抽出された日本の電気系企業 312 社のネットワークを図 1 に示す．同じエンティティの集合であっても，関係の種類や関連度の尺度が異なることにより，異なる社会ネットワークが得られることが分かる．

4. ランキングの学習

この段階では，Web から抽出した社会ネットワークと与えた予測目的のランキングに基づいて，エンティティのランキングを学習する．ここで，ネットワークから直接生成されるランキングを「内部ランキング」とする．例えば，i 番目のネットワークの内部ランキングを $R^{(G_i)}$ で表す．そして，予測目的のランキングを外部ランキングとし $\hat{R}$ で表す．本研究では，社会ネットワークに基づく３つのランキング学習方法を提案する．

4.1 手法 1: 単純比較法

ここでは単純に，各ネットワークのランキングを企業価値のランキングとそれぞれ比較し，相関が高いときの関係の種類とランキングの指標を調べる．この手法は単純であるが，複数の関係からネットワーク分析を行う最初の段階でもある．すなわち，ここでは，それぞれのネットワークをランキングし，その内部ランキングと外部ランキングとの相関を比較する．直感的に，もし内部ランキング $R^{(G_i)}$ が最も高い相関を表す場合，関係 i が最適なパラメータに入り，エンティティのランキングを予測するためのもっとも適切な指標であると言える．

$$\theta = \underset{i\in m}{\operatorname{argmax}}\, Cor(R^{(G_i)}, \hat{R}). \tag{1}$$

ネットワークのノードをランキング手法は様々である：今回は，社会ネットワークの中心性のランキングとして，あるノードが持っている関係の数を計算する次数中心性 (R_D) と，ネットワーク中の特定のノードから他のノード同士の関係をどれぐらい媒介しているかを表す媒介中心性 (R_B)，および，ネットワーク中の特定のノードから他のノードに対してどれぐらい近い位置にいるかを表す近接中心性 (R_C) を用いる．ほかに，マルコフ遷移ネットワークにおける有名なランキング方法である *PageRank* (R_P) を用いる．これらのランキング手法の異なる性質により，単純比較法を $\{i, j\} \in \theta$ ($i \in m$, $j \in n$) の最適パラメータを見つける方法に拡張することができる．すなわち，対象とするランキング $\hat{R}$ と相関が最も高い場合の i 番目ネットワークの j 番目の内部ランキング $R_j^{(G_i)}$ を見つけることになる．

$$\theta = \underset{i\in m, j\in n}{\operatorname{argmax}}\, Cor(R_j^{(G_i)}, \hat{R}) \tag{2}$$

4.2 手法 2: 関係の組合わせ法

多くのネットワークに基づくランキング手法では，対象とするネットワークを単一の関係で構成されていると仮定している．しかし，現実世界には様々な関係が存在していて，これらの関係が重なって何らかの結果を生み出している．関係の組合わせ法では，複数の関係を重み付き線形和で組合わせることで，得られるネットワークにおけるランキングを学習する．つまり，企業の価値はそれぞれの関係によって独自に決められるのではなく，複数の関係が同時に影響するという考えで影響の具合を見つける方法である．

まず，各関係の社会ネットワークをそれぞれの重み w_i $(i \in m)$ の線形和で組合わせる（但し，$\sum w_i = 1$）ことで，1つの組み合わせネットワークを得る．そして，中心性とランキング手法により組合わせネットワークのランキングを計算し，外部ランキングとの相関をフィードバックすることで，最適な組み合わせの重み θ を得る．

$$\theta = \underset{w_i, j}{\operatorname{argmax}}\, Cor(R_j^{\sum_{i \in m} w_i G_i}, \hat{R}) \tag{3}$$

ほかに，NetRank の考えに基づいて，関係の組合せ法を実装する試みも行う．Agarwal らにより提案されている NetRank 法は，マルコフ遷移における遷移確率を関係ごとに設定することで，最適な関係の遷移確率を求める [Agarwal 06]．ここで，各関係は正の遷移確率 $\beta(i) > 0$ をもち，この場合の遷移行列 A は下記のようになる．

$$A(y,x) = \begin{cases} \alpha \frac{[\beta(i(x,y)) \in E]}{OutWeight(x)} + (1-\alpha) r_y, & e \in V_0 \\ r_y, & \text{otherwise} \end{cases} \tag{4}$$

ここで，$\mathrm{OutWeight}(x) = \sum_y \beta(i(x,y))$ で，A は重み β の関数である，遷移確率 p は $p = Ap$ を満たして，対象のランキングを最も反映する $\{\beta_i\}$ を見つける問題になる．我々は，それぞれの関係の重みをランダムにサンプリングして得られるパラメータを用いて，対象ランキングとの相関をフィードバックすることで，最適なパラメータを求める．

4.3 手法 3: ランキングの組合わせ法

ランキング組合わせ法では，ネットワークを組合わせずに，それぞれのエンティティの各ネットワークにおけるランキングをその企業の特徴量とし，これらの特徴量を組合わせることで，エンティティのランキングを予測する．この場合，エンティティのネットワークにおける関係的・構造的特徴が，エンティティ自身の属性として働く．従って，この手法の目的は，特徴量の最適な組合わせのパラメータ θ（すなわち，各特徴量の重み）を見つけることである．

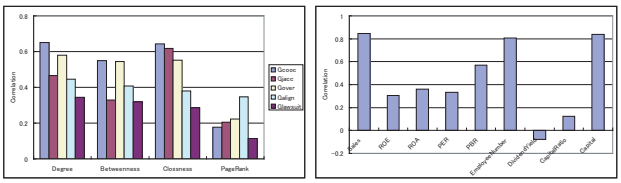
$$\theta = \underset{w_{i,j}}{\operatorname{argmax}}\, Cor(w_{i,j} \cdot R_j^{(G_i)}, \hat{R}). \tag{5}$$

ここで，$R_j^{(G_i)}$ は，i 番目ネットワークの j 番目の内部ランキングを表し，それぞれのエンティティは，$m * n$ 次元の特徴量のベクトルで表す．今回は，SVM-regression 学習機を用いて，最適なパラメータ $w_{i,j}$, $(i \in m, j \in n)$ を見つけて，ランキングを予測する．

5. 評価実験

本章では，Web から抽出[*1]した社会ネットワークに基づいて，エンティティのランキングを予測する試みを行う．

*1 検索エンジンは MSN（http://jp.msn.com/）を用いる．



(a) network rankings (b) fundamental rankings

図 2: 各ネットワークのランキングと企業価値との相関.

まず，日本の電気系企業 312 社を対象に，Web 上から 5 つの社会ネットワークを抽出する．[松尾 05] らの手法により，共起のネットワーク（G_{cooc}），Jaccard ネットワーク（$G_{jaccard}$），Overlap ネットワーク（$G_{overlap}$）を構築し，[金 07] の手法を用いて，提携関係のネットワーク（G_{align}）と訴訟関係のネットワーク（$G_{lawsuit}$）を構築する．つぎに，253 人の研究者を対象に，Web 上から 5 つの社会ネットワークを抽出する．同じく [松尾 05] らの手法により，G_{cooc}，$G_{jaccard}$，$G_{overlap}$ を構築し，そのうちの Jaccard ネットワークに対してさらに C4.5 で獲得したルールにより，同所属関係のネットワーク（G_{affi}）と同プロジェクト関係のネットワーク（G_{proj}）を構築する．そして，これらのネットワークに基づいて，企業と研究者のランキングを学習し予測を行う．

本実験では，3 交差検定を行う．そして，予測段階で得られるランキングと正解のランキングとの相関の平均を結果として示す．

5.1 企業間関係のネットワークに基づく企業のランキングの予測

企業の価値を表す指標として今回は時価総額[*2]を用いる．

まず，Web から抽出された 5 つの企業間関係のネットワークを様々な指標でランキングすることで，内部ランキングを得る．社会ネットワーク分析で良く使われている次数中心性 (R_D)，媒介中心性 (R_B)，近接中心性 (R_C) のランキングと，グラフランキングの代表的なアルゴリズム PageRank(R_P) を用いる．そして，単純比較法により企業価値を予測する．図 2(a) に，5 種類のネットワークを 5 つの方法でランキングした内部ランキングと企業の時価総額のランキングとの相関を示す．Web 共起や提携関係が訴訟関係より高い相関を表すことが分かる．また，同じ関係においても，次数中心性と近接中心性が時価総額と相関が高い．図 2(b) は，各ファンダメンタル指標[*3]におけるランキングと企業価値のランキングとの相関を示す．

次に，関係の組合わせ法では，まず，各関係のネットワークをそれぞれの重みの線形和（但し，$\sum w_i = 1$）でコンバインしたネットワークに対してランキングした結果と外部ランキングとの相関を調べる．表 1 では，Web 上における名前の共起関係が企業の時価総額のランキングに対する影響が大きいことが分かる．そして，近接中心性が高い相関を示している．すなわち，他の企業と Web 上で近い距離にいるほど，企業のランキングを向上させることができる．次に，関係の組合せ法を，NetRank の手法を用いて実験する．ネットワークをの遷移確率をそれぞれ関係の種類ごとに設定し，企業の価値のランキングを選好順位として与えることで，関係の組合わせの重みを学習

*2 時価総額は，ある企業の株価に発行済株式数をかけたものであり，企業価値を評価する際の指標である．時価総額が大きいということは，その企業の資金調達力が高いことを意味する．本手法を通して，上場されていない海外の企業や小さい企業でも，その企業が持つ関係構造から時価総額を予測することが可能である．

*3 売上高，自己資本利益率（ROE），総資産利益率（ROA），株価収益率（PER），純資産倍率（PBR），従業員数，配当利回り，自己資本比率，および資本金などのファンダメンタル指標を用いる．

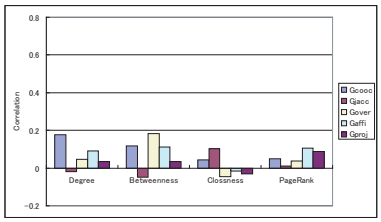


図 3: 各ネットワークのランキングと研究者の論文数ランキングとの相関.

させる．表 2 に示しているように，５つの関係の β_i の重みで最大 0.408 の相関を得ることができる．これらの重みは，各関係の企業価値における影響度を表しており，提携関係や名前の共起関係が企業に対する影響力が大きいことが分かる．

表 1: 各関係を w_i の重みで組み合わせたネットワークのランキングと企業価値との相関.

R_j	w_{cooc}	w_{jacc}	w_{over}	w_{align}	$w_{lawsuit}$	Cor
R_D	0.5	0.2	0.1	0.2	0.0	0.494
R_B	0.7	0.0	0.1	0.2	0.0	0.430
R_C	0.4	0.5	0.0	0.0	0.1	0.626
R_P	0.5	0.0	0.1	0.1	0.3	0.386

表 2: 各関係に β_i の遷移確率を付与した場合に得られるネットワークのランキングと企業価値との相関.

β_{cooc}	β_{jacc}	β_{over}	β_{align}	$\beta_{lawsuit}$	Cor
0.66	0.00	0.14	0.10	0.00	0.408

最後に，ランキングの組合わせ法では，各ネットワークの中心性の指標をその企業の特徴量として，SVM 学習を行う．そして，企業のファンダメンタル指標を特徴量とした場合と，両方を組合わせた場合の予測結果と比較する．表 3 で示しているように，関係だけを用いて企業の価値を予測することは，企業の財政状況をそのまま評価するファンダメンタル指標より劣れるが，両方を組合わせることでより良い予測が可能であることが分かる．これは，企業のネットワークにおける関係的・構造的特徴が，企業自身の属性として働くことを示唆する．

5.2 研究者ネットワークに基づく研究者ランキングの予測

学術論文は，複数の研究者で共著されることが多い．研究者の社会ネットワーク上で良い位置にいる研究者のほど，多くの論文を書けると言えるだろう．すると，論文の書く生産性を向上させるような研究者のネットワークは存在するのであろうか？

図 3 は，それぞれのネットワークの内部ランキングと外部ランキング（論文数のランキング）との相関を表す．結果から見ると，論文の数は，この５つのネットワークの内部ランキングと相関が低い．次に，ネットワークの組合わせ法を用いた（NetRank 法を用いる）場合，最適なパラメータは，$\{\beta_{cooc}, \beta_{jacc}, \beta_{over}, \beta_{affi}, \beta_{proj}\}$ は，$\{0.74, 0.00, 0.01, 0.14, 0.01\}$ の場合で，予測結果 0.207 の相関しか達してない．最後に，ランキング組合わせ法を用いた場合，少しだけ高い相関 0.326 を表す．

これらの結果からは，研究者の論文数によるランキングは，今回使った５つのネットワークと関連性が低いことが分かる．その理由として，研究者の論文の数はそもそも関係に依存する指標でない可能性と，今回用いた社会ネットワークが研究者のランキングを表すに適切でない可能性がある．今後は，どのようなランキングが関係の影響を受けやすいか，どのような関係がランキングに影響しやすいかなどを検討する必要があるだろう．

表 3: ファンダメンタル指標とネットワーク指標，および両方における企業のランキングと企業価値との相関

$Cor(W \cdot R_j^{(G_i)}, \hat{R})$	$Cor(R^{(F)}, \hat{R})$	$Cor(R^{(BOTH)}, \hat{R})$
0.512	0.612	0.644

6. Conclusion

本稿では，Web から抽出した社会ネットワークに基づいてエンティティのランキングを予測する方法を提案した．単純比較法，ネットワークの組合わせ法，ランキングの組合せ法の３つのランキング学習手法を提案することで，企業と研究者のランキングをそれぞれの Web 上のネットワークから学習した．企業の価値分析に良く使われているファンダメンタル指標のほか，企業を取り囲む関係構造が企業の価値に強く影響することが分かった．研究者の場合は，論文の数が必ずしも今回用いた関係と相関があるとは限らない．今後は，より多様なランキングと多様な社会ネットワークを用いて提案手法を適応してみる．

参考文献

[Agarwal 06] Agarwal, A., Chakrabarti, S., and Aggarwal, S.: Learning to rank networked entities, in *Proc. KDD'06* (2006)

[Chang 00] Chang, H., Cohn, D., and McCallum, A.: Creating customized authority lists, in *Proc. ICML2000* (2000)

[Getoor 05] Getoor, L. and Diehl, C. P.: Link Mining: A survey, *SIGKDD Explorations*, Vol. 2, No. 7 (2005)

[Jin 07] Jin, Y., Matsuo, Y., and Ishizuka, M.: Extracting Social Networks Among Various Entities on the Web, in *ESWC2007* (2007)

[Kautz 97] Kautz, H., Selman, B., and Shah, M.: The Hidden Web, *AI magazine*, Vol. 18, No. 2, pp. 27–35 (1997)

[Matsuo 06] Matsuo, Y., Mori, J., Hamasaki, M., Ishida, K., Nishimura, T., Takeda, H., Hasida, K., and Ishizuka, M.: POLYPHONET: an advanced social network extraction system, in *WWW2006* (2006)

[Mika 05] Mika, P.: Flink: semantic web technology for the extraction and analysis of social networks, *Journal of Web Semantics*, Vol. 3, No. 2, pp. 211–223 (2005)

[Scott 00] Scott, J.: *Social Network Analysis: A Handbook (2nd ed.)*, SAGE publications (2000)

[Staab 05] Staab, S., Domingos, P., Mika, P., Golbeck, J., Ding, L., Finin, T., Joshi, A., Nowak, A., and Vallacher, R.: Social networks applied, *IEEE Intelligent Systems*, Vol. 20, No. 1, pp. 80–93 (2005)

[金 07] 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)

[松尾 05] 松尾 豊, 友部 博教, 橋田 浩一, 石塚 満：Web 上の情報からの人間関係ネットワークの抽出, 人工知能学会論文誌, Vol. 20, No. 1 (2005)