

# Web上の情報から人間関係ネットワークの抽出

友部 博教

## 1 はじめに

我々は現実世界においても、または Web 上のオンラインの世界においても多種多様なコミュニティに存在している。そのコミュニティは、大学の一研究室の様に数名から十数名ほどで成り立つ小さいものから、何千人という会員を抱える学会などの学術組織まで、様々である。そのコミュニティにおいて、所属するメンバーを特徴づけるのは、人間関係は、所属するメンバーを特徴づけるのに大きな役割を持っている。たとえば、コミュニティの中心人物ならば彼の発言は他のメンバーに大きな影響を与えるものになるだろう。またコミュニティがいくつかのクラスタに分かれていれば、それが同じ興味を持つ集まりであることを知ることができる。

人間関係を知っていればコミュニティにおけるコミュニケーションも活発にすることができる。たとえば、コミュニティにおける友人の重要度について知ることができる。また、あまり面識のない人物と出会ったとき、自分とはどのような人間関係でつながっているのか知ることができる。

本稿では、学会における人間関係を自動的に抽出する手法について提案する。この方法は、従来の社会学において多くの質問から人間関係を抽出するのではなく、Web上の情報を探索することによって人間関係を抽出する。人間関係を考慮することによって、我々は人間関係ネットワークを構築する。この人間関係ネットワークにおいて、ノードはコミュニティ内のメンバーを表し、エッジは二人の人間の人間関係を表すことになる。また、エッジに付加されたラベルは、人間関係の種類を表している。このネットワークによって、我々は二人の間の関係の情報を知ることができる。

我々のシステムは第 17 回情報処理学会全国大会にて学会支援として運用された。

## 2 ノードとエッジの抽出

### 2.1 基本的なアルゴリズム

人間関係ネットワークを構成するメンバーはあらかじめ決められているとする。つまり、ネットワークのノードは所与である。たとえば、JSAI2003 などの学会の参加者の氏名は、開催に先立って公開されている。また特定の学会誌の過去の論文の著者リスト、情報系の研究者のリストなどを入手すれば、特定の学会や分野における研究者の氏名リストを入手することが可能である後述するように、同姓同名の問題に対処するために、参加者の氏名に加え所属情報も得る。また、これらの情報は、Web上に公開されている過去の学会のプログラムから獲得することができる。

次に、ノード間にエッジを付与する処理を行う。

“松尾豊 石塚満”

と入力する (両者は AND の関係である)。ここでヒット件数が多ければ、氏名が共起する傾向が強く、関係が強いであろうことが推測される。このように、本手法では基本的に、Web 文書における氏名の共起の強さによって関係の強さを推測する。

### 2.2 共起の強さを正確に知る

氏名の共起の強さを知るために、両者の名前の AND をとりヒット件数を得ることは有用である。しかし、それを単純に関係の強さの推測地とするのは問題がある。

共起の強さを測るために、共起頻度以外にもさまざまな指標がある。集合の類似度、重なり具合を表す指標として、Dice 係数や Simpson 係数のようにさまざまなものが提案されている [5]。

ここでは、次式で表される閾値付き Simpson 係数を用いた。

同発表関係 同じ研究会や国際会議で発表する  
(した)関係

$$R(X, Y) = \begin{cases} \frac{|X \cap Y|}{\min(|X|, |Y|)} & \text{if } |X| > k \& |Y| > k, \\ 0 & \text{otherwise.} \end{cases}$$

ここでは、氏名「X」と氏名「Y」の単独でのヒット件数をそれぞれ  $|X|$ 、 $|Y|$ 、ANDをとったときのヒット件数をそれぞれ  $|X \cap Y|$ と表記する。ただし、 $R(X, Y)$ は「X」と「Y」の関係の強さを表す関係であり、 $K$ は閾値である。JSAI2003の場合、 $k=30$ とした。つまり単独でのヒット件数が30件以下の人はエッジが張られない。(孤立ノードになるので、ネットワークから除外する。)  $R(X, Y)$ はネットワークを構築する際に、閾値より高ければエッジをはり、そうでなければエッジを張らないという基準を用いる。また、エッジの長さとして用いることもできる。

### 2.3 同姓同名の問題

Web上には同姓同名の人が多数存在し、対象となる人物以外の人物も検索にヒットするために、単独でのヒット件数が増えることがある。この場合、二つのノードの関係の強さ  $R(X, Y)$ を正確に得ることが不可能になるため、所属情報を用いた絞込みを行う。例えば、対象となる人物が産業総合研究所の松尾豊氏の場合

“松尾豊 産業技術総合研究所”

というクエリを用い検索する。JSAI2003の場合には、学会のプログラムに記載されている所属情報を用いた。

### 2.4 エッジラベルの抽出

社会的関係の種類として、本論文では研究分野に特有の次のようなクラスを定める。これを、エッジのラベルとしてネットワークに付与する。

共著関係 共著の論文がある関係

同研究室関係 同じ研究室や研究所のメンバーなど所属が同じである(あった)関係

同プロジェクト関係 同じプロジェクトや委員会など、組織をまたがる同グループに所属している(いた)関係

以後、誤解のない範囲で、「共著」「研究室」「プロジェクト」「発表」と略記する。関係のクラスは複数取り得る。さて、このような関係を抽出するために、まず検索エンジンに

“X AND Y”

をクエリとして入力し、上位5ページを取得する次に、それぞれのページから表1にあるような属性の値を抽出する。属性GroTitle、GroFFive属性は、そのページが何に関するページであるかを判断するためのものであり、別に定義した語群を用い、語群Aがタイトルに出現するかどうか(GroTitle(A)属性)、語群Bが最初の5行に出現するか(GroFFive(B)属性)などを表す。各語群は手動で設定する方法もあるが、できるだけ自動化するために、あらかじめ正解クラスの付与されたページを用い、各クラスごとにTF-IDF値の上位後を語群としている。

表 2: 属性に利用する語群の例

Group	Words
A	出版 (publication), 論文 (papers) など
B	メンバー (members), 研究室 (lab) など
C	プロジェクト (project) など
D	ワークショップ (workshop) など
E	学会 (society), 団体 (association) など
F	教授 (professor), 専攻 (major) など

これらの属性からクラスを判別するルールが得られればよい。これは、属性からクラスを予測するルールを学習する問題となる。そこで本研究では、C4.5[4]を用いて判別ルールを生成する。ランダムに抽出した275ページを手で正解クラスを付与し、これを訓練例として用いた。獲得したルールを表3に示す。

## 3 人間関係ネットワークの評価

### 3.1 ラベルの精度

表4は判別されたラベルの適合率を表す。評価には、人間関係ネットワーク上のランダムに選んだ200エッジを手でクラス分けしたものを正解例として求めた。共著ラベルと発表ラベルは他のクラスに比

表 1: ページの特徴量

属性	説明	値
NumCo	二人の氏名の共起回数	zero, one, or more_than_one
SameLine	二人の氏名が同じ行に一度以上出現するか	yes, or no
Rel	関係の強さが閾値以上か	yes, or no
FreqX	X の出現頻度	zero, one, or more_than_one
FreqY	Y の出現頻度	zero, one, or more_than_one
GroTitle	タイトルに語群 (A-F) が出現するか	yes or no (それぞれの語群に対して)
GroFFive	最初の 5 行に語群 (A-F) が出現するか	yes, or no (それぞれの語群に対して)

表 3: 獲得した判別ルール例

クラス	判別ルール
Coauthor	SameLine = yes
Lab	(NumCo = more_than_one & GroFFive(F) = no)
Proj	(FreqX = one & GroTitle(B) = yes)
Conf	(FreqY = more_than_one & GroTitle(D) = yes)

べて高い適合率を得た。表 3 をみればわかるように、この二つのクラスは非常に簡単なルールを持っている。

表 4: ラベルの適合率

ラベル	適合率
共著	91.8% (90/98)
研究室	70.9% (73/103)
プロジェクト	74.4% (67/90)
発表	89.7% (87/97)

表 5 は判別されたラベルの再現率を表す。ここで示されるように、発表ラベルの再現率が低くなっている。これは、コミュニティのメンバーが参加することのできる研究会が多く存在し、その研究会ごとに Web ページの特徴が異なっているためであると考えられる。

表 5: ラベルの再現率

ラベル	再現率
共著	97.8% (90/92)
研究室	86.9% (73/84)
プロジェクト	91.8% (67/73)
ラベル	67.4% (87/129)

### 3.2 アンケートによる評価

JSAI2003 の後、人間関係ネットワークに関するアンケート調査を行った。調査対象者は、JSAI2003

に参加登録した人の中から 141 人全員を対象者とした。82 名から回答を得、回収率は 58%であった。一人あたり各 20 人の相手との関係を質問した。質問は、一人の相手あたり各 15 問で「同じ研究室や部署など 30 人規模の組織に同時期に所属していたか」や「研究発表を聞いたり論文を読んだことがある」といった内容である。

表 6 に、JSAI2003 で表示したネットワークのエッジラベルに対して、アンケートから得た回答を正解とした適合率および再現率を示す。また、表 7 は、抽出した全関係に対する適合および再現率であり、ネットワーク中にエッジラベルとして表示していないものも含む。言い換えれば、関係の強さ  $R(X, Y)$  が閾値以下のノードペアに対するラベルも含んでいる。

表 4 と表 5 と比較して、表 6、表 7 は、適合率、再現率ともに低い値になっている。この理由として、回答者が共著やプロジェクトの関係を忘れていたり、記述もれしているなどの可能性があることや、プロジェクトの定義としてより広いものを想定しており、再現率が低くなっているといった、アンケートの回答に関する問題がある。しかし、最も大きな原因、特に再現率が低いことに対する原因として考えられるのは、この手法は、すべての Web ページを網羅的に分析しているわけではないことや、そもそも、すべての情報が Web 上にあるわけではないということである。そもそも、Web 上にない情報から関係を把握することはできないので、本手法のアプローチ

が再現率に対して限界があることは明らかである。しかし、関係の強さと併せて用いることで、表 6 で示したように 80%程度 の適合率で共著、研究室、発表などの関係を抽出できると言うことは、学会におけるコミュニケーション支援という目的には有用であろう。また、本論文で用いた方法はシンプルなものであり、属性の取り方や学習の方法を工夫することで、さらに高い精度を得ることも可能であると考えられる。

表 6: JSAI2003 ネットワークにおけるエッジラベルのアンケートによる評価

クラス	適合率	再現率
共著	89.0% (81/91)	32.1% (81/252)
研究室	78.3% (72/92)	18.7% (72/385)
プロジェクト	50.0% (9/18)	3.0% (9/300)
発表	79.5% (35/44)	6.5% (35/538)

表 7: 抽出された全エッジラベルのアンケートによる評価

クラス	適合率	再現率
共著	78.5% (135/172)	53.6% (135/252)
研究室	55.6% (109/198)	28.3% (109/385)
プロジェクト	20.3% (60/296)	20.0% (60/300)
発表	39.9% (222/556)	41.3% (222/538)

## 4 関連研究

Referral Web [2] は Web 上で対象人物へのつながりを発見するプロジェクトである。しかし、我々の手法ではまず最初にコミュニティのメンバーのリストを抽出し、そこから人間関係ネットワークを構築するという点で異なる。また関係の種類としてラベルも付加している。

村田らは検索エンジンで検索された件数を用いて Web ページ間の関係を発見する手法を提案した [3]。この手法でも検索件数を用いているが、さらにラベル判別を行い、関係の種類にかんしても言及している点で異なる。

Dan Brickley や Libby Miller は FOAF(Friend of A Friend) という RDF vocabulary を開発し、それを用いて人間関係ネットワークを作成した [6, 1]。Foaf はそれぞれのユーザが自分の Web サーバ上に Foaf ファイルを作成し URL を共有する。このシス

テムではユーザは自分自身の人間関係(誰と知り合いか)を記述しなければならない。しかし、このシステムでは自動的にユーザの人間関係を抽出することができる。

## 5 おわりに

コミュニティにおける人間関係は、そのコミュニティをより緊密にするためには重要な情報となる。本論文では人間関係ネットワークを Web から抽出する手法を提案したが、例えばどういう人間関係の人が近くにいるときにはどういう支援を行えばいいのか、人間関係ネットワークの位置とユーザの活動状況の関係はあるのか、活発な学会はどのような人間関係ネットワークはどういう特徴を持っているのかなど、今後、様々な研究が可能であると考えられる。

## 参考文献

- [1] Foaf: the 'friend of a friend' vocabulary. <http://xmlns.com/foaf/0.1/>.
- [2] H. Kautz, B. Selman, and M. Shah. Referral web: Combining social networks and collaborative filtering. *Communications of the ACM*, 30, 1997.
- [3] T. Murata. Finding related web pages based on connectivity information from a search engine. In *Proc. 10th WWW Conf.*, 2001.
- [4] J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, CA, 1993.
- [5] E. Rasmussen. Clustering algorithms. *Information Retrieval: Data Structures & Algorithms*. William B. Frakes and Ricardo Baeza-Yates (Eds.), Prentice Hall, 1992.
- [6] XML watch: Finding friends with xml and rdf. <http://www-106.ibm.com/developerworks/xml/library/x-foaf.html>, 2002.