Web 上の情報からの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 [2]。

ここでは、次式で表される閾値付き Simpson 係数を用いた。

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

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

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 2: 獲得した判別ルールの例

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)

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 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

同発表関係 同じ研究会や国際会議で発表する（した）関係

これらのラベルを抽出するために、まず対象となるエッジが結ぶ二つのノードの氏名を検索エンジンのクエリとして入力する。そして獲得できたページ (ここでは Google のスコア上位 5 ページ) から、ページを表す特徴を抽出する。そこで本研究では、C4.5[1] を用いて、これらの特徴を属性とした判別ルールを生成する。ランダムに抽出した 275 ページを人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 2 に示す。

3 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] J. R. Quinlan. *C4.5: Programs for Machine Learning.* Morgan Kaufmann, CA, 1993.

[2] E. Rasmussen. Clustering algorithms. *Information Retrieval: Data Structures & Algorithms. William B. Frakes and Ricardo Baeza-Yates (Eds.), Prentice Hall*, 1992.